SURVEYING SUBJECTIVE PHENOMENA

1

CHARLES F TURNER & ELIZABETH MARTIN
Editors

SURVEYING SUBJECTIVE PHENOMENA

VOLUME 1

SURVEYING SUBJECTIVE PHENOMENA

VOLUME 1

CHARLES F. TURNER AND
ELIZABETH MARTIN, EDITORS

Panel on Survey Measurement of Subjective Phenomena
Committee on National Statistics
Commission on Behavioral and Social Sciences and Education
National Research Council

RUSSELL SAGE FOUNDATION NEW YORK

THE RUSSELL SAGE FOUNDATION, one of the oldest of America's general purpose foundations, was established in 1907 by Mrs. Margaret Olivia Sage for "the improvement of social and living conditions in the United States." The Foundation seeks to fulfill this mandate by fostering the development and dissemination of knowledge about the political, social, and economic problems of America. It conducts research in the social sciences and public policy, and publishes books and pamphlets that derive from this research.

The Foundation provides support for individual scholars and collaborates with other granting agencies and academic institutions in studies of social problems. It maintains a professional staff of social scientists who engage in their own research as well as advise on Foundation programs and projects. The Foundation also conducts a Visiting Scholar Program and a Postdoctoral Fellowship Program.

The Board of Trustees is responsible for oversight and the general policies of the Foundation, while the immediate administrative direction of the program and staff is vested in the President, assisted by the officers and staff. The President bears final responsibility for the decision to publish a manuscript as a Russell Sage Foundation book. In reaching a judgment on the competence, accuracy, and objectivity of each study, the President is advised by the staff and a panel of special readers.

The conclusions and interpretations in Russell Sage Foundation publications are those of the authors and not of the Foundation, its Trustees, or its staff. Publication by the Foundation, therefore, does not imply endorsement of the contents of the study. It does signify that the manuscript has been reviewed by competent scholars in the field and that the Foundation finds it worthy of public consideration.

THE NATIONAL RESEARCH COUNCIL was established by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and of advising the federal government. The Council operates in accordance with general policies determined by the Academy under the authority of its congressional charter of 1863, which establishes the Academy as a private, nonprofit, self-governing membership corporation. The Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in the conduct of their services to the government, the public, and the scientific and engineering communities. It is administered jointly by both Academies and the Institute of Medicine. The National Academy of Engineering and the Institute of Medicine were established in 1964 and 1970, respectively, under the charter of the National Academy of Sciences.

The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the Councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the panel responsible for the report were chosen for their special competences and with regard for appropriate balance.

This report has been reviewed by a group other than the authors according to procedures approved by a Report Review Committee consisting of members of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine.

Copyright © 1984 by Russell Sage Foundation. All rights reserved. Printed in the United States of America. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher. This work may, however, be reproduced in whole or in part for the official use of the U.S. Government on the condition that copyright notice is included with such official reproduction.

Library of Congress Catalog Number: 83-61131
International Standard Book Numbers: 0-87154-882-8 (Volume 1),
0-87154-883-6 (Volume 2), 0-87154-881-X (set)

Panel on Survey Measurement of Subjective Phenomena

ROBERT P. ABELSON, Department of Psychology, Yale University

BARBARA A. BAILAR, Bureau of the Census

MARIAN BALLARD (Administrative Secretary), National Research Council staff

THERESA J. DEMAIO (Research Associate), National Research Council staff

OTIS DUDLEY DUNCAN (Chair), Department of Sociology, University of Arizona

BARUCH FISCHHOFF, Decision Research, Eugene, Oregon

LESTER R. FRANKEL, Audits and Surveys, Inc., New York, New York

WILLIAM H. KRUSKAL, Department of Statistics, University of Chicago

MICHAEL B. MACKUEN, Department of Political Science, Washington University

CATHERINE MARSH, Social and Political Sciences Committee, Cambridge University

ELIZABETH MARTIN* (Research Associate), National Research Council staff

SARA B. NERLOVE, School of Urban and Public Affairs, Carnegie-Mellon University

HOWARD SCHUMAN, Department of Sociology and Institute for Social Research, University of Michigan

TOM W. SMITH, National Opinion Research Center, University of Chicago

CHARLES F. TURNER (Study Director), National Research Council staff

NOTE: The Panel on Survey Measurement of Subjective Phenomena conducted this study under the auspices of the Committee on National Statistics: Stephen E. Fienberg, Jean D. Gibbons, Zvi Griliches, Clifford G. Hildreth, Nathan Keyfitz, Leslie Kish, Gary G. Koch, Paul Meier, Lincoln E. Moses, Ingram Olkin, Burton H. Singer, Judith M. Tanur, and Edward R. Tufte.

*Member of the panel, January-May, 1980; subsequently, member of the staff.

Note on Authorship

For each major section of this volume, the list below identifies the persons who took primary responsibility for original drafting of the text. The panel reviewed all sections and all have been edited in light of panel reactions and the comments of outside reviewers appointed under procedures prescribed by the Report Review Committee of the National Academy of Sciences. It is possible, therefore, that the persons responsible for the original draft of a section may not endorse every statement in the text as published in this volume. Nor does panel review imply literal and complete endorsement of all text by every member. It is in the nature of collaborative efforts that participants may have more or less serious reservations about portions of a collective product. Panel members were given and took the opportunity of publishing formal statements of amendment or dissent in regard to matters they considered important enough to warrant such treatment (see Appendix A). The purpose of the following listing, therefore, is primarily to give credit for individual contributions but not to assign final responsibility for the published text. It should also be noted that while the following list covers major sections of this volume, it is frequently the case that paragraphs or pages by other hands may appear within these sections.

Chapter 1 Duncan (1.2, 1.3), Fischhoff (1.4), Turner (1.1, 1.4, 1.5).

Chapter 2 DeMaio, Marsh, Turner. This chapter and Chapter 3 were a collaborative effort; the nature of this collaboration was such that it would be difficult to divide responsibilities or identify an order of precedence. Section 2.1 was contributed by Duncan.

Chapter 3 DeMaio, Marsh, Turner. (See note above.)

Chapter 4 Kruskal (4.1), Bailar and Frankel (4.2), Abelson (4.3).

Chapter 5 Schuman (5.1), Smith (5.2), Turner (5.2, 5.3).

Chapter 6 Duncan.

Chapter 7 MacKuen (7.1), Fischhoff (7.2), Nerlove (7.3).

Chapter 8 Martin.

Chapter 9 Martin (9.1, 9.2, 9.4), Abelson (9.3).

Chapter 10 This chapter is a collective product of the panel and staff. It presents our recommendations and is a direct result of our deliberations. Original drafting was done (mainly) by the chair, and revisions were made by the editors.

Appendixes DeMaio, Marsh and Turner (B, C), DeMaio and Turner (D), Smith and Turner (F, G), Martin (H). The material presented in Appendix E is reprinted with the permission of Union Carbide Corporation.

The editors divided primary responsibility for this volume as follows: Turner, Chapters 1–6; Martin, Chapters 7–9; Chapter 10 was jointly edited.

Contents

Contents, Volume 2 *xi*

Preface *xiii*

Part I Background *1*

1 Introduction *3*

1.1 Origins of the Study *3*

1.2 Domain of the Study: Objective and Subjective Phenomena *7*

1.3 Subjectivity and "Social Facts" *10*

1.4 Putting Surveys and Their Problems in Perspective *14*

1.5 Organization of the Volume *18*

Part II Uses and Abuses of Surveys *23*

2 The Development and Contemporary Use of Subjective Surveys *25*

2.1 The Development of the Survey Method *25*

2.2 The Survey Enterprise *28*

2.3 The Use of Subjective Survey Measurements *31*

2.4 The Effects of Surveys and Polls *51*

3 Standards of Practice and the Misuse of Surveys *61*

3.1 Attempts at Setting Survey Standards *62*

3.2 Indicators of the Quality of Survey Practice and Reporting *68*

3.3 Misuses of Polls and Surveys *73*
3.4 Conclusions *90*

Part III Measurement Issues *95*

4 Measurement and Error: An Introduction *97*
4.1 Some Fundamentals *97*
4.2 Measurement Variation in Surveys: Examples of the Effects of Respondents, Coders, and Interviewers *106*
4.3 Psychological Measurement: An Introduction to the Subjective Domain *117*

5 Variability in Survey Measurements of Subjective Phenomena: Empirical Results *129*
5.1 Response Effects With Subjective Survey Questions *129*
5.2 Reproducibility of Survey Measurements of Subjective Phenomena *148*
5.3 Coordinated Measurement and the Control of Error *166*

6 Measurement and Structure: Strategies for the Design and Analysis of Subjective Survey Data *179*
6.1 Objects and Subjects *181*
6.2 Relating Measurements of Subjective and Objective Phenomena *191*
6.3 Response Structures *197*
6.4 The Latent Trait Approach in Survey Research: The Rasch Measurement Model *210*

Part IV The Survey Interview Process *231*

7 Conceptual Ambiguity in Surveys *235*
7.1 The Concept of Public Opinion *236*
7.2 Using Surveys to Clarify Concepts: The Example of Risk *245*

7.3 Techniques for Exploring Frames of Reference in Surveys *251*

8 The Role of the Respondent *257*
8.1 The Interviewer's Approach to the Respondent *258*
8.2 The Respondent's Concept of His Role *262*
8.3 Social Expectations *271*
8.4 The Presence of Others During the Interview *272*
8.5 Implications *273*

9 The Question-and-Answer Process *279*
9.1 Rules for Asking and Answering Questions *280*
9.2 The Individual Survey Question as a Source of Bias *283*
9.3 Contextual Influences Within Sets of Survey Questions *287*
9.4 The Tasks Posed by Survey Questions *295*

Part V Improving Survey Measurement of Subjective Phenomena *303*

10 Recommendations *305*
10.1 Improving Public Understanding of Surveys and Polls *306*
10.2 Upgrading Current Survey Practice *310*
10.3 Advancing Survey Measurement and the Scientific Use of Survey Data *319*

Appendixes *335*

A Separate Statements *337*
B Notes on the Size and Scope of the Survey Industry *341*
C Details of Newspaper Study *353*

D On the Review of Federal Surveys *369*
E Details of Survey Results Cited in Union Carbide Example *371*
F Intersurvey Comparisons: Early Work *377*
G Intersurvey Comparisons: Recent Work *391*
H Scheme for Classifying Survey Questions According to Their Subjective Properties *407*

References *433*
Name Index *473*
Subject Index *483*

Contents

Volume 2

Part I Measurement of Subjective Phenomena in the Social Sciences *1*

1 Attitude Measurement in Psychology and Sociology: The Early Years *3*
Jean M. Converse

2 Utility in Economics: A Survey of the Literature *41*
J. G. Tulip Meeks

3 The Use of Survey Data in Basic Research in the Social Sciences *93*
Stanley Presser

Part II Quasi-Facts *115*

4 The Subjectivity of Ethnicity *117*
Tom W. Smith

5 Measuring Employment and Unemployment *129*
Barbara A. Bailar and Naomi D. Rothwell

6 Housing Research: Conceptual and Measurement Issues *143*
Sandra J. Newman

Part III Nonsampling Sources of Variability *157*

7 Why Do Surveys Disagree? Some Preliminary Hypotheses and Some Disagreeable Examples *159*
Charles F. Turner

8 Nonattitudes: A Review and Evaluation *215*
Tom W. Smith

9 Social Desirability and Survey Measurement: A Review *257*
Theresa J. DeMaio
10 The Manner of Inquiry: An Analysis of Survey Question Form Across Organizations and Over Time *283*
Jean M. Converse and Howard Schuman

Part IV Some Statistical Models for Error and Structure in Survey Data *317*
11 Some Statistical Models for Analyzing Why Surveys Disagree *319*
Clifford C. Clogg
12 Rasch Measurement in Survey Research: Further Examples and Discussion *367*
Otis Dudley Duncan
13 Measurement Error in Surveys *405*
Judith T. Lessler

Part V Putting Survey Measurements in Context *441*
14 Reality, the Press, and Citizens' Political Agendas *443*
Michael B. MacKuen
15 Mass Media, Contraceptive Behavior, and Attitudes on Abortion: Toward a Comprehensive Model of Subjective Social Change *475*
James R. Beniger
16 The Popularity of Presidents, 1963–80 *501*
Michael B. MacKuen and Charles F. Turner
17 Cultural Indicators and the Analysis of Public Opinion *547*
Elizabeth Martin
18 Do Polls Affect What People Think? *565*
Catherine Marsh

Preface

Surveys are a relatively recent invention. Their modern form evolved from the experiences of election polling in the 1930s and from breakthroughs made during the same decade in developing the statistical theory of sampling. The mating of sampling theory and systematic interview techniques made it theoretically possible to estimate the characteristics of very large populations by studying relatively small samples.

Although the speculative use of surveys to make election forecasts led to some early disappointments (most notably in the 1948 presidential election), the survey enterprise has expanded greatly in size and influence during the ensuing decades. Today tens of millions of Americans are surveyed annually, and a veritable avalanche of newspaper stories, television broadcasts, and government and scholarly publications report the results of these surveys. Business and government agencies routinely conduct surveys for a wide variety of purposes. And in the social sciences, surveys have become one of the most commonly used methods of empirical research.

The amazing growth of the survey enterprise since the 1930s has been accompanied by persistent concerns about the reliability, validity, and usefulness of survey data. These concerns have been voiced by persons both inside and outside of the survey field. Such concerns provided the general impetus for the present study.

Unlike the Social Science Research Council's review of the 1948 election forecasts, this study did not arise from a single dramatic failure, but rather it was prompted by several events that caused some disquiet among people who collect and use survey data. One such event was the discovery of several instances in which seemingly equivalent survey measurements made at approximately the same time produced surprisingly different results. The discovery of these anomalies raised questions initially about the reliability, and ultimately about the meaningfulness, of commonly used survey data. The discovery of these anomalies coincided with a vigorous protest about the conduct of some particular surveys and with an apparent

decline in public willingness to participate in surveys (see Chapter 1 and American Statistical Association, 1974; Steeh, 1982).

These developments led to discussions within and among people from several institutions including the National Research Council-National Academy of Sciences, the Social Science Research Council, and the National Science Foundation.* These discussions led to the initiation of the present study. Because the events precipitating the present undertaking (mainly) involved attitude and opinion surveys, this study focuses on the use, reliability, and meaningfulness of survey measurements of attitudes, opinions, and other subjective phenomena. (The domain and background of the study are more fully described in Chapter 1.)

The Panel on Survey Measurement of Subjective Phenomena was convened in January 1980 under the auspices of the Committee on National Statistics and with the financial support of the National Science Foundation. The panel included social scientists, statisticians, and survey researchers from academic, government, and commercial survey organizations. And, to provide a productive mix of experiences and energy levels, a large complement of young researchers was invited to participate in the panel's work.

The panel held five plenary sessions (as well as assorted working group meetings) between January 1980 and February 1981. This volume is the main intellectual product of the panel's work. The first draft of this volume was completed in June 1981; review, editing, revision, redrafting, and rereview were completed in June 1982. The panel was aided in its work by the individual contributions of authors commissioned to undertake special studies (the reports of their work comprise Volume 2). The panel has also published separately a summary of its findings and recommendations (*Survey Measurement of Subjective Phenomena: Summary Report*, Washington D.C.: National Academy Press). That summary is designed for nonspecialists and is intended to make the problems and progress of the survey enterprise known outside of the profession.

Almost all of the participants in this study were actively involved in collecting or analyzing survey data, or both, but their different intellectual and

*In November 1977, a small group met to begin exploring some of the issues treated in this report. That meeting was hosted by the Institute for Research in Social Science at the University of North Carolina and attended by James A. Davis, Otis Dudley Duncan, Elizabeth Martin, Frank Munger, Robert Parke, Mark Schulman, Howard Schuman, Tom Smith, D. Garth Taylor, and Charles Turner. That meeting planted a seed that was nurtured through a long period of development and review. David Goslin, Sara Kiesler, Edwin Goldfield, and Margaret Martin at the National Research Council; Eleanor Sheldon and Robert Parke at the Social Science Research Council; and Murray Aborn at the National Science Foundation provided valuable assistance during that period.

organizational perspectives encouraged the application of a variety of approaches to the problems of measuring subjective phenomena in surveys. We have tried to integrate and to preserve those different perspectives in editing this volume. The note on authorship (on page vi) and the description of the panel's procedures for adopting recommendations (see Chapter 10) describe the nature of the panel's collaboration.

While almost every participant in this study had some professional stake in the quality and reputation of survey research, the panel approached its work with a shared interest in exploring alternatives to current survey practice and in developing new perspectives and improved methods for survey measurement and analysis. The panel's discussions did not reveal great satisfaction with contemporary survey practices or with the extent of our own understanding of the characteristics of the survey as a measurement instrument. Panelists shared the perception that there is both a great deal of survey practice that could be improved given the present state of knowledge and a great deal that we must begin to learn about fundamental problems of social measurement in order to improve survey research in the future.

We are grateful for the support provided by the National Science Foundation and for the advice given during the early stages of this study by Murray Aborn. We are also grateful to the Russell Sage Foundation for their subsequent support. The latter support made possible a conference and a workshop, held on December 14–15, 1981, to discuss the panel's findings with a large audience of producers and users of survey data. We are also grateful to the Russell Sage Foundation for the publication of these two volumes.

The panel's work was also aided by the Gallup Organization, the Opinion Research Corporation, the *Washington Post* poll, the General Social Survey program of the National Opinion Research Center, and the Question Form, Wording, and Context Project of the Survey Research Center at the University of Michigan. These organizations used their own resources to conduct special experiments that assisted the panel in its work.

The panel also profited from a discussion with Saunders Mac Lane and from critiques of its work by the many reviewers appointed by the Committee on National Statistics and the Assembly of Behavioral and Social Sciences of the National Research Council. In addition, at the request of the editors, James A. Davis provided a detailed critique that helped greatly in our revisions. Naomi D. Rothwell, Stanley Presser, and Miron Straf also provided helpful comments. None of these reviewers, of course, is responsible for the conclusions of this study, but we are grateful for their thought-

ful assistance. We are also indebted to Chris McShane and Elaine McGarraugh for their patient help in preparing this manuscript and to Maura Grant and Priscilla Lewis at the Russell Sage Foundation for their valiant struggles with the army of gremlins that sought to sabotage the typesetting and production of these volumes.

Finally and most crucially, we wish to thank Otis Dudley Duncan for the support and encouragement he gave to us throughout this study. We also wish to express the gratitude that we and the other members of the study owe to him for his willingness to serve as chairman and to contribute his time and talents to this work. As with many subjective phenomena, we acknowledge the existence of this debt and our gratitude to him with no hope of finding words adequate to express them.

Charles F. Turner and Elizabeth Martin

PART I

BACKGROUND

Hasn't it often occurred that instruments originally invented for record and computation have inadvertently so extended the concepts of the entity they were invented to measure (concepts of space, etc.) in the mind and imagination that employed them, that they may metaphorically be said to have extended the original boundaries of the entity measured?

HART CRANE

Primary responsibility for original drafting of this part of the panel's report was assumed by Duncan (1.2, 1.3), Fischhoff (1.4), and Turner (1.1, 1.4, 1.5). See Note on Authorship, page vi.

1

Introduction

1.1 Origins of the Study

Questions about the reliability, validity, and relevance of survey measurements are not new, and we are not the first committee to study those questions. In 1945, the National Research Council (NRC) joined with the Social Science Research Council to establish a Joint Committee on the Measurement of Opinion, Attitudes and Consumer Wants. With financial support from the Rockefeller Foundation, a committee of distinguished social scientists, statisticians, and practitioners was established under the chairmanship of Samuel Stouffer.[1] Among the major topics identified by this group at its first meeting was "the validity of statements, opinion and information furnished by respondents." The minutes of this group's first meeting go on to note that "[this] was considered to be a difficult and long-range problem but a fundamental one for the committee to be concerned with."[2] The four publications directly stimulated by the committee did not, however, concentrate on this question. Rather, they made important contributions toward codifying the theory of sampling (Stephan and McCarthy, 1958), examining the effects of the personal interview situation on survey responses (Hyman et al., 1954), and studying the problems of inference that arise in longitudinal studies (Anderson, 1954; Lipset et al., 1954).

The recent revival of interest in the "fundamental" issues reflects the confluence of several forces. Probably most important of these has been the

apparent success of pollsters and survey researchers in making converts. Surveys have become a ubiquitous component of social life in America. Not only do the press and private industry routinely quote their results, but the federal statistical system has become increasingly committed to the use of survey measurements of opinions and attitudes.

Historically, official statistics in the United States were largely the domain of demographers, economists, and statisticians. Inquiries made by the Bureau of the Census, for example, generally focused on assessments of the size and distribution of the population and a variety of other phenomena that are, at least theoretically, objective, that is, amenable to independent corroboration, such as age, income, and characteristics of industrial establishments. In recent years, however, published national statistics have come to include a growing complement of social statistics on public attitudes, opinions, and so forth (see Executive Office of the President, 1973; U.S. Department of Commerce, 1977; Bureau of the Census, 1980a). One of those recent volumes argued that such measures provide a vital addition to traditional statistics (U.S. Department of Commerce, 1977:XXVI):

> The basic reason for including such [subjective] measures in this report despite the difficulties in their interpretation is that they offer a vital dimension in developing a comprehensive description of the condition of our society and the well being of its members. The bulk of the information presented [in this report] relates to people's objective situation or condition—their jobs, their incomes, their health status, etc. The main purpose of the attitudinal measures is to provide some insight as to how people perceive certain aspects of their conditions. Such data are an essential source of information on people's values and aspirations.

For similar reasons, the recent series of reports by the National Science Board (1973, 1975, 1977) on the state of science in the United States has incorporated data on public attitudes toward science and technology.[3]

Similar trends are also evident around the world. For example, since 1970 the British government has issued an annual report, *Social Trends*, that pays considerable attention to measures of subjective phenomena. Indeed, the authors of this report have argued that "the way forward lies not in adding more measures of conventional hard statistics, but rather in supplementing the existing ones by adding . . . a dimension of the satisfaction (happiness, contentment, psychological well-being, etc.) felt by those who constitute the community" (Abrams, 1973:36). A recent United Nations report (1975:32) echoes this concern.

This growing use of survey measurements of subjective phenomena has again raised questions concerning the validity and reliability of such measurements. Developments or events in four specific areas in 1976–78 served as immediate catalysts for the present study.

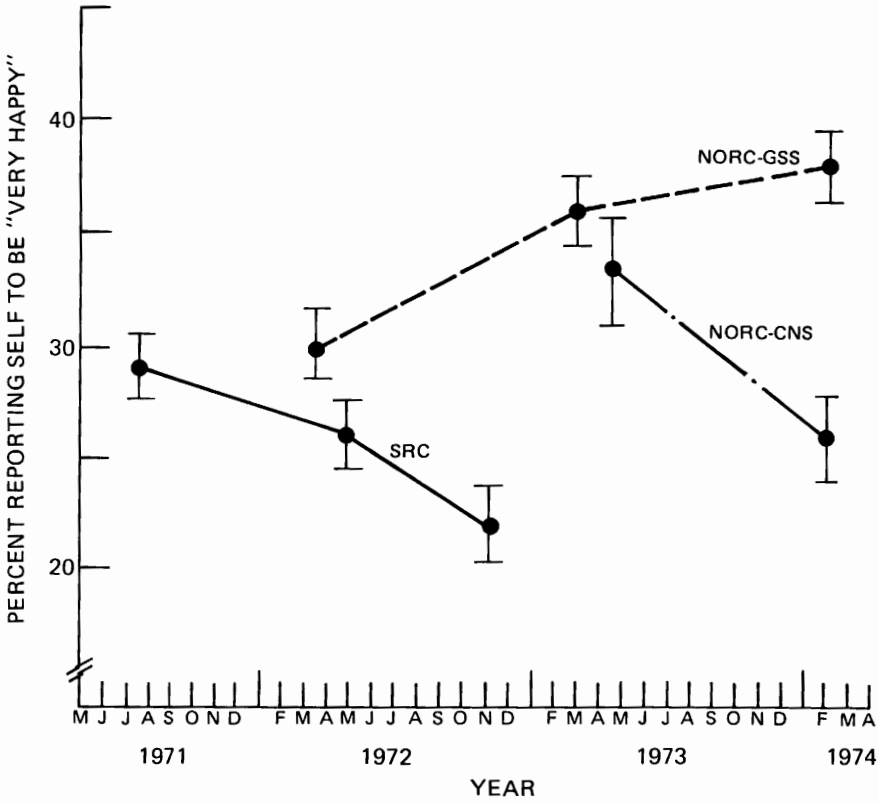
1. *Measurements of Public Confidence in National Institutions.* In 1976–77 an attempt was made to use a series of measurements of public confidence in the leaders of national institutions in a National Research Council report (Kiesler and Turner, 1977; Turner and Kiesler, 1981), but comparison of allegedly equivalent measurements made by two survey organizations showed substantial discrepancies in both the levels of reported confidence and the trends across time in those measurements (Turner and Krauss, 1978; Smith, 1981). While the data were not included in the NRC report, they were widely reported in the media and were incorporated in such federal publications as *Social Indicators* and *Science Indicators.* Indeed, even a recent presidential commission both used and republished these data without commenting upon their fallibility (President's Commission on a National Agenda for the Eighties, 1980:103).

2. *Measurements of Public Attitudes Toward Science.* Surveys of public attitudes toward science that have been commissioned by the National Science Board for the *Science Indicators* series have been the subject of considerable discussion. Questions have been raised (for example, by Mac Lane, 1978; La Porte and Chisolm, 1980), on technical grounds, that some survey questions are misleading or biased and, on philosophical grounds, that the surveys reify public opinion. There are doubts, too, as to whether most people have definite attitudes and knowledge about science. In addition to these questions, evidence has emerged that the *Science Indicators* surveys contained significant anomalies. For example, the responses across time to a question purporting to measure public support for spending on science and technology suggested a very sharp decrease between 1974 and 1976. This result was not consistent with other measurements made during the same period. A change in the content and ordering of survey questions in the 1978 *Science Indicators* survey, however, created a likely (although unproven) artifactual explanation for this decline (see Turner, Chapter 7, Volume 2).

3. *Happiness and the Quality of Life.* Indicators of public happiness, satisfaction, and the perceived quality of life have recently been the subject of serious analysis in economics, sociology, and psychology (see, for example, Easterlin [1974]; Davis [1975b]; Campbell et al. [1976]; Andrews and Withey [1976]). A question asking respondents about their "happiness" has been employed as a reference criterion for validating developmental work on subjective indicators of the quality of life. However, as Figure 1.1 illustrates, the trends obtained by different survey programs over time were not always convergent.[4] While plausible explanations have been advanced to explain these discrepancies, the discrepancies themselves and the specula-

FIGURE 1.1

Survey measurements of self-reported happiness made by three survey research programs



NOTES: Estimates are derived from sample surveys of the noninstitutionalized population of the continental United States age 18 and older. Error bars demark approximately 1 standard error around sample estimates (assuming sampling variance for these clustered samples to be twice that for simple random samples).

Although the questionnaires show a minor divergence in the wording of this question, data from these two sources have been treated as a unitary time-series in other publications (e.g., Campbell, Converse, and Rodgers, 1976; Andrews and Withey, 1976).

QUESTION: "Taken all together, how would you say things are these days—would you say that you are very happy, pretty happy, or not too happy?" (NORC)

"Taking all things together, how would you say things are these days—would you say you're very happy, pretty happy, or not too happy these days?" (SRC)

SOURCES: NORC General Social Survey (GSS), Codebook, 1972–74; NORC Continuous National Survey (CNS) [cycles 1 and 10 + 11]. See Turner, Chapter 8, Volume 2. SRC national data from time-series presented in Campbell, Converse, and Rodgers (1976). Survey dates from Campbell, Converse, and Rodgers (1976); Andrews and Withey (1976); and J. Varvra (personal communication).

tiveness of the explanations raised questions about the meaningfulness, reliability, and error structure of these widely used indicators.

4. Survey of University Faculty. A 1977 survey of American professors (Ladd and Lipset, 1977, 1978) provoked a spirited public controversy concerning the adequacy, accuracy, and appropriateness of survey measures that purported to reflect the respondents' attitudes and opinions (Lang, 1978, 1981). Many of the survey questions were attacked as "ambiguous," "meaningless," or tending to "prejudice" the issues.[5]

The problems illustrated by the preceding examples all appear to involve the influence of nonsampling factors on survey results.[6] It was unclear at the outset of this study whether our understanding of the potential effects of such nonsampling factors in survey data was sufficient to specify and control the relevant factors in the future. Nor could one be confident that policy makers and other users of such survey data appreciated the possibility of systematic errors in such measurements and the implications that such errors might have for their own decision making. Indeed, there were even doubts as to whether error was an appropriate concept to use in the discussion of most of these problems.

It was against this background that the present study was launched. Our aim was to stimulate improvements not only in the practice of survey research and the understanding of the processes which produce its measurements (and its errors), but also to stimulate improvements in the use of research findings by the general public and by clients with all kinds of specialized interests.

1.2 Domain of the Study: Objective and Subjective Phenomena

Our central concern in this study was the poll or sample survey as used both in the social and behavioral sciences and in journalism, marketing, and politics. We followed fairly closely the usage of a committee of the American Statistical Association that recently issued a useful pamphlet *What Is a Survey?* (Ferber et al., 1980):

> Any observation or investigation of the facts about a situation may be called a survey. But today the word is most often used to describe a method of gathering information from a number of individuals, a "sample," in order to learn something about the larger population from which the sample has been drawn. Thus, a sample of voters is surveyed in advance of an election to determine how the

> public perceives the candidates and the issues. A manufacturer makes a survey of the potential market before introducing a new product. A government agency commissions a survey to gather the factual information it needs in order to draft new legislation. For example, what medical care do people receive, and how is it paid for? Who uses food stamps? How many people are unemployed? [p. 3]
>
> Surveys of human populations also provide an important source of basic social science knowledge. Economists, psychologists, political scientists and sociologists obtain foundation or government grants to study such matters as income and expenditure patterns among households, the roots of ethnic or racial prejudice, comparative voting behavior, or the effects of employment of women on family life. [p. 5]

Among surveys, we were especially interested in those surveys, or portions of surveys, that deal with subjective phenomena. (For a review of surveys in general, see Tanur, 1981.)

Subjective phenomena are those that, in principle, can be directly known, if at all, only by persons themselves, although a person's intimate associates or a skilled observer may be able to surmise from indirect evidence what is going on "inside." Objective phenomena are those that can be known by evidence that is, in principle, directly accessible to an external observer. Often that evidence is actually a matter of record, although the relevant records may not be easily sampled for the population of interest.

Thus, if an individual is asked to name a favorite author, to state whether the draft system is fair, to indicate how many children she wants to have, to identify an area in which she or he is afraid to walk at night, or to say whether he or she has ever wished to belong to the opposite sex, the information sought is subjective. But if an individual is asked whether she or he has served in the Armed Forces, how many children she has ever borne, how far she or he lives from the place of work, or to state his or her sex, the information is objective, even though its measurement may rely solely on the respondent serving as an informant rather than on records or reports of observers. When the accuracy of that measurement is questioned, its validity may be tested by referring to records or seeking corroboration from other informants. In household surveys, one respondent may often serve as informant for all members of the household, even for such personal but observable behavior as smoking habits, but such proxy reports ordinarily would not be accepted for subjective phenomena.

We recognize that for some subjective phenomena there is the possibility of substituting behavioral observations for direct questions, particularly when there is presumed reason to distrust the latter. Economists, for example, tend to rely on the principle of "revealed preference," letting actual behavior in the marketplace, such as purchases of goods and services, serve as an indicator for tastes or preferences (see Scitovsky, 1976). Some critics of survey research have urged that greater use be made of "unobtrusive

measures," such as taking inventory of the contents of garbage cans, in place of asking people about their preferences for various foods (particularly alcoholic beverages), on the grounds that biases excessively compromise the validity of respondents' answers (Webb et al., 1966).

However clear the distinction between objective and subjective phenomena may be in principle, there is much blurring in practice. Whether a person is "unemployed," under the definition currently used by federal statistical agencies, turns in part on whether he or she was "seeking work" during a specified period, and no proof of overt activity (such as visiting an employment office) is required to support the respondent's (or proxy respondent's) definition of "seeking work." There are many such "quasi-facts"—for example, ethnic group membership, condition of housing, and being the victim of a criminal act—that similarly allow latitude for the respondent's definition of the criterion. For the sake of economy, surveys often obtain what are at best estimates or at worst mere guesses about matters that are in principle factual or objective, such as how many times the respondent attended religious services in the preceding year or total income of the members of the household.

With these considerations in mind, we note that the realm of the subjective can enter into a survey in two distinct ways: the *information* sought may, itself, pertain to subjective phenomena (rather than factual or objective phenomena); or the *method of securing information* may call for an estimate or impression reported by someone in a more or less favorable position to know the facts, rather than relying upon objective determination by some well-described procedure (see discussion in Fienberg and Goodman, 1974:74–76). Our main concern is with the first category: phenomena that are intrinsically subjective. As just defined, such subjective phenomena are, in principle, directly observable only by respondents (subjects) themselves. However, in view of the importance of quasi-facts in surveys and the enormous possibilities for distortion of reports of objective phenomena, we also give some attention to these hazards of surveys (see Chapter 5 in this volume and Bailar and Rothwell, Chapter 5; Smith, Chapter 4; and Newman, Chapter 6 in Volume 2).

There are at least three general uses for data on subjective phenomena, uses that lead survey organizations to collect such data on a large scale. First, if one wants to forecast some class of behavior, a possible strategy for doing so is to ask people about their intentions or plans—whether they intend to go to college, whether they expect to purchase an automobile in the next several months, for whom they plan to vote, how many children they expect to have, and so on. One can also try to get at the dispositions or motives underlying such decisions and use the evidence of these attitudes to project the size of a market or the outcome of an election. Second, many (though not all) students of politics, economics, demography, and social behavior work with concepts and theories in which subjective phenomena

constitute key explanatory variables. For such work, data on subjective phenomena are needed to understand the behavior for which explanations and theories are sought. Third, some researchers and writers on social indicators stress the importance of subjective social indicators in any attempt to develop comprehensive social reports on the quality of life (Campbell, A., 1976; Campbell, Converse, and Rodgers, 1976; see also Bureau of the Census, 1980a, for an example of the use of subjective measures in a social report). In the United States, of course, the Declaration of Independence affirms the "pursuit of happiness" as one of our rights. One way to inquire how well we are doing in maintaining that heritage is to investigate the state and trend of "happiness." Moreover, in order to interpret social indicators based on objective data, subjective data are also needed to disclose the goals that people may have in mind: for example, how much pollution is too much?

1.3 Subjectivity and "Social Facts"

Since the notion of subjectivity in survey measurements is critical to our inquiry, further analysis of that notion is useful at this point. As an example, suppose each of the following is a true statement: (a) The planet Saturn has rings. (b) The X family had a total money income of $27,500 in year 1980. (c) Mr. X, when asked in a survey, "What is the smallest amount of money a family of four needs to get along in this community?" replied, "$250 a week."

Statement (a) is clearly a matter of objective fact, although any knowledge of how such a fact is established and comes to be scientifically accepted will concede a role to human perception and interpretive processes, both of which may be influenced by social interactions among observers and analysts. But, however great the role of the human factor in this sense, one would not consider statement (a) to be one about subjective phenomena.

Statement (b) in one sense is a similarly objective fact in the sense that, in principle, relevant records might be produced that would be similarly interpreted by competent observers. Although one often uses someone like an adult member of the X family as an informant concerning the family's income—and does not call in other observers or have recourse to records of transactions—that is a matter of convenience rather than principle. But, still, there is a difference between (a) and (b). Both concern objective phenomena. But fact (b) is a social fact, which is to say that its *meaning* incorporates implicit assumptions about society that are shared by some more or less extensive group of people. Part of the implicit meaning of

statement (b) is that a world of purveyors outside the X household is prepared to exchange various goods and services for their dollars; otherwise there would be no such thing as "income." The phenomenon itself has an irreducible subjective aspect (Parsons, 1937:46). But inasmuch as one can count on the willingness of people to exchange goods and services for money, money has an agreed-upon meaning, and the units in which it is measured as well as the documents (currency, etc.) involved in exchange are likewise a matter of agreement. Hence, income is an "objective" phenomenon in much the same sense that the word "mason" (one who builds with brick, stone, or the like) has an objective meaning. In both instances, the meanings are carried by people, but the similarity of meanings carried by different people is great enough so that the implicit subjectivity may escape notice.

In statement (c), we are, *ostensibly*, introducing still a different sort of subjective element. At least for a casual observer (such as an interviewer), there is no way to find out what Mr. X thinks is enough money to get by on but to ask him. Or, at any rate, if one attributes some amount to him, he can deny the attribution without any possibility of being shown up as a liar (while if he lies too much about his own income, the Internal Revenue Service may well show him up). So, on the face of the matter, statement (c) concerns a subjective phenomenon in the fairly strict sense of the word we suggested at the outset: subjective phenomena are those that, in principle, are directly observable only by persons themselves. But there is a bit more to be said.

The Gallup Poll has often used the question cited in statement (c) since 1946, and the resulting data have been studied by Rainwater (1974:Table 3–4). He computes the mean (average) of all responses in each year's survey (Y_t, for the tth year) and compares it with the national disposable personal income per family (X_t) computed from national income statistics. Data are available for 18 years in the 1946–69 period. For those 18 years, Rainwater finds that the linear function

$$Y_t = .515\, X_t$$

provides an excellent fit to the observations ($R^2 = .967$). Over the period in question, real income actually increased more than 50 percent in the United States. But each year's progress in real income was matched by an increase in the amount perceived as "what a family needs," so that the proportionality of the foregoing equation tended to remain stable. Much other evidence (see Chapter 6) supports the proposition that people's sense of what is "needed" or what is "satisfactory" depends on the prevailing level of income, and should the latter change, the perception or standard soon adjusts to the new level. Hence, although Mr. X is indeed speaking for

himself and reporting a subjective phenomenon, that phenomenon itself is shaped by a powerful social consensus, or what Durkheim (1895/1938:XLIX) termed "collective representations." It should be noted that the several hundred Messrs. X who respond to the Gallup Poll need not be aware of the social pressure that is shaping their standards of living and that there is room for idiosyncratic variation in their concepts for structuring of individual consciousness by the social milieu. Statement (c), then—or, rather, the distribution of a set of such responses—is an excellent example of both a "social fact" in Durkheim's sense and a subjective phenomenon in our terms.

Consider a fourth hypothetically true statement: (d) Mrs. X, when asked in the General Social Survey,[7] "Taken all together, how would you say things are these days—would you say that you are very happy, pretty happy, or not too happy?", said, like somewhat more than half of all respondents, "pretty happy." The preceding discussion will have alerted us to the possibility that there is more social content to her response than might appear at first sight. No doubt the notion of what happiness means is one that Mrs. X has largely acquired from her milieu, and she may well ask herself how a person of her age, sex, family status, economic level, community and neighborhood location, and so on, would be expected to respond when "things" are "taken all together." Moreover, there is something very suggestive of a norm of response in the fact that the modal category is the middle one and the response distribution shows no clear trend from year to year. Still, the question invites a personal response. The referent is not a somewhat hypothetical "family of four" (which might differ appreciably from the size of the X family itself) but the respondent—"you"—herself.

The happiness question is indeed prototypical of what advocates of subjective social indicators propose in calling for "direct monitoring of key social-psychological states of the population" (Campbell and Converse, 1972:9). These writers cite perceptions of complex social situations, expectations, hopes, and frustrations of members of the society as basic "subjective data." And Campbell (in Campbell and Converse, 1972:442) proposes a concern with the quality of personal experience and cites as relevant aspects thereof the frustrations, satisfactions, disappointments, and fulfillment that must be assessed or evaluated by the people experiencing these feelings or reactions if they are to be ascertained at all (except in the very indirect way that Durkheim [1895/1938:8] proposed, that is, via the analysis of rates of marriage, birth, and suicide).[8]

There are traps here for the unwary. To be sure, when "social-psychological states" are at issue, the individual presumably *is* the uniquely well-qualified informant. But the fact that respondents are generally compliant and quite willing to engage in introspection does not mean that they do their job well. On some topics a person may be a poor judge of his own

motives, and thus it is no good asking him how his mind works (Nisbett and Wilson 1977; but see also Ericsson and Simon, 1980). Presumably, if disposed to do so, a respondent can tell the interrogator *what* she* intends to do or whether she has made up her mind about something, but not necessarily *why* she'll do it. And while a respondent can no doubt state what he wants in the short run, it is most dangerous to assume that he'll be satisfied if he gets it.

Can any person really assess how much affect or emotion suffuses his definition of a situation, for example, how "strongly" he feels about some public issue? The case is not proven.[9] And certainly there are serious problems, as careful analysts have noted, with data collected by means of the "vague quantifiers" that are often attached to survey questions to provide "measures" of emotional intensity (Bradburn and Miles, 1979). Indeed, Ericsson and Simon's (1980:215) caveat about the use of verbal report data in psychological experiments applies equally well to survey measurements of subjective phenomena:

> Accounting for verbal reports, as for other kinds of data, requires explication of the mechanisms by which the reports are generated, and the ways in which they are sensitive to experimental [or survey] factors (instructions, tasks, etc.).

This point may be well illustrated by contrasting a survey report that reads, "Most people *say* they favor handgun control," with one proclaiming "most people favor handgun control." Such a simple insertion brings strong emphasis to the limitations of the method of inquiry, and it alerts one to the problems that can arise in using such inquiries to predict behavior. (For an example, see Schuman and Presser's [1977–78] attempt to reconcile public support for gun registration in national surveys and legislative inaction on the same issue.)

The extent to which the answer to a question calling for introspection is valid cannot be known a priori; the fact that only the subject herself can know if she is happy does not mean that her report on happiness constitutes a trustworthy scientific datum. A persistent theme of the critics of research on attitudes and opinions (see the classic statements of McNemar [1946] and Blumer [1948]) has been precisely that survey researchers too often take the easy way out. As a result, the research is superficial in regard to both the psychological structuring of responses and the social determinants of individual dispositions.

**Note*. To reduce the ubiquity of male pronouns in the text, we have tried to use both feminine and masculine pronouns for indefinite references. An exception to this rule occurs in Part IV, where we use the feminine form to refer to survey interviewers; this usage is intended to remind readers of the fact that survey data are collected by a labor force that is overwhelmingly female.

As in all scientific inquiry, there is a paradox of observation: to collect the data to be used to test a theory one must presuppose a lawfulness of the behavior involved in making the observation (Reiss, 1980). "Like all existential dilemmas in science . . . the paradox is resolved by a process of approximation" (Kaplan, 1964:54) or iteratively. The more one understands about behavior, the more reliable and valid will be the data collected, and the better the data, the more their analysis will enhance understanding.

1.4 Putting Surveys and Their Problems in Perspective

A large amount of survey data concerning subjective phenomena is generated more or less continuously. At the same time, strong skepticism about the value of such data is expressed from time to time. And even convinced users of these data are likely to acknowledge problems in demonstrating their reliability, validity, and relevance. All experienced research workers are aware, for example, that the proportion of respondents giving a "yes" response on some question can sometimes be modified considerably by altering the wording of the question. Sometimes this knowledge is used in a deliberate way to bias the outcome of a poll, and when this occurs, or is alleged to occur, the news may discredit the entire field of survey research and the polling industry as a whole. There is need, therefore, for greater knowledge about such matters, not only on the part of the public and the consumers of poll and survey results, but also on the part of survey technicians and analysts as well.

There is a similar danger that the evidence of error in data about subjective phenomena that we review might lead to unwarranted and wholesale rejection of survey evidence on important social issues and problems or to ill-advised efforts to police the practice of polling (see, for example, the proposal by the General Accounting Office, 1978). Indeed, looking at our litany of difficulties, some readers may come to wonder whether anything at all can be learned from such surveys or whether surveys have a useful role to play in social thought and policy. The following considerations have led us to answer "yes" to those questions.

FALLIBILITY OF MEASUREMENT IN OTHER SCIENCES

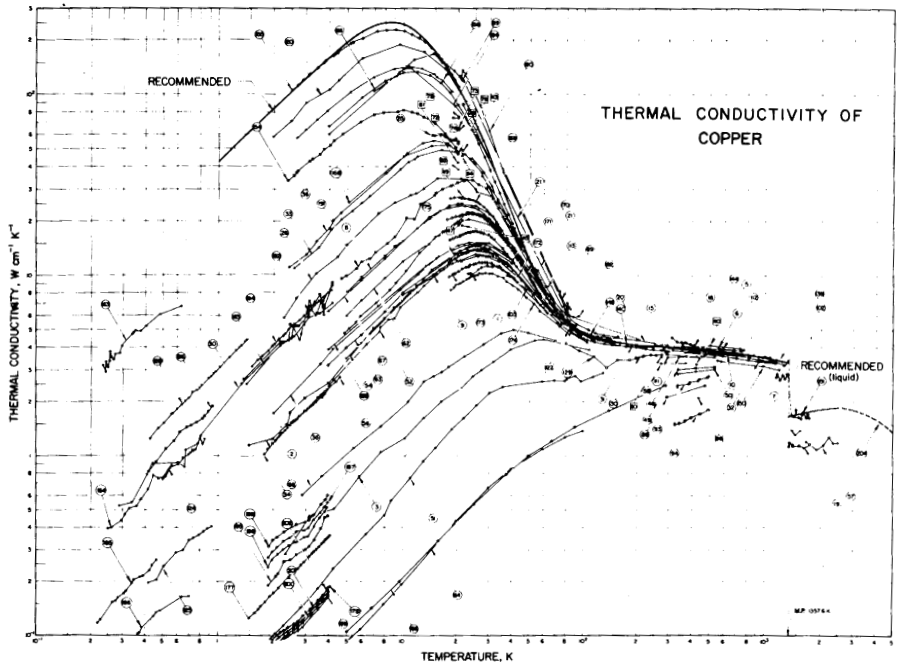
As we previously noted, fallibility and error are not confined to the subjective realm, and public misunderstanding of such errors is often quite substantial (as evidenced by the controversy surrounding the population statis-

tics of the 1980 census). Furthermore, just as fallibility of measurement is not limited to the subjective domain, neither is it limited to survey measurements or social statistics. Recent reviewers (Hunter, 1977; Lide, 1981) have used the data presented in Figure 1.2 to demonstrate the variability that exists among measurements of such elementary physical phenomena as the thermal conductivity of copper. Commenting on this figure, Hunter (1977:2) noted that ". . . although each analyst measured a physical quality that did not vary with location or time, it is clear that a remarkable variability attended the measurements." Moreover, he concluded (p. 2):

> The variation in attempting to evaluate the same physical constant is obvious. This example is not unusual. Similar plots of thermal conductivity as a function of temperature for approximately 400 common metals and material can be found in a supplement to the *Journal* (Ho et al., 1974). Nor is the observed variation in the measurement of "thermal conductivity" unique among physical parameters. . . .

FIGURE 1.2

Display of all reported measurements of the thermal conductivity of copper as a function of temperature



NOTE: Each set of connected points corresponds to the data given in a single literature source. The recommended curve is based on an evaluation by Ho et al. (1974).
SOURCE: Lide (1981). Reprinted by permission.

Not only can physical measurements vary wildly, but even well-publicized "discoveries" in the physical sciences have sometimes been shown to be experimental artifacts. For example, between 1963 and 1974 more than 500 articles in journals (including *Science* and *Nature*) discussed a supposed new substance: anomalous water or polywater. Although it resembled ordinary water, polywater was alleged to have a greater density, a reduced freezing point, and an elevated boiling point, among other anomalous properties. In the end, however, it was discovered that this "new substance" was nothing more than an impure solution of ordinary water (see Franks, 1981; Eisenberg, 1981).

Even measurements of geological events thought to be important precursors of major earthquakes have sometimes been called into question. For example, Jackson, Lee, and Liu (1980) recently published data that suggest that improper calibration of measuring devices accounts for the disturbing reports that there was a significant uplift of a portion of southern California between 1959 and 1974 and an apparent sinking between 1974 and 1977.[10] Common biological measurements have also shown similar fallibility.[11]

While measurement problems abound in the nonsocial and nonsubjective realms, there are also some noteworthy institutional safeguards that have been developed to deal with such measurement fallibility. For example, explicit programs of interlaboratory research and coordination of measurements have been developed by analytical chemists (see Youden and Steiner, 1975). Measurements are made with standard samples that should, in principle, always yield the same result. The variability of measurements across repeated trials made under "identical" conditions (that is, in the same laboratory, by the same scientist, using the same measurement procedure), those made by different scientists, those produced in different laboratories, and those produced by different analytical procedures allow for a better understanding of the error structure of such measurements. Not only do such procedures permit a more realistic appraisal of the significance of interlaboratory comparisons of measurements but they also permit the identification of "rugged" analytical procedures that produce low variability across laboratories, trials, and experimenters.

The systematic application of experimental procedures in a coordinated interlaboratory program of measurements provides one important paradigm that might be (but seldom has been) applied in survey research (see Chapter 5). The physical and chemical sciences provide a further example in their compilations of extant measurements, for example, the *Journal of Physical and Chemical Reference Data*, and in the role played in this work by the National Bureau of Standards (NBS), which operates programs intended to ensure the usefulness of the International System of Units (SI), such as those pertaining to length (the meter), mass (kilogram), and time

(second), as well as the many other measures derived from the SI units. Its programs provide the decentralized laboratories of federal, state, local, and private organizations with technical and operational guides to measurement procedures and tolerances (see Hunter, 1980). Since these measurements are often used in law, government regulations, and private contracts, the existence of an official reference bureau for such measurements is a practical necessity.[12] Since survey research does not presently have similar institutional mechanisms, we subsequently recommend a variety of institutional safeguards that may provide survey research with analogs to the quality control procedures developed in other areas of scientific measurement.

ALTERNATIVES TO SURVEYS

Insofar as subjective phenomena are important to public opinion and public policy, people will find some way to ask and answer the treacherous questions that they often entail. As a result, any suggestion of doing without surveys must consider what would come in their stead. Political commentators are one alternative. Rather than engaging in any sort of systematic sampling, they try to determine how "the people" view life, what people would answer if they were asked directly. Although the answers such commentators provide may reflect a sensitive understanding of the issues involved, these answers are inevitably colored by personal and political prejudices.

Another alternative for doing without surveys is doing without verbal questions. Many economists argue, for example, that one should look only at behavior. In this view, life conducts its own surveys, presenting an unending series of problems that people resolve in actions that can be interpreted as revealing their underlying preferences or values. Although based on an appealing argument, this position neglects the ways in which life poses questions in incomplete and even misleading forms. Some people argue, for example, that the aim of advertising and marketing is to induce consumers to assess their own preferences in a biased manner. When information is withheld or misrepresented, the interpretation of people's "economic behavior" is complicated further. Even when the market formulates its questions and presents its options in a balanced and forthright manner, the resultant behavior reveals at best people's underlying values only to the extent that one subscribes to the economic model of human rationality. This model involves controversial prescriptive assumptions, incompletely substantiated descriptive claims, and numerous ad hoc assumptions, regarding, for example, people's beliefs about one another's bargaining power, their sophistication regarding governmental monetary policy, or the implicit constraints they place on the set of available options (see, for example, Meeks's

review of economic theories of utility in Chapter 2, Volume 2). Finally, analysis of observed behavior can provide only a tenuous guide to attitudes that have yet to be tested by (and expressed in) real-life decisions.

Nonetheless, both of these alternative strategies have important contributions to make to the understanding of subjective phenomena. Commentators can, for example, offer creative hypotheses as to what bothers people, free from survey researchers' penchant for restricting discussion to issues for which data are available. The observer of behavior can discern something about how important various values are when the time comes for action, as opposed to how important those values appear to be in the ruminations of people who have not yet been forced to act on what they say they want. Yet insofar as one wants to know what people think and believe about their world and themselves, there is no substitute for asking them directly.

The knowledge obtained by asking is both unique and necessary, making it a vital complement to the insights provided by other procedures. And, therefore, improvement in the practice and understanding of the survey process is desirable. It is to this end that we have tried to summarize what is known, describe what is possible, and indicate, as best we can, those directions that may prove fruitful in the future.

1.5 Organization of the Volume

A scientific panel that is charged with the task of reviewing the current theory and practice in an applied field is inevitably drawn to concentrate on the frontiers of practice, to review work that is exemplary, and to make suggestions about what the future ought to hold. In most of this volume we approach our charge from this perspective. It is also valuable, however, to attempt to assess the actual scope of the field and to examine both typical current practice and those aspects of current practice that must be considered as negative examples or abuses. This is the task of the following two chapters.

We begin by describing the development and contemporary scope of the survey enterprise (Chapter 2) and by reviewing some major attempts to set standards for the practice and reporting of survey research (Chapter 3). In Chapters 4 through 6 we review fundamentals of measurement and examine their application to the survey measurement of subjective phenomena. In Chapters 7 through 9, the social and psychological aspects of the survey process are reviewed. Finally, in Chapter 10, we present the panel's recommendations for upgrading survey practice and advancing the state of the art of survey measurement and the scientific use of survey data.

In our second volume we present 18 empirical studies and theoretical essays on the survey measurement of subjective phenomena. These papers were commissioned by the panel in support of its work. (The titles of the papers are listed following the table of contents for this volume.) These papers cover in depth a range of issues that could only be summarily reviewed in the present volume. Throughout this volume readers will find numerous cross-references to the papers in Volume 2.

Notes

1. According to a memorandum from S. S. Wilks, the vice-chairman (October 9, 1945, archives of the National Academy of Sciences), the original members appointed in 1945 included: P.G. Agnew (American Statistical Association); E. Battey (Compton Advertising); H. Cantril (Princeton University); A. Crossley (A. Crossley, Inc.); W.E. Deming (Bureau of Budget); R. Elder (Lever Brothers); G. Gallup (American Institute of Public Opinion); P. Hauser (Department of Commerce); C. Hovland (Yale University); P. Lazarsfeld (Columbia University); R. Likert (Department of Agriculture); D. Lucas (New York University); E. Roper (Roper Organization); W. Shewhart (Bell Laboratories); F. Stanton (Columbia Broadcasting System); S. Stouffer (chairman of the committee, University of Chicago and War Department); C. Warwick (American Society for Testing of Materials); S. S. Wilks (vice-chairman of committee; Princeton University). At the time the committee was discharged (August 5, 1954), the roster of members showed one additional member, G.F. Hussey (American Standards Association).

A further list included with an annual report of 1945–46 activities lists Frederick Stephan as a member of the committee's staff. In a memorandum of January 8, 1948 to NAS President Detlev Bronk, Elbridge Sibley, executive secretary for the committee, noted the appointment of three subcommittees to advise directors of the research projects sponsored by the committee. Project directors and committee members were: (1) study of sampling, S.S. Wilks (chairman); F. Stephan (study director); W. Cochran; W.E. Deming; R. Franzen; M. Hansen; R. Jessen; P.F. Lazarsfeld; A. Politz; J.S. Stock; (2) study of use of panels, D.B. Lucas (chairman); P.F. Lazarsfeld (study director); S. Barton; W.E. Deming; A.R. Eckler; R.F. Elder; C. Hart; C. Hovland; T. Koopmans; F. Stephan; W. Wilbur; (3) study of interviewer effects, F. Stephan (chairman); C. Hart (study director); A. Crossley; W.E. Deming; G. Gallup; P.F. Lazarsfeld; R. Likert; E. Roper.

2. Memorandum from S.S. Wilks, October 2, 1945, p. 2; archives of the NAS.

3. Interest in this topic derives in part from the belief that financial support for science, external constraints on research, and the recruitment of young people into the sciences depend upon the public's perception of scientific activity (see National Science Board [1975:145] and subsequent discussion in the text).

4. We would note that responses for the "very happy" category are commonly reported for these measurements. One might intuit that the "not too happy" category would provide more reliable measurements because the semantic boundary between it and the other two categories seems clearer than that between "very happy" and "pretty happy." (This intuition is supported by the results of an analysis of response patterns obtained with this question; see Clogg, Chapter 11, Volume 2.) However, focusing on this category does not eliminate the discrepan-

cies between these survey measurements (see Chapter 5 for a comparison of responses to the complete set of categories).

5. Subsequent to the initial critique by Lang, a number of flaws in the sample design and execution have been noted (Lang, 1981).

6. When a probability sample of a population is drawn, most customary statistics (for example, means, proportions, etc.) have an estimable probability of deviating (by any fixed amount) from the value that would have been obtained if the statistics had been based upon the entire population rather than the sample. This variability in statistics based on samples is referred to as *sampling error*. In statistics, variability and bias arising from *other* sources is commonly referred to as *nonsampling error*. It has been suggested that the semantic treatment of all other types of error as a residual category (that is, nonsampling error) reflects the relatively underdeveloped state of thinking on that topic (Turner, 1980:43).

7. The General Social Survey (GSS) is a survey conducted regularly since 1972 by the National Opinion Research Center (NORC) under a grant from the National Science Foundation. It was designed to provide a national data resource for social scientists, and it includes a great variety of subjective measurements—most of which repeat measurements made in previous decades. It thereby provides the opportunity for the study of social changes in the United States.

8. What Campbell has in mind evidently is what Mead (1934:5) stated in discussing the relevance of a psychological point of view to the analysis of the social act:

> That which belongs (experientially) to the individual *qua* individual, and is accessible to him alone, is certainly included within the field of psychology, whatever else is or is not thus included. This is our best clue in attempting to isolate the field of psychology.
>
> The psychological datum is best defined, therefore, in terms of accessibility. That which is accessible, in the experience of the individual, only to the individual himself, is peculiarly psychological.

Campbell and Converse presumably would not be proposing to construct *social* indicators concerning "psychological states" unless the latter reflect or respond to social factors and conditions or have some kind of structuring based on the social relationships in which survey respondents (like all of us) are involved. So there is nothing to be gained by arguing the issue of social versus psychological phenomena. The subjective phenomena of interest are indeed psychological in Mead's sense. Whether they are social facts in Durkheim's terms has yet to be firmly established, but the demonstration of patterning of responses by the social characteristics of respondents or by the shifting of response distributions synchronously with documented variation in social conditions (as in the example from Rainwater) goes a long way toward providing the demonstration. Of course, working out causal relationships is just as great a challenge here as it is in, say, meteorology or cancer epidemiology, so that the social determination of subjective states is never axiomatic nor commonplace. In a careful study (see Stinchcombe, Jones, and Sheatsley, 1980b) the causal labyrinth may become intricate indeed.

9. In addition, at least some psychologists recognize the phenomena of repression or denial, and survey research has not yet offered much evidence of the efficacy of projective questions in bringing latent affect to the surface in the survey interview.

10. Since these changes in elevation would "if correct, have surprising and important implications," the dispute over these measurements is of great significance to geologists and others concerned about the future of Southern California (Jackson, Lee, and Liu, 1980).

11. Examples include data collected by the Center for Disease Control that show a large variation in the estimates made by different laboratories of the amount of lead present in identical samples of blood. For a sample of blood with a putative lead concentration of 41 μg/dl, 100 cooperating laboratories produced measurements that covered a range of 33 to 55 μg/dl; this result prompted the reviewer to observe (Hunter, 1980:870): "Clearly, whatever the true amount of lead in a sample, the variability demonstrated [in these measurements] guarantees numerous false alarms or perhaps more important, when the true level is high—nonalarms."

12. There are more than 5,700 voluntary standards, procedures, specifications, and codes for measurements of all sorts. These have been approved by the government for use in contracting and federal regulation (U.S. Office of Management and Budget circular 43 FR-48-51; see Hunter, 1980).

PART II

USES AND ABUSES OF SURVEYS

Once upon a time, runs an old Wendish legend, there was a peasant girl of Dehsa who lay on the grass at midday and slept. Her betrothed sat by her. . . . Then the noon-woman came and questioned him. Each time he answered her she put fresh questions to him. As the clock struck one his heart stopped beating. The noon-woman had questioned him to death.

ELIAS CANETTI, Crowds and Power

Primary responsibility for original drafting of this part of the panel's report was assumed collaboratively by DeMaio, Marsh, and Turner. Section 2.1 was contributed by Duncan. See Note on Authorship, page vi.

2

The Development and Contemporary Use of Subjective Surveys

2.1 The Development of the Survey Method

At least two rather distinct origins of modern polls or surveys can be discerned. The term "survey" was used in the nineteenth and early twentieth centuries for any kind of (supposedly impartial) "first-hand investigation" of a community or group. "Social Surveys" were given this broad definition in an article in the 1934 edition of the *Encyclopaedia of the Social Sciences*, which refers to Le Play's family budget studies, Booth's monumental survey of poverty and working-class life in London, and subsequent studies in this vein. That encyclopedia also includes an article on "straw vote," defined as "an unofficial canvass of an electorate to determine the division of popular sentiment on public issues or on candidates for governmental office." The author, Claude E. Robinson, indicates that the "principle of sampling" has been employed in straw votes in the United States since the early 1900s.

Many of those straw votes were sponsored by leading newspapers and magazines; they served as an important kind of election news and, in some instances, as an adjunct to campaigns to build circulation.

One of the most famous straw votes was the *Literary Digest* poll, which dealt with issues, such as the soldier's bonus and prohibition, as well as with candidate preferences. The disastrous failure of this poll to forecast correctly the 1936 presidential election has sometimes been credited with the demise of the classic straw vote. In many textbooks, that failure is attributed to the use of telephone directories and automobile registration lists as the frame of the very large sample (more than 2 million ballots were returned of 10 million distributed); this was presumably fatal because of a shift in voting patterns that accentuated the correlation between candidate preference and socioeconomic level. Something of this sort presumably did occur, because the *Literary Digest* poll had a good record of forecasting the elections from 1916 to 1932. But the major difficulty was not a matter of the sampling frame; it was the reliance on volunteer responses. Parten pointed out that auxiliary surveys by the *Literary Digest* in selected communities, where voter registration lists served as the sample frame, showed the same sort of bias as the national results: "So even with a better sample source . . . the *Literary Digest* poll would have failed because of the unrepresentative character of mail respondents" (Parten, 1950:393).[1]

Concomitantly with the *Literary Digest* fiasco, several, then new, public opinion polling organizations produced election forecasts that correctly named the winner. Their samples were much smaller than that of the *Literary Digest*, a few thousand cases, but supposedly selected "scientifically." "The interviewers were each given an assigned quota specifying the number of men, women, and persons of various economic levels to be interviewed. These restrictions on the selection of respondents were believed to control to a very considerable extent the biases that might result if interviewers chose their respondents freely" (Mosteller et al., 1949). (Although such quota sampling is not esteemed by sampling experts in statistics, its introduction ironically roughly coincided with what Kruskal and Mosteller [1980:187] describe as "the great Neyman article of 1934, 'On the two different aspects of the representative method . . .'." That article has proved to be a watershed in the slow introduction of lucid probabilistic ideas into the domain of statistical surveys, a process still far from complete today.)

During the 1930s, there were also a great many federal surveys that, among other things, exemplified the state of the art of probability sampling and helped to develop it. In 1940, important parts of the census were taken or reported on a sample basis. Also in 1940 the Monthly Report on the Labor Force was initiated; it continues to this day. And in 1939 the Division of Program Surveys had been set up in the U. S. Department of Agriculture

(USDA) to ascertain farmers' attitudes about government programs and other such topics (although these latter surveys did not use probability sampling at the outset).

Academic survey research was institutionalized in the period following World War II. The postwar report on the wartime activities of the Research Branch of the Information and Education Division of the War Department (Stouffer et al., 1949–50) opened up fascinating methodological problems and exemplified the systematic exploitation of social psychological concepts in survey analysis. Meanwhile, Paul F. Lazarsfeld had founded the Bureau of Applied Social Research at Columbia University, following up his experience in bringing methods hitherto used in market research into the academic orbit at the Office of Radio Research. These organizations did not have a national survey staff. But in 1946 the Survey Research Center, led by Rensis Likert, was established at the University of Michigan; in effect, the USDA Division of Program Surveys was just transferred to that campus. And under the direction of Clyde Hart in 1947, the National Opinion Research Center (founded in 1941) became affiliated with the University of Chicago. Many other university-based survey laboratories have subsequently been developed, as well as unaffiliated nonprofit survey centers combining features of university-based organizations and commercial polls and market research agencies. In addition, several major survey research organizations in the federal government were established or expanded, for example, National Center for Health Statistics, and, of course, the Bureau of the Census. (For an account of the interplay of academic attitude research and polling, see Converse, Chapter 1, Volume 2; for a detailed history, see Converse, 1981.)

A crisis in the development of the public opinion polls followed their failure to forecast the winner of the 1948 presidential election. But, in contrast to the *Literary Digest* episode of a dozen years earlier, the failure was not so dramatic because the election itself was so close. A thorough review of the sources of the failure (including but not limited to the methods of sampling) led to a series of recommendations by an expert committee of the Social Science Research Council. One-third of a century later, that committee's recommendations remain pertinent because the problems of polls and surveys are generic and perennial; election forecasting is just one exercise in which those problems emerge clearly.

We present here the major recommendations of that committee (Mosteller et al., 1949:292–293):

1. To improve the accuracy of polls, increased use should be made of the better techniques now available, particularly in sampling and interviewing. Since the reduction of any part of the error greatly increases the chances of a successful

forecast, the committee urges that pollers exert every effort to adopt more reliable techniques.

2. Increased attention should be paid to the development of research on each step of the polling operation to attempt to improve methods used in opinion research. This would include research on sampling methods, interviewer bias, concealment of opinions, selection and training of interviewers, etc. Experimental studies of such problems should be planned well in advance of elections. Many of these experiments can be incorporated into regular survey operations and as a part of more general cooperative studies of specific communities.
3. Research should be expanded on the basic sciences, particularly social psychology and political science, which underlie the analysis of voters' behavior. Even if perfect sampling of individuals is employed, we now know too little about voting intentions, factors affecting change in opinion, prestige effects, and similar topics to predict who will translate his opinion into actual voting.
4. In view of the increasing amount of emphasis being placed on public opinion polls, the Committee considers it very important for the public to be effectively informed about the limitations of poll results so it can interpret them intelligently. It urges that polling organizations, newspapers and magazines, and social scientists who work with poll results help provide the public with more information about polls, their interpretations and limitations.
5. There should be more effective cooperation between research workers interested in opinion measurement on common problems of methodology, underlying theory, and research design, including more studies of validity and reliability than have been made heretofore. Such cooperation would benefit all parties through the wider dissemination of their findings, and more efficient use of research resources and opportunities for experimentation.
6. More extensive training facilities and opportunities for practical experience under effective supervision should be provided for students who may be interested in research careers in this field, which includes political behavior, psychological research, and opinion measurement and statistical methods.

2.2 The Survey Enterprise

THE SCOPE OF THE ENTERPRISE

The nineteenth century budget studies and community surveys were largely concerned with the "facts" of income, expenditure, patterns of working life, housing conditions, and the like. The straw votes were preoccupied with ascertaining candidate choices, a subjective phenomenon, to be sure, but a rather specialized one. However, as noted before, some of these early polls ventured into what is now called public opinion research. The polls established in the 1930s quickly expanded their domain from election fore-

casts and current public issues. In 1936 the Gallup Poll was asking questions such as "What do you think is the ideal family size?" and "Do you think there will be another serious depression?" In the work of the Division of Program Surveys (Parten, 1950:21)

> Not only were efforts made to ascertain the types and incidences of various attitudes but the factors producing the attitudes and the effects of these attitudes on behavior also were probed. The extent to which people were informed about various aspects of federal programs was determined because of the relationship between attitude and knowledge or understanding of the purposes of the programs.

Thus, polls and surveys, in the modern sense of those terms, have always given some attention to subjective phenomena, and in many cases such phenomena were of primary interest. And so a major academic survey organization recently found it possible to issue a large volume recording trends in social attitudes for 1947–1978 (Converse et al., 1980). And responses to a number of individual attitude questions are available at irregular intervals from the 1930s to the present time (National Opinion Research Center, 1978, Appendix M; and also Southwick, n.d.). This means that the art of survey research today is based on considerable experience with such subject matter.

The remarkable range of topics on which subjective survey data are routinely gathered today is reflected both in government publications (for example, Executive Office of the President, 1973; U.S. Department of Commerce, 1977; Bureau of the Census, 1980a; U. K. Central Statistical Office, 1970–1981) and in popular compilations of survey data (Tarshis, 1979; Noelle and Neumann, 1967) on the views and characteristics of the "average" person. The dust jacket of one such publication (Tarshis, 1979) indicates the scope of the topics covered; it invites readers to learn

> . . . what the latest surveys, polls and lifestyle studies tell us about the not very average American people . . . How happy we are. How honest we are. How often we have sex. Our musical tastes. How often we go to McDonald's. How much we drink. How much television our kids watch. How many of us commit adultery.

The rapid expansion of survey research evidenced by such publications has been facilitated by the increased literacy and numeracy of the population, by the ease of transportation that allows interviewers to get to the homes of respondents, by the almost universal access to the telephone that makes possible less costly forms of interviewing, and by the increased power and decreased costs of automated information processing.

SIZE OF THE INDUSTRY

While it is evident to those of us working in survey research that the enterprise has expanded in the last decades, statistical evidence to support these perceptions is meager. Although some fragmentary evidence is available on the revenues of specific survey organizations (see Appendix B, Table B.1), there is no reliable information that would allow one to make comprehensive estimates of the past or present size of the survey industry. Conjectural estimates have placed the industry's gross revenues at around $4 billion in 1978 (see Appendix B for further details). Needless to say, there is no way of deriving separate estimates for surveys that make subjective measurements.

Two sources provide some information that allows us to infer the size of the enterprise as a whole. One survey, conducted under the auspices of the Committee on National Statistics (National Research Council, 1979), provides estimates of how frequently American adults *report* being surveyed—by telephone, in person, and by mail. Using these data, we estimate that a minimum of 100 million survey interviews were conducted between 1971 and 1976 in the United States. The majority of these surveys involved subjective measurements, for example, of attitudes toward a product or political views. Another survey has estimated that more than 28 million survey interviews were conducted by telephone during 1980 (Walker Research, cited in *Marketing News*, 1981). In Britain it has been estimated that 39 percent of the British public have been surveyed (Bowen, 1979).[2]

As we will note in Section 2.3, a veritable avalanche of stories reporting the results of such polls and surveys is published by the news media. During a single one-month period, we documented the distribution of more than 200 million copies of poll stories in America and more than 50 million copies in Britain.[3] (These include *only* stories we retrieved in our search of the media; we estimate the true totals to be 6 to 10 times as great.)

An increase in the size of the survey industry is suggested by the evidence that is reviewed in Appendix B. First, the number of organizations conducting surveys has increased, largely because of the establishment of in-house survey units by television stations and newspapers. Second, new techniques for surveying have made it possible to do more surveys and to provide almost instantaneous survey measurements. One particularly important technological innovation is computer-assisted telephone interviewing (CATI). A growing number of private and academic survey organizations use such systems, and it is likely that CATI will be adopted by some government survey programs in coming years.[4] The availability of CATI technology is likely, we suspect, not only to decrease the cost and time required to conduct polls and surveys and hence to increase the number of polls, but also to increase the use of instant polls to evaluate the public's views on contemporary issues.

2.3 The Use of Subjective Survey Measurements

In this section, we discuss in varying degrees of detail what we have been able to discover about the current use of subjective surveys. In presenting this material, we aim to provide readers with a more detailed and comprehensive picture of the contemporary survey enterprise. The material that results from our own research on the use of subjective survey data is presented in greatest detail, although even in this case we have relegated technical discussions of data gathering methods to appendices. To organize our discussion we have divided our domain into three broad sectors: government, academic, and commercial. We have also included two additional types of use that do not conveniently fit within these sectors: the use of surveys in legal and quasi-legal proceedings and the use of surveys in journalism.

GOVERNMENT USES

Surveys sponsored by the U.S. government[5] are conducted by the Bureau of the Census (together with other statistical agencies such as the National Center for Health Statistics) and by private organizations working under contract to various federal agencies. A parallel division occurs in Great Britain, although there is a greater dependency on the private sector, which conducts the fieldwork in some official British surveys.

The government of the United States has made quasi-factual survey measurements for a long time.[6] However, the U.S. government has not been a consistent sponsor of explicitly subjective measurements. The aversion of the Census Bureau (in recent times) to measuring inherently subjective phenomena is reflected in their protracted resistance to asking women how many children they expect to have. This resistance persisted for many years even though such measurements were thought to be potentially useful in making population projections (Goldberg, Sharp, and Freedman, 1959). It was not until 1967 that the Census Bureau agreed to collect such data, and four years elapsed between the first such measurement and the inauguration of the annual Current Population Survey supplement on birth expectations (see Bureau of the Census, 1975:5).

Statements by Census Bureau representatives make explicit the reluctance of the Bureau to delve into the subjective domain, although by the late 1960s this resistance was often proving unsuccessful. Addressing a conference organized by the Social Science Research Council, Daniel Levine described the Census Bureau's reluctant acquiescence to growing demands for the inclusion of subjective measurements in its surveys; he noted (Levine, 1968:7):

> With respect to social and psychological questions dealing with attitudes, expectations and/or opinions, the Bureau's position in surveys in which the Bureau itself is the prime sponsor has been one of limiting their use.

By 1968, however, there were a growing number of instances in which the Census Bureau had consented to ask questions about subjective phenomena. Indeed, the Bureau had even approved a survey asking respondents about their personal happiness.[7]

In the years since 1968 subjective measurements have come to play a substantial role in government survey research. In fiscal year 1979, more than one million survey interviews conducted by the Demographic Surveys Division of the Bureau of the Census incorporated some measurements of *explicitly* subjective phenomena; total costs for these surveys exceeded $50 million (see Appendix B). These surveys included such questions as:

> How would you rate this house (building) as a place to live — would you say it is excellent, good, fair or poor? [Annual Housing Survey, national sample, 1979, Question 110]
>
> Which of these four statements best describes your (family's) ability to get along on your (its) income?
>
> –I (we) always have money left over
> –I (we) have enough with a little extra sometimes
> –I (we) have just enough, no more
> –I (we) can't make ends meet
>
> [National Longitudinal Survey: mature women sample, 1979, Question 30]

In addition to the subjective measurements included in sample surveys, several subjective and quasi-factual items were also included in the 1980 census of the population, for example: "How well does this person speak English? (Very well, well, not well, not at all)" (Question 12c).[8]

Besides the surveys conducted by the Census Bureau, the government sponsors a wide range of surveys that are carried out under contract by private survey organizations. We attempted to estimate the full extent of the government survey enterprise by examining the Office of Management and Budget (OMB) and General Accounting Office (GAO) dockets for clearance of government sponsored surveys.[9] We found 228 active surveys on the OMB docket and one survey on the GAO docket.

Following review of the questionnaires in the OMB files, we classified the surveys into three categories:[10] (1) subjective—surveys containing one or more measurements of subjective phenomena; (2) quasi-factual—surveys containing one or more quasi-factual (and *no* subjective) measurements; and (3) factual—surveys containing no subjective or quasi-factual measurements. The definition of subjective used in this classification is that given in Chapter 1: subjective phenomena are those that, *in principle*, can be

directly known, if at all, only by persons themselves. Factual measurements are in principle verifiable without reference to respondents' interpretations. Quasi-factual measurements allow latitude for the respondent's definition of the criterion for the (factual) behavior or event in question: unemployment, crime victimization, ethnicity. (See Bailar and Rothwell, Chapter 5; Newman, Chapter 6; and Smith, Chapter 4 in Volume 2, for further discussions.) In considering our results it should be borne in mind that surveys coded in the subjective category usually included quasi-factual and factual measurements as well; surveys coded in the quasi-factual category generally contained only a few such measurements with the remainder being strictly factual.

Table 2.1 presents statistics on the total and average numbers of interviews and respondent hours involved in each category of government survey. Contrary to our expectations, surveys that included subjective measurements greatly outnumbered surveys of the other two types (145, compared with 49 factual and 8 quasi-factual). However, the number of surveys, by itself, is a misleading indicator of the predominance of surveys including subjective measures since such surveys are generally conducted on a much smaller scale than either factual or quasi-factual surveys. The average number of interviews per year[11] in subjective surveys (7,786) was small compared to factual surveys (41,700) and minute compared to quasi-factual (268,245) ones. Moreover, direct government involvement in subjective measurement was often reduced by contracting surveys out to private research firms: 70 percent of the surveys containing some subjective measures were conducted by private contractors, compared with 47 percent and 38 percent, respectively, for factual and quasi-factual surveys.

Although surveys including subjective questions are characterized by a smaller average number of interviews, the interviews themselves are generally much longer than the interviews in other kinds of surveys. Interviews lasted an average of almost half an hour, compared with less than 10 and 13 minutes for quasi-factual and factual interviews.[12] Thus, overall, the total amount of time required for federal surveys involving subjective measurements is greater than that for quasi-factual or factual surveys (507 compared with 337 and 415 thousand hours, respectively).

The purposes for which different types of government surveys are taken vary. The surveys included in the OMB docket included almost equal numbers of surveys used in evaluating government programs ($n = 91$) and those identified as generating "general purpose statistics" ($n = 90$); a smaller number were conducted for research purposes.[13] Examples of each kind of survey, as described in their OMB abstracts, follow:

> *General Purpose Statistics*: 1981 Health Interview Survey (OMB-0937-0021). "A national sample survey . . . providing statistical information for health re-

TABLE 2.1

Indicators of the Magnitude of Government Survey Activity and the Characteristics of Government Surveys

Type of Survey	Number of Surveys	Total Interviews[a] (per year)	Total Hours of Interviewing (per year)	Average Number[a] of Interviews per Survey	Average Number[a] of Hours per Survey	Average Number[a] of Minutes per Interview	Proportion of Surveys Conducted by Private Contractor
Factual	49	2,001,595	415,385	41,700	8,654	12.6	.47
Quasi-factual	8	2,145,960	337,930	268,245	42,241	9.6	.38
Subjective	145	1,097,803	507,981	7,786	3,603	27.6	.70
All surveys	202	5,245,358	1,261,296	26,626	6,403	14.4	.63

NOTE: Tabulation includes government surveys with active OMB clearance as of November 17, 1980; figures summarize estimates provided to OMB for each survey. (These figures exclude forms related to the 1980 Decennial Census, for which estimates are extremely large and not representative of other surveys; there were 13 such census forms, which involved an estimated 181,867,850 interviews and 31,196,874 total hours of respondents' time.)

[a]Estimates of hours and interviews were not available for 1 factual survey and 4 subjective surveys.

searchers, planners and policymakers on the amount, distribution and effects of illness and disability and the services rendered for or because of such conditions. Annual supplements address topics such as smoking habits, often to meet the needs of other agencies." (Number of interviews per year: 40,000; Sponsor: Department of Health and Human Services, Office of Assistant Secretary for Health.)

Research: Household Survey for Monroe County Georgia (OMB-1901-0163). Survey "to interview residents of Monroe County, Georgia, site of . . . a coal-fired power plant, to determine their characteristics, opinions concerning the area, the quality of public and private services, and their views on the likely impact of [the power plant]." (Number of interviews per year: 345; Sponsor: Department of Energy.)

Program Evaluation: Survey of Indochinese Resettlement (Operational Feedback —6th and 7th Waves) (OMB-0960-0163). Survey "under Public Law 94–23, Indochina Migration and Refugee Assistance Act of 1975, to determine how well the Indochinese refugees are adjusting to life in the U.S. and [the] types of private and government services they are receiving or may require. This survey is used to solicit this information for program planning, developing policies, and implementing . . . resettlement." (Number of interviews per year: 2,400; Sponsor: Social Security Administration.)

We found that the range of topics covered in government-sponsored surveys, as categorized by OMB, was quite broad. There were, however, a few subject areas in which a majority of surveys were concentrated. For instance, 50 of the surveys were on the general topic of "health"; 74 percent of them included subjective measurements. Three other topics were each the focus of more than a dozen surveys; for two of them ("community development" and "public assistance and other income supplements") there was a preponderance of surveys which included subjective measurements. Data on the surveys by subject areas are presented in Table 2.2. In general, surveys involving subjective measurements covered somewhat different topic areas than factual and quasi-factual surveys. For example, while most of the surveys on ground transportation and commerce were wholly factual, most of the surveys on training and employment, community development, health, public assistance and welfare, and basic science research included subjective measurements.

Surveying Foreign Populations. The U.S. government, particularly through the U.S. Information Agency (USIA), conducts a large number of subjective surveys in foreign lands.[14] These surveys are designed both to measure audience reaction to American broadcasts and also to provide the U.S. government with information on foreign sentiments toward U.S. policies. So, for example, Adler and Wertman's (1981) review of European public opinion toward the North Atlantic Treaty Organization (NATO) cites

TABLE 2.2

Subjects Investigated by Different Types of Government Surveys (percentage distribution)

OMB Code for Subject Matter of Survey	Type of Survey				
	Factual	Quasi-Factual	Subjective	Total	(N)[a]
Health	26%	0%	74%	100%	(50)
Advancement and Regulation of Commerce	56	0	44	100	(18)
Community Development	0	0	100	100	(14)
Public Assistance and Other Income Supplements	7	7	86	100	(14)
Training and Employment	22	0	78	100	(9)
General Science and Basic Research	0	12	88	100	(8)
Ground Transportation	62	0	38	100	(8)
Other (30 categories each having a frequency of 7 or less)	22	8	70	100	(80)

NOTE: Tabulation includes government surveys with active OMB clearance as of November 17, 1980; figures summarize estimates provided to OMB for each survey. (These figures exclude forms related to the 1980 Decennial Census; there were 13 such census forms, which involved an estimated 181,867,850 interviews and 31,196,874 total hours of respondents' time.)

[a]N: Number of surveys. (Note that subject matter code for one survey was not available.)

USIA-sponsored surveys conducted in five European nations (Belgium, Great Britain, Italy, Netherlands, and West Germany) between October 1980 and March 1981. Wertman estimates (personal communication, October 6, 1981) that the Office of Research of USIA conducted between 40 and 50 foreign surveys during 1980 and that it employed approximately 25 professionals in the design, analysis, and interpretation of public opinion data.

Surveys of foreign populations can raise serious questions of propriety. For example, during the Vietnam War, the National Academy of Sciences accepted a congressionally mandated (P.L. 91–441) contract from the U.S. Department of Defense for a study to assess the effects of herbicides on the land and people of South Vietnam; as part of that study surveys and other social research were conducted in Vietnam. The results were published in

a monograph: *The Effects of Herbicides on South Vietnam, Part B: Beliefs, Attitudes and Behavior of Lowland Vietnamese.* The authors of that report introduced their study by noting (Murphy et al., 1974:I–5):

> Bombing, intravillage fighting, being overrun by the National Liberation Front or by the government of South Vietnam forces, and economically devastating effects of the war often dominate people's lives. Thus, feelings from many sources can become channeled on herbicides, and conversely feelings originally generated by herbicides can masquerade as feelings about some other topic.

The study was the subject of great controversy in the social science community, and the role played by the Academy was strongly criticized (see, for example, Boffey, 1975:Chapter 7). The participation of social scientists in this Department of Defense funded project was questioned at the outset by the American Anthropological Association (1971:40–41), whose members feared the uses to which such data might be put and the potential dangers posed to respondents (see Issacs, 1971; Martin, 1971).

Government surveying of the attitudes and opinions of foreign populations is not a uniquely American activity. For example, the German Armed Forces Institute for Social Research (Sozialwissenschaftliches Institut der Bundeswehr) has conducted surveys to assess the public's fear of war, their regard for the military, their willingness to pay additional taxes (or cut social spending) to "provide financial resources for the maintenance of the fighting power of the U.S. Armed forces," and so forth (Schönborn, 1981). In the fall of 1979, the Institute conducted simultaneous surveys on these topics in both Germany and the United States.

Should Governments Collect Such Data? Serious concerns have been expressed about the appropriateness of government agencies collecting subjective data from their own (not to mention foreign) citizens. Such concern about state interference in the private affairs of citizens is not unique to the twentieth century. In the eighteenth and nineteenth centuries, battles were waged in Europe (often against the authority of the church) to obtain civil registration of such objective phenomena as births, marriages, and deaths. Some government projects of that period that attempted to monitor subjective aspects of the social world caused such a public furor that they have never been repeated (for example, the British census of religious belief and attendance of 1851). Indeed, surveys of morale of the kind conducted by government survey organizations in America during World War II (Stouffer et al., 1949-50) were attempted by the British government, but the surveys had to be abandoned following public outcries against the "Cooper-snoopers" (Cooper was Minister of War at the time).

The policies of governments concerning the propriety of such official sur-

veys are varied. Some Western governments are quite sensitive about this question. Opinion polling in socialist states has a very mixed history; active programs have proliferated at times in Eastern Europe, and at other times they have been suppressed (see, Connor and Gitelman, 1977; Kramer, 1981).[15]

One way in which some Western governments may have tried to sidestep the political controversy engendered by subjective surveys has been through support for programs conducted under the auspices of international organizations. The Social Indicators Program of the Organization for Economic Cooperation and Development (OECD) is one such example.[16] However, even in this forum serious objections have been raised. At a September 1979 meeting of an OECD coordinating group, the Swedish representative argued that (Erikson, 1980:3–4):

> The purpose of a social report is . . . to describe the state and change of social conditions in a nation as correctly as possible. It should make it possible for the citizens to evaluate the social development. A given development could be valued differently by different persons, depending on their value orientations. Some might (for example) especially look to economic change while others emphasize changes in health situation. It is for those making judgments of less interest how others evaluate the development; it is the outcome in terms of their own value orientation that is crucial. Evaluative indicators thus have no place in a social report. . . .
>
> [The public's] evaluations and values should enter the political process not by their being asked as questions in surveys and opinion polls, but through partaking in the political discussion.

In accord with the Swedish position, the OECD social indicators program is not expected to include subjective measurements.

Besides this aborted OECD initiative, the other major examples of international progams of subjective survey measurement are the various collaborative programs of the European Economic Community (EEC).[17] The United Nations report, *Toward a System of Social and Demographic Statistics*, evidences some international support for such programs of subjective measurement; it observes that (United Nations, 1975:32):

> In dealing with social questions, however, we may also be interested in subjective information relating to how much people in general know about an issue, how much importance they attach to it and what kind of solution they think would be desirable.
>
> Public opinion surveys provide a means of obtaining some light on such matters. It would be interesting to know, for instance, what issues are commonly regarded as major problems and how the ranking of these issues changes with time.

Our panel considered the appropriateness of government measurements of subjective phenomena, but it could not reach a consensus on this issue. Appendix A presents two positions on the question.

ACADEMIC USES

Basic research, the quest for solutions to theoretical (as opposed to practical or applied) problems, is generally the bailiwick of academic institutions. Surveys are one vehicle for basic research in several of the social sciences, particularly sociology, political science, psychology, and economics; Glock (1967) has provided a detailed review of survey use in these disciplines.

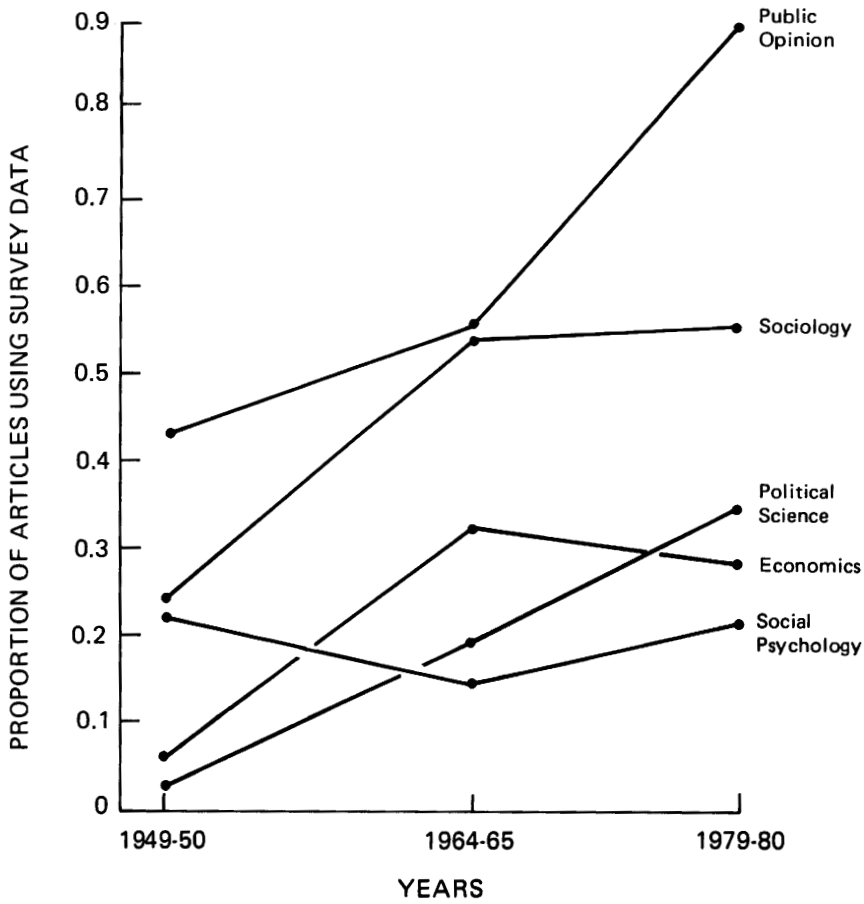
Trends in Use. Content analyses of academic journals attest to the increasing importance of surveys as sources of data for basic research. Alwin and Stephens (1979), for example, found increased use of survey research data over the last three decades in the major American journals of sociology. As part of its work, our panel commissioned a review of core research journals in political science, social psychology, economics, sociology, and public opinion research (see Presser, Chapter 3, Volume 2). This review found a substantial increase over the last three decades in the use of survey data in all disciplines except social psychology (see Figure 2.1). Subjective phenomena (attitudes, in particular) were found to be frequent topics of survey investigation.

While it is widely agreed that universities provide a suitable intellectual setting for theoretical inquiries, they may not always provide an optimal setting for undertaking survey research with its requirements for institutional continuity, a large staff, heavy managerial commitments, and an assurance of stable funding. Thus, there has been a tendency in many countries for survey research programs to be affiliated with semi-independent institutes (attached with varying degrees of strength to a university).

Academic researchers contemplating a survey research project have a choice of several alternatives for arranging their data collection. One possibility is to subcontract the data collection using the interviewers, the coders, and sometimes the executive expertise of a commercial market research or opinion polling company. This type of arrangement is not always an easy one (see Payne, 1979; Abrams, 1980; and Harrop, 1980, for different views on such arrangements), and is more common outside the United States.[18] Another, more prevalent alternative is to enlist the help of an institute specifically designed to conduct survey research. Examples of these are the large and well-established organizations like the National Opinion Research Center (NORC) at the University of Chicago and the Survey Research Center (SRC) at the University of Michigan. (Revenue and employment information for the largest of these institutes is included in

FIGURE 2.1

Proportion of articles in core journals which report survey data, 1948–80.



NOTE: To Table 2.3 for list of journals; from Presser, Chapter 3, Volume 2.

Appendix B). Another possibility is to formulate a problem that is amenable to investigation using archived data.

In their review of sociology journals, Alwin and Stephens (1979) found a trend over the last three decades away from academic researchers organizing their own surveys toward their using specialized fieldwork agencies, omnibus surveys, and secondary analysis. The review we commissioned (see Presser, Chapter 3, Volume 2) reached a similar conclusion for both sociology and political science. Table 2.3 shows the sources of the survey data found in published research articles for 1949–50, 1964–65, and

TABLE 2.3

Sources of Survey Data in Articles Published by Core Journals

Source of Survey Data	Academic Fields[a]				
	Sociology	Political Science	Economics	Social Psychology	Public Opinion
			1949–50 and 1964–65		
Authors	46.5%	67.6%	6.9%	91.3%	48.5%
Governments	35.0	17.6	70.7	0.0	4.4
ISR	4.5	2.9	12.1	4.3	2.9
Other	14.0	11.8	10.3	4.3	44.1
Total*	100%	100%	100%	100%	100%
(N)	(157)	(34)	(58)	(46)	(68)
			1979–80		
Authors	27.3%	38.5%	6.3%	92.0%	42.2%
Governments	34.3	3.1	64.6	0.0	2.2
ISR	7.0	40.0	12.7	4.0	17.8
Other	31.5	18.5	16.4	4.0	37.8
Total*	100%	100%	100%	100%	100%
(N)	(143)	(65)	(79)	(75)	(45)

SOURCE: Presser, Chapter 3, Volume 2.
NOTE: The periods 1949–50 and 1964–65 have been combined partly because the patterns are quite similar in both and partly because the number of cases in the earlier period is small.
[a]Journals surveyed were: *American Economic Review*, *Journal of Political Economy*, *Review of Economics and Statistics* (Economics); *American Journal of Sociology*, *American Sociological Review*, *Social Forces* (Sociology); *American Political Science Review*, *American Journal of Politics*, *Journal of Politics* (Political Science); *Journal of Personality and Social Psychology* (Social Psychology); *Public Opinion Quarterly* (Public Opinion).
*Some totals do not sum to exactly 100.0% due to rounding.

1979–80. This tabulation indicates that the sources of survey data in economics and social psychology were relatively constant over the period studied. Researchers in sociology and political science, however, have moved away from collecting their own data. In political science, this movement has involved a growing use of survey data produced by the Institute for Social Research (ISR) at the University of Michigan. In sociology, it has largely involved a movement toward data produced by the National Opinion Research Center (NORC) at the University of Chicago, by ISR, and, to an even greater extent, by other nongovernmental survey organizations.

A major factor underlying these trends has been the growth in size and accessibility of archives of survey data. The growing wealth of data available to contemporary social researchers, together with improved means to effi-

ciently manipulate it, has profoundly affected the practice of academic social research during the last two decades. Data banks such as those of the Inter-university Consortium for Political and Social Research, the Roper Public Opinion Research Center, and the Louis Harris Data Center, all in the United States, the Zentralarchiv für Empirische Sozialforschung at the University of Cologne in Germany, and the Social Science Research Council Survey Archive at the University of Essex in England have grown rapidly.

As a condition of funding, research sponsors may now require that survey data be deposited in these archives at the conclusion of a research project. This encourages wider use of expensively acquired survey data and is one form of dissemination of survey results. It also provides a research resource that permits verification of published studies and permits other researchers to undertake alternative analyses of the same data. Often, however, all the information necessary for responsible use of the data is not included in the archives. So, for example, the original researcher may be aware of limitations of the original data collection, but these limitations may fail to be adequately documented when the data are deposited at the archives. Thus, a secondary analyst can lack the information needed to make full disclosure of the relevant details of the data, and so the archived data will be poorly described when the secondary analyses are published. Documentation is not always complete even for the best of survey programs. The General Social Survey, for example, began in 1972 as a national data program for the social sciences, but did not begin routinely reporting the response (or refusal) rates for its surveys until 1978.

Nonetheless, archiving is a desirable practice that holds the promise of improving survey practice and encouraging more efficient use of survey data. Improving documentation for archived data is a necessary step to ensure their continuing value. (See Recommendation 17, Chapter 10, on maintaining, improving, and enlarging archives of survey and auxiliary data.)

Funding. Data on the funding of academic survey research are scarce. Alwin and Stephens (1979) report that, in sociology, federal and private foundations and universities are currently the major supporters of survey research (see Table 2.4). Unfortunately, we cannot use these data to infer a trend over time in funding patterns in sociology, and we have no suitable data for survey research conducted in other social science disciplines.

USE OF SURVEYS IN LEGAL PROCEEDINGS

The use of surveys in courtroom and quasi-legal settings does not represent a distinct category of use in terms of its sponsorship or the auspices of data collection. It does, however, represent a sufficiently distinctive and important use to merit separate note.

TABLE 2.4

Sources of Funding for Survey Research in Sociology

Source of Funding	Percentage of Authors Reporting Each Source of Support
Federal government	78
State government	29
Private foundation or university	68
Private commercial	3

SOURCE: Alwin and Stephens (1979: Table 2).
NOTE: Figures are from a review of articles using survey data (n = 40) published in 1975 in the *American Journal of Sociology* and the *American Sociological Review*. Percentages sum to more than 100 because some articles mentioned multiple sources of funding.

Court Uses. One of the earliest uses of public opinion surveys in a criminal trial occurred in 1952 in support of an attempt to change the venue of a trial in which a Southern black youth was charged with the rape of a young white woman.[19] The case stirred a swift and violent community reaction. According to a contemporary newspaper account, the report of the rape was followed by

> . . . mob violence in the small Lake County farming town of Mascotte, resulting in the burning of three Negro homes and the shooting up a small colony of Negro homes in nearby Stuckey's Still. The National Guard was dispatched to quell the rioting.[20]

The defendant's attorneys asked the court to move the trial to another jurisdiction arguing that community passions and prejudicial publicity jeopardized the defendant's right to a fair trial. With support from the National Association for the Advancement of Colored People (NAACP), the defense commissioned Elmo Roper to conduct an attitude survey. Roper interviewed 518 white persons in the county and reported that 43 percent believed that the defendant was definitely guilty of the crime. The court, however, denied the motion and dismissed the survey as hearsay evidence, that is, evidence reported by someone other than the original speaker.

This procedural objection, of course, applies to all survey research. Survey reports are by definition the second-hand or hearsay evidence of anonymous witnesses: rather than having witnesses testify under oath in court, a third person reports on what he has been told. In recent years, however, this objection to the use of survey data as evidence in court has been overruled in most courts (Collins, 1978:163), and survey research presently pro-

vides one of the prime sources of courtroom testimony as to community prejudice (see Bonora and Krauss, 1979:Chapter 2). Indeed, the 1968 American Bar Association Project on Standards for Criminal Justice has recommended that (American Bar Association, 1968:112–113):

> A motion for change of venue or continuance shall be granted whenever it is determined that because of the dissemination of potentially prejudicial material, *there is a reasonable likelihood that in the absence of such relief, a fair trial cannot be had.* This determination may be based on such evidence as qualified public opinion surveys or opinion testimony offered by individuals, or on the court's own evaluation of the nature, frequency and timing of the [publicity]. A showing of actual prejudice shall not be required. [emphasis in original]

In recent years, public opinion surveys have been a key part of attempts to change the venue of highly publicized criminal cases (see, for example, McConahay et al.'s [1977] description of work in the case of *North Carolina* v. *Joan Little*). They have also been used to assist attorneys in vetting prospective jurors (see Bonora and Krauss, 1979; Schulman et al., 1973; Zeisel and Diamond, 1976; Berk, Hennessey, and Swan, 1977; Marsh, 1979; Saks, 1976; Seltzer and Turner, 1981), and they have played a prominent role in judicial decision making concerning the impanelment of jurors in death penalty cases. In *Hovey* v. *Superior Court of Alameda County* (California Supreme Court, H-1440, decision of August 28, 1980), for example, it was argued that the exclusion from jury service of persons who were absolutely unwilling to impose the death penalty resulted in the selection of a jury from a pool of eligible jurors who were more conviction prone than the community at large. The social science evidence was presented in 17 court days of testimony by seven expert witnesses, and the transcript is over 1,200 pages long. The evidence included a number of attitude surveys which, the appellant claimed, demonstrated that "death-qualified" jurors represented a more conviction prone population which was also less likely to defer (at least verbally) to various constitutional and legal safeguards afforded to defendants in criminal trials (Ellsworth, Cowan and Thompson, 1979; Goldberg, 1971; Jurow, 1971; Wilson, 1964).

Opinion surveys are also frequently introduced as evidence in various types of civil litigation—particularly in cases involving claims of trademark infringement. In the latter cases the substantiation of a claim of public confusion between two brand names may turn upon the demonstration of such confusion in a survey sample (Collins, 1978).

Regulatory Actions. Attitude surveys have also served as both the subject of government regulatory actions and as a tool used by regulators, particularly the Federal Trade Commission (FTC).

Regulatory cases involving surveys have arisen from complaints of misleading advertising, such as: "9 out of 10 doctors recommend. . . ." For example, in 1980 the FTC ruled on a complaint that Litton Industries had made unfair and deceptive statements in a nationwide advertising campaign. Litton's advertisements (published, for example, in the, *Wall Street Journal*, October 25, 1976) claimed that "76% of the independent microwave oven service technicians surveyed recommend Litton."

The FTC found the Litton surveys on which the advertisements were based to be "fatally defective" because the population from which the survey took its sample was incomplete. Litton used as its sampling frame a compilation of agencies that serviced or sold Litton products. Based on the fact that a substantial number of Litton dealers were included and that "respondent's own experts have shown that the answers of a servicing dealer would be biased in favor of a brand which it sells" (Federal Trade Commission, 1980a:22), the Commission concluded that "[Litton] did not have a reasonable basis for advertising that their surveys revealed certain preferences for Litton among independent service agencies. . . ." (Federal Trade Commission, 1980a:25).

Survey research has been also been used by regulatory agencies to monitor the need for and impact of federal regulations. As of March 1, 1980, the FTC's impact evaluation group had 14 survey research projects in progress. One study attempted to assess "what assumptions consumers make about warranty protection when given varying degrees of information about the warranty coverage offered" (Federal Trade Commission, 1980b:2). Another study attempted "an analysis of consumer awareness, beliefs, and attitudes towards Listerine." This latter study sought to determine whether FTC-mandated corrective advertising had succeeded in dispelling the false notion (induced by previous advertising) that Listerine mouthwash prevents colds.

A particularly interesting example of the importance of survey research in the legal arena comes from Great Britain. Under the 1975 Employment Protections Act, the Advisory, Conciliation, and Arbitration Service (ACAS) is required to discover the opinions of employees of any establishment in which there is a dispute about whether a union should be recognized. This legal requirement that opinion be tested has brought with it the inevitable legal commentary on the methods and findings of survey research. For example, in one dispute between an insurance company and a union, the company disputed the results of an ACAS survey because the option of union representation was listed first on the ballot. At trial, the company presented a witness from a survey research organization who testified that the order of presentation of response alternatives can influence the outcome of a poll.

A high court judge, Mr. Justice Browne-Wilkinson, ruled that in this

instance order effects were implausible (Reeves, 1979; see Chapter 5 for a review of research on this topic). It is eloquent testimony to the growing social and political importance of subjective survey measurements that methodological disputes about the effects of question order are sometimes a subject for litigation.[21]

COMMERCIAL SURVEY RESEARCH

A large proportion of commercial survey research is market research of one kind or another. Market research holds a central place in the economies of most industrialized Western nations; product manufacturers collect systematic information about the public's attitudes, perceptions, and buying habits, both to guide production and to plan advertising and public relations campaigns. Not only do market researchers conduct a great number of surveys, but, indeed, many aspects of survey methodology as well as psychological and sociological theory have been advanced by market researchers: quota sampling, attitude scaling, and the theory of consumer behavior, to name a few.

One recent technological innovation produced by market researchers is a form of cable television (known as QUBE) that has been used on an experimental basis for conducting surveys. QUBE uses a cable television channel to allow subscribers (that is, respondents) to communicate with a central office. It thus overcomes a major disadvantage of telephone interviewing, namely, the inability to provide the respondent with visual stimuli. While this system overcomes a disadvantage of telephone surveys, it is not without its own special problems. Most important is the uniqueness of the self-selected population that is sampled in any survey using QUBE; only persons in households that subscribe to the service, who are watching that television channel at the time, and who choose to respond, are included in the sample. Despite this inherent bias, QUBE has expanded to several metropolitan areas, and QUBE surveys have been reported in several national news stories. One example is NBC's report of public reaction to President Carter's major address on energy and the state of the nation (see *Washington Post*, July 16, 1979:A16).

Although market research innovations influence the general practice of survey research, the vast majority of commercial market research surveys are never made public. One experienced market researcher who served on this panel conjectured that only about 1 of every 20 surveys that are conducted by commercial agencies is ever made public. The other 19 surveys are proprietary information collected by corporations to test consumers' preferences for products, their responses to advertising, their perceptions of corporations, their political opinions and behavior, and so forth.

There are many questions about survey practice in the commercial field

to which one would like systematic answers. For example, it would be interesting to know how subjective survey measures are typically used in the development, launching, and advertising of a new product, and it would be fascinating to evaluate the success of such subjective measurements when they are used as the basis for marketing decisions. One would also like to know more about the extent and uses of "corporate image research" conducted by companies that commission surveys to obtain information about the attitudes of various important publics. Unfortunately, even if it were feasible to draw a sample of such uses, it is doubtful that one could obtain sufficient details about the surveys to investigate the topic thoroughly. Market research studies are costly, and they are typically treated as confidential business information by their sponsors. Survey organizations are often forbidden by their contracts with sponsors from releasing such information (see note 3 in Chapter 3 for an extreme instance of such restraints on public disclosure).

While we do not have information on 95 percent of commercial survey research, we do have some information about the tip of the iceberg, the 1 poll in 20 that is published. During the summer of 1980, we conducted an exploratory study of the surveys and polls that were reported in major newspapers in both the United States and Great Britain, and this has allowed us to make some tentative statements about the commercial surveys that enter the public domain (see Appendix C for details of study). Before discussing the results of this study, it is appropriate to note that the polls reported in the media appear to be a special subset of commercial research. None of the five largest commercial survey agencies (as measured by gross revenue; see Appendix B, Table B.1) was mentioned at all in our study of newspaper reports of polls. It has been asserted that published polls such as the syndicated Gallup Poll and Harris Survey are not conducted for their own commercial value but rather for their advertising value in attracting more lucrative business (Wheeler, 1976; Dionne, 1980f). If so, it appears that the advertising incentive that journalistic polls offer is most attractive to small- and medium-sized companies. In any case, polls reported in the media are a somewhat special subset of the commercial survey world.

Based on our study, we estimate that more than 200 million copies of newspapers containing poll results reached the American public and 70 million copies reached the British public in a one month period.[22] Of course each of these poll reports need not have been read, but even so we were surprised by the sheer volume of poll reports to which the public is exposed.[23]

The newspaper stories we amassed in our files provide information as to the topics reported in polls that were deemed newsworthy by newspapers in Great Britain and the United States. For each poll reported in our file of newspaper stories we (multiply) coded the topic(s) about which questions

were asked. The distribution of topics is shown in Table 2.5. The first column of Table 2.5 gives the unweighted distribution; this can be thought of as an indicator of the proportion of polls which ask about a given topic (provided that the topic is considered newsworthy by *any* newspaper). In the second column the topics covered in a poll were weighted by how often each poll was cited and the circulation of the newspapers in which it was cited; as before, the results were summed across all polls. Each percentage in column 2 can be thought of as an indicator of the probability that someone reading randomly through the output of newspaper printing presses during this period would encounter a story based on a poll that contained questions on this topic (versus other poll topics).

As Table 2.5 indicates, the subject matter of the surveys and polls reported in the press is extremely varied. However, nearly all of the polls published in the newspapers reported measurements of subjective phenomena. Of the 71 British and 253 American polls reported in our file of newspaper stories, only 7 and 5 of them, respectively, referred solely to factual or quasi-factual matters. Clearly it is what people think rather than what they do that makes for punchy stories about polls.

For the most part the topics of polls in the two countries were similar. The largest difference between the polls reported in the British and American press is for election polls. It should surprise no one to learn that reports of national and local election polls predominated in the American press in July 1980. (If anything, the surprise is that 6 percent of the British clippings reported American election polls.) These data suggest that the reporting of items that are a staple of opinion polling (presidential popularity, evaluation of the problems in the country today, and general level of support for particular parties) is overshadowed by candidate preference polls during an election campaign. (A comparable study conducted in a nonelection period would be necessary, of course, to provide direct evidence of this.)

Other differences in poll reporting reflect political differences between Great Britain and the United States. The emphasis on trade unions, pay increases, and attitudes toward pending legislation in the British polls reflected British concerns with Prime Minister Margaret Thatcher's economic policies. In addition, the annual Trades Union Congress, which discussed pay restraint and attitudes toward the government's proposed changes in the legislation relating to trade union activity, took place during the time period we studied.[24]

The greater expression of public concern over matters of defense (and to a lesser extent international affairs) in the American poll reports may reflect the fact that these stories were written during the 1980 election campaign, in which military defense and international relations were major issues. The differences between Great Britain and the United States in the mention of satisfaction with commercial products reflects the fact that the British clip-

TABLE 2.5

Percentage Distributions for Detailed Topics of Polls Reported in Newspapers in the United States and Great Britain

	Percentage Distribution of Topics			
	Unweighted		Weighted by Circulation[a]	
Topic	**U.S.A.**	**G.B.**	**U.S.A.**	**G.B.**
Popularity of president/prime minister	3.9%	5.3%	1.5%	8.3%
Popularity of other leader	4.6	5.3	4.1	6.8
Election polls	26.6	6.1	29.2	5.0
Problems in country	3.1	6.1	6.2	8.0
Party support	1.8	4.6	3.9	4.1
Popularity of government (in Great Britain)	—	1.5	—	2.7
Current legislation	0.9	9.2	0.6	10.8
Unions	0.0	6.9	0.0	10.2
Pay increases	0.0	5.3	0.0	11.6
Monetary policy	4.6	4.6	6.5	9.2
International affairs	5.1	3.0	6.0	2.2
Religion	1.8	2.3	1.4	0.4
Sports and leisure	3.6	4.6	1.8	1.6
Commercial products	0.9	6.9	0.1	1.7
Crime	0.6	0.0	0.1	0.0
Drugs	0.0	0.8	0.0	1.3
Race	1.8	0.0	2.4	0.0
Issues affecting women	5.4	1.5	5.5	0.0
Unemployment	3.3	3.0	3.1	1.9
Inflation	5.1	2.3	7.2	7.2
Defense	5.4	0.0	6.8	0.0
Other social issues	11.2	13.0	8.2	4.4
Consumer confidence measures	1.2	0.0	1.8	0.0
Business confidence measures	1.5	0.8	0.8	0.3
Other (nonsocial) issues	7.6	6.9	2.8	2.5
Total number of polls	253	71	—	—
Total number of newspaper stories	1,038	168	—	—

NOTE: Stories were (multiply) coded to show *all* topics covered in each poll; the resulting distribution was then rescaled to show the percentage of total mentions given to each topic. Appendix C presents details of the study and subject coding. The period covered is July 1–31, 1980, for the United States and July 31–September 11, 1980, for Great Britain.

[a]Poll topics were weighted in proportion to the combined circulation of all newspapers reporting the poll.

pings included stories from trade magazines, which sometimes reported market research studies.

In order to summarize the subjects treated in poll reports in our file of newspaper stories, we coded each poll into broader categories intended to reflect the dominant subject matter of the poll as a whole. A crude distinc-

tion was made between broad categories of social, economic, and political matters. If possible, topics were further coded into more specific categories: unions, energy, and elections. Table 2.6 presents the results of this summary (see table notes for details of coding). The relative predominance of trade union related topics in Great Britain and of election and other politi-

TABLE 2.6
Characterization by Broad Subject Matter Groups of Poll Reports in the United States and Great Britain

	Percentage of Circulated[a] Poll Reports	
Subject Area Code[b]	**U.S.A.**	**G.B.**
Union-related	0.0%	43.0%
Energy	2.6	0.1
Other economic	14.1	7.1
Election	38.8	14.7
Political (other than election)	23.2	17.1
Other social	17.4	5.9
Other nonsocial	3.8	12.2
Total number of polls (unweighted)	253	71
Total number of newspaper stories	1,038	168

NOTE: The period covered is July 1–31 for the United States and July 31–September 11, 1980 for Great Britain.

[a]Weighted in proportion to the combined circulation of all newspapers reporting the poll.

[b]A single variable was created to indicate the subject of the poll. The values of the variable were coded hierarchically in the following order of precedence:

"Union-related" category includes polls relating to unions and pay increases;

"Energy" category includes polls relating to energy;

"Other economic" category includes polls relating to attitudes toward monetary policy, employment, inflation (but not pay increases), consumer and business confidence measures (but not energy);

"Election" category includes candidate preference and other election polls;

"Political" category includes polls relating to presidential or prime ministerial popularity. The popularity of other leaders, problems in the country, party support, "government" popularity (in Britain), attitudes toward current legislation, international affairs, and defense;

"Other social" category includes polls relating to attitudes toward crime, drugs, race, women's issues, religion, sports and leisure, and other social issues;

"Other nonsocial" category includes polls relating to a variety of topics such as favorite foods, favorite airports, school cafeterias and hospital foods.

Thus, economic subjects were subdivided into union-related, energy, and other economic topics. A topic was considered political only if it could not also be classified as economic.

cal (but noneconomic) matters in the American media during this period is quite clear.[25]

2.4 The Effects of Surveys and Polls

It is clear from our review that surveys of subjective phenomena have become a ubiquitous part of contemporary life, and it is likely that technological innovations, such as computer-assisted telephone interviewing and two-way cable hookups, will spur the growth of the survey enterprise and shorten the elapsed time between taking survey measurements and reporting them. Today, virtually instant polls on controversial issues are becoming a growing part of daily life. And survey research has come to play an important role not only in journalism, but also in government data gathering, basic social science research, and even legal decision making.

Since the scope of the survey enterprise is so large, one might expect its impact on the public would be correspondingly large. The crucial word here is "expect": actual documentation of the influence of polls, whether properly or improperly conducted, is rare, although it is widely asserted that polls have major influence. Sean MacBride (1980), who served as president of the UNESCO Commission for the Study of International Communication, for example, has asserted that:

> In democratic countries where there are a multiplicity of political parties, political parties now *have* to organize opinion polls to insure that their policies are in conformity with the trends of public opinion. These public opinion polls are already exerting considerable influence in the shaping of policy by parties. The art of the public opinion poll as a method for influencing policy is only still in its infancy.

Some anecdotal evidence of such influences is available. Elizabeth Drew (1979) claims, for example, that poll results played an important role in President Carter's actions during summer of 1979 when he diagnosed a "national malaise" and fired several members of his cabinet (see also, Converse, 1979; Von Hoffman, 1980). In addition, presidential popularity measurements are frequently alleged to be key determinants of presidential action—and inaction (see MacKuen and Turner Chapter 16, Volume 2; Brody and Page, 1975; Steele, 1978). Brody and Page (1975:136–137) assert: "We can be sure that Presidents themselves are avid readers of the polls," and historical evidence (see Steele, 1978) suggests that presidents as far back as Franklin Roosevelt have followed the polls.

Recent commentators on American politics have observed an increase in

the use of polls in the political process. Sabato (1981), for example, notes that while "perhaps, good politicians can rely on instinct . . . virtually none are left who trust instinct alone. The political pollster and his surveys are omnipresent . . . ." Knowledgeable observers of Washington politics have observed an enhanced role of polling in presidential politics. Bonafede (1981:2,184), for example, argues:

> . . . the increased reliance of Presidents on [pollsters'] skills is reflected in the White House operational structure. Presidents have used professional pollsters at least since Franklin D. Roosevelt, but it was not until Carter made Caddell . . . an informal member of his White House advisory staff that they became ensconced as part of the presidential establishment . . . .
>
> Perhaps more than any other Administration, the Reagan White House uses polling, public opinion analyses and media and marketing research in the decisionmaking process and the selling of the presidency.

In addition to their use by governments and politicians, polls have also been used on occasion as tools in actions *against* governments. During the McCarthy era, Samuel Stouffer (1954:234) delivered a presidential address to the American Association for Public Opinion Research in which he observed:

> Proof of past Communist subversion and espionage in the United States arouses fears that it may be happening now . . . All these fears, real and imagined, are fanned by irresponsible politicians into a flame which could destroy our heritage of civil liberties handed down through long centuries of Anglo Saxon tradition.
>
> To combat the threats from without and the threats from within, all the resources are needed of people dedicated to freedom armed with the best in technology. One of these resources is the social sciences, especially as they find expression in the tools for the analysis of public opinion.

Stouffer's (1955) monograph, *Communism, Conformity, and Civil Liberties*, and Rosenberg, Verba, and Converse's (1970) *Vietnam and the Silent Majority: The Dove's Guide* are important examples of the use of survey research in opposition to government actions.[26] Polling done by the Solidarity labor movement in Poland in 1981 may provide another such example (Karatnycky, 1981).

While examples and anecdotal evidence are available, it is difficult to find convincing scientific evidence of the nature or magnitude of the effects of polls and surveys in contemporary society. (See Recommendation 18, Chapter 10, on the need for research on this topic.) Research has been conducted on the function of the media in setting the agenda for political debate. Through its growing role in the media, polling has a powerful po-

tential for such agenda setting—it may create the impression that the issues polled about are indeed the topics of public opinion or the focus of current controversy. Political scientists have begun to test the agenda-setting hypothesis by comparing public concerns as expressed in opinion polls with external measures. Funkhouser (1973) found that the American public's view of what was important to the nation during the 1960s correlated with newspaper coverage of topics more strongly than it correlated with measures of "real world" conditions. Moreover, MacKuen, Chapter 14 (Volume 2), and Beniger, Chapter 15 (Volume 2), have produced striking evidence that the path of influence is from the media to public opinion. MacKuen, for example, reports that for most issues there is a clear *lag* period: changes in issue salience in the media are followed by changes in issue perception by the public. And to the extent that polls become part of the news, they, too, may influence the public agenda.

One instance in which polls had a more immediate and direct influence on public policy should be mentioned here. When the League of Women Voters in 1980 had to decide which presidential candidates it would invite to take part in its televised debates, it relied on opinion polls of public preferences for the various candidates (League of Women Voters, 1980a; 1980b). This use of opinion polls was criticized by a number of people and organizations, including the National Council on Public Polls (1980b), but the League wanted a putatively "objective" criterion on which to base its decision. Its use of opinion poll data had the effect of sparing the League the discomfort of having to publicly decide who was a serious candidate for president.

Other possible influences of the polls in the political arena are not so easily documented. The abundance of electoral polls conducted by various organizations for various purposes, and with sometimes inconsistent results, has prompted speculation and disagreement about their influence on voters and politicians. While some pollsters, including Louis Harris, admit that polls probably affect politicians indirectly through their campaign contributions, others, including Robert Teeter, feel that a candidate's popularity, rather than the polls themselves, is the crucial factor influencing campaign contributions (Thornton, 1980). In addition to potentially influencing a candidate's ability to raise enough money to finance a political campaign, polls may also influence the decision of voters (for example, whether or not to vote as well as for whom to vote). The argument about such effects, particularly so-called bandwagon effects, has been debated for some time (for a full discussion see Marsh, Chapter 18, Volume 2). As the number of election polls increases, the possibility for and extent of any such effects would seem to increase accordingly.

In addition to the abundance of political polls, measurements of opinion on a wide variety of topics are available, as we discovered in our clipping

study and as can be seen by flipping through the Gallup Opinion Index. The effect that this "mirror-image of society" aspect of polling has on the public is unknown, but possible influences can be imagined. The temptation may exist, for example, to use survey results as a substitute conscience –to justify some course of action or ideology by the fact that X percent of (some) population feels the same way. The extent to which this occurs is unknown, but the likelihood of such an effect increases, we suspect, as the volume of published poll reports increases.

The size and importance of the survey industry, and its growth, heightens concerns for the integrity of the data that are disseminated. In the following chapter we review some major attempts to set standards for survey practice and evaluate the impact of those standards on survey practice.

Notes

1. The reason for the poll's failure has been reconsidered by Bryson (1976), albeit without reference to Parten's discussion. He holds that repetition of the conventional account, whose original source he was unable to locate, amounts to the perpetuation of a myth. A myth persists when each repetition echoes uncritically the previous repetition and when retelling it offers a gain (or, at any rate, no loss) to the teller. As Bryson (1976:184) notes, "Pedagogically, there is no more useful or enjoyable device than the horrible example." The *Literary Digest Story* serves the purposes of textbook authors very well. It might be asked whether other accepted diagnoses of survey defects similarly have a questionable basis in fact.

Link (1980) endorses Bryson's conclusion about the effect of nonresponse on the *Literary Digest* poll and provides some interesting new evidence on the electoral projections derived from the poll. Link shows that adjustment of the 1936 projections to reflect state-by-state inaccuracies of the polls' predictions in the 1932 presidential election would have produced a very different result. Link makes the ratio adjustment

$$\hat{P}'_{i,1936} = \hat{P}_{i,1936}\left(\frac{P_{i,1932}}{\hat{P}_{i,1932}}\right)$$

where $\hat{P}_{i,y}$ is the poll's unadjusted estimate of the proportion of votes cast for Roosevelt in state, i, in year, y; $P_{i,y}$ is the actual proportion voting for Roosevelt; and $\hat{P}'_{iy}$ is the adjusted estimate). Link (1980:55) finds that this adjustment would have moved the poll's projections for 19 states from the Republican to the Democratic side of the ledger with the result that "the electoral projection would have pointed to a Roosevelt landslide (or at least a substantial win) instead of the Landon landslide predicted by the unadjusted data." (For an earlier treatment of these data using a similar approach, see Wallis and Roberts [1956:485].)

2. As with the other data we review in subsequent chapters, the problems of measurement and recall inherent in such estimates are formidable. These estimates, thus, should be treated with caution. (See Appendix B for further details.)

3. As defined in Section 2.4, the number of copies is the sum of the number of stories weighted by the circulation of the paper publishing the story.

4. As their name implies, CATI systems use a computer to assist the interviewer. Interviewing is done by telephone from a central location, and the computer assists by dialing telephone numbers, selecting the proper sequence of questions to be asked, coding the data, checking for consistency between responses, and so forth (see Dutka and Frankel, 1980). Use of telephone rather than face-to-face interviewing eliminates travel and reduces the total time and costs for a survey (see Groves and Kahn, 1979, for further discussion). For government surveys, given the mandate of the government, it will be necessary to use supplementary samples to provide coverage of households that do not have telephones.

5. Because of limited resources, we restricted our analysis to surveys conducted by or for the federal government. However, state and local governments also conduct many surveys involving subjective measurements. For example, in the state of North Carolina (Office of State Budget and Management, 1980:1)

> . . . statewide citizen survey has been conducted annually or semi-annually since the fall of 1976. . . .
>
> Topics of interest and suggested questions for each survey are provided by the Office of State Budget and Management as well as by other agencies and offices. Questions included on the survey attempt to . . . describe characteristics, conditions, and opinions of North Carolinians across the state. Resulting data has served many functions. Planning, budget allocation, policy-making, and program evaluation efforts within state government have benefited from the data gathered through the survey.

While we are aware of other state survey programs, we do not know of any reliable compilation of state and local activities of this type.

6. A number of these implicitly subjective measurements (unemployment, ethnic group, housing quality) are discussed in detail in Volume 2.

7. The Retirement History Survey, which was conducted biennially from 1969 to 1979 for the Social Security Administration, asked respondents: "Taking things all together, would you say you're very happy, pretty happy, or not too happy these days?"

8. The 1980 census also included the following subjective and quasi-factual measurements: "What is this person's ancestry?" (Question 14); "Does this person have a physical, mental or other health condition which has lasted 6 or more months and which (a) *Limits* the kind or amount of work this person can do at a job? (yes, no), (b) *Prevents* this person from working at a job? (yes, no), (c) *Limits or prevents* this person from using public transportation? (yes, no)" (Question 19); "Has this person been looking for work during the last four weeks? (yes, no)" (Question 26a); "What is the value of this property, that is how much do you think this property (house and lot or condominium unit) would sell for if it were for sale?" (Question H-11). Question H-11 was asked of all home owners; the other questions were included in questionnaires sent to about 20 percent of households. In addition to these questions, the 1980 census required judgments as to the number of rooms and the number of bedrooms in the household's living quarters and about whether unrelated individuals in the same household were "partner/roommate(s)" or "other nonrelative(s)." (See Bailar and Rothwell, Chapter 5; Newman, Chapter 6; and Smith, Chapter 4 in Volume 2 for a discussion of the subjective components of some of these quasi-factual measurements.)

9. At the time of our study (fall 1980) all government forms requesting information from the public, including those for government-sponsored surveys, had to be submitted for approval to the forms' clearance offices of OMB or GAO. Appendix D presents a detailed description of the procedures we used in our review of the OMB and GAO dockets.

10. We reviewed summary abstracts, and, if necessary, the complete OMB clearance file and questionnaire were also reviewed to determine whether subjective measurements were made in the survey. Thirteen cases were excluded from the analysis because a determination could not be made on the basis of the abstract alone and the OMB clearance file was not available at the time of our inquiry. In addition, 13 forms related to the 1980 census were excluded. Hence, we had a total of 202 surveys for our analysis.

11. This is referred to in the OMB docket as the "number of responses" and is defined as: the product of the estimated number of respondents per year times the number of reports annually by each respondent.

12. A separate tabulation (not shown in Table 2.1) indicates that this pattern remains stable regardless of whether the surveys are conducted by government or private contractors.

13. One might speculate that program evaluation (in contrast with research or general statistics) involves an inherently subjective element, and that surveys conducted for this purpose might incorporate subjective measures to a greater extent than surveys conducted for other purposes. This hypothesis is partly supported by the data we have collected. Government surveys whose purpose was evaluation (as identified in OMB records) were more likely than general purpose statistical surveys to make subjective measurements (81 percent versus 50 percent). Surprisingly, of those surveys making subjective measurements, program evaluation surveys were (slightly) more likely to be conducted by government agencies than by private contractors (36 percent versus 30 percent). In contrast, the relatively small number of surveys conducted for research purposes were more likely to make subjective measurements (84 percent), and 92 percent of these "subjective" surveys were carried out by private contractors.

A cursory review of the questionnaires used in federal surveys revealed wide variation in the number of subjective questions that appeared in the surveys categorized as subjective. No differences concerning the number of subjective questions in surveys conducted within and outside of the government were formally documented, although it was our impression that subjective questions seemed to be more heavily concentrated in government surveys that were conducted by private contractors.

14. We do not know precisely when such foreign survey activity began; we do know that extensive polling was done during the postwar occupation. As Jensen (1977:14) notes:

> In occupied Germany, the U.S. Army sponsored an elaborate series of polls to monitor acceptance of defeat and the success of [de-Nazification]. As for postwar occupied Japan, one former OWI analyst recommended . . . "If we are really trying to reeducate Japan, [it] should be prefaced by intensive research to locate social and political blind spots which must be eradicated before democratic thought is possible."

15. In the Soviet Union, *Pravda* recently carried an article (Safarov, 1981) that appeared to urge increased efforts to monitor Soviet public opinion. That article argued that

> Public opinion is a sensitive barometer, and the evidence it provides, if carefully analyzed, can tell us a great deal about the deep processes of social life, processes that are in some cases barely discernible or just emerging.

> Often this "barometer" provides timely and precise warnings about the contradictions and conflict situations.
>
> . . . Measuring public opinion and making it known in connection with such contradictions and situations is the proper way to keep the political system functioning on the basis of popular consent.

The article continued by noting one important stimulus for this concern for public opinion measurement and by offering a caveat about the limitations of any "generalizations . . . drawn from a single poll":

> At the 26th CPSU [Communist Party of the Soviet Union] Congress, Comrade L. I. Brezhnev particularly noted, with reference to the conditions in Poland, how important it is for the party—and for strengthening the Party's leading role—to listen attentively to the voice of the masses. . . .
>
> . . . it is no accident that special attention is being paid to the scientific level of public opinion studies. . . . Of course, any information on the state of public opinion can prove valuable and useful, but one must nonetheless recognize that serious conclusions require precise, objective, full and repeatedly verified information on public opinion.
>
> At the present time, the problem of studying the effectiveness of public opinion has become particularly urgent. It is the researcher's job, moreover, to [measure] both the degree of satisfaction of workers' needs and the time that elapses between the moment when the criticism and suggestions are first voiced "from below," and the time when the responsible agencies take the requisite action.

The People's Republic of China is not thought to have the capacity to conduct national or regional sample surveys (Frankel, 1981), but *Beijing Review* recently carried the report of an "opinion poll of the young" reporting the responses of 1,000 youth "select[ed to provide] as broad and diverse a group as possible" (Li Ming, 1981:22). Questions asked included "How do you view the socialist system?" (This question drew a notable number of critical responses.)

16. This program originally included subjective measurements. In 1973, the OECD enunciated a list of 24 social concerns, three of them subjective: students' satisfaction with education, employees' job satisfaction, and confidence in the administration of justice (OECD, 1973). Indeed, many OECD participants argued in favor of including subjective measures in all areas of concern—a strategy followed in recent U.S. social indicators reports (for example, U.S. Department of Commerce, 1977; Bureau of the Census, 1980a). An OECD working party worked through the 1970s to develop social indicators in the various domains of concern (OECD, 1976). Several comprehensive surveys have been undertaken by individual OECD member nations.

17. The best known of these are the Eurobarometer surveys, which have been conducted regularly in the EEC member countries since 1973. Subjects covered in these surveys include such general topics as respondents' satisfaction with their lives, standards of living, occupations and so forth. More specific issues, for example, childbearing intentions, have also been covered, as well as political attitudes. Questions have also been asked about EEC elections, European integration, satisfaction with various political institutions, and expectations of war. Besides these regular surveys, the statistical office of the EEC has also undertaken subjective surveys on health, housing, and working life in anticipation of building a series of subjective social indicators in these areas.

18. In Great Britain, one particularly nettlesome aspect of the relationship between academic researchers and survey research organizations involves bidding and formal contracting for survey projects. It is common practice in that country for academic researchers to draw up a specification for a piece of survey research that they then submit to competitive bidding. (In

many respects this practice and its consequences parallels that of government-contracted surveys in the United States.) In such situations, one is generally under great financial pressure to accept the lowest bid, which may often be half the price of the most expensive one. However, the large discrepancies in cost often stem from the costing of substantially different jobs. It takes a very carefully designed specification to produce comparable cost estimates, and, therefore informed decisions about value for money. Guidelines do exist to help researchers to specify their job in sufficient detail (for example, Marsh, 1976), but the more detailed the specification the more time and effort it takes to estimate costs.

Moreover, the system may not be in the interest of maintaining expertise in general. The rationale of a bidding system, the costs associated with it, and the demand for mobility of resources that the system generates in the performer (contractor) universe may have an impact on methodology. (A large amount of pre-specification may, for example, limit the development of unusual or novel research designs and favor economical execution of routine work. Conversely, contractors face the conflicting need for clear differentiation of their firm's approach from that of others.) Survey research firms may cope with the many strains inherent in this system by developing broadly applicable techniques rather than highly specific expertise, by relying on personnel with broad skills rather than subject matter experts, and by operating through "contingent hiring practices and other devices which maximize flexibility" (Biderman and Sharp, 1973). To overcome some of the undesirable features of this system, it has been suggested (Social Research Association, 1980) that research companies be paid a small sum to prepare detailed cost estimates for a job.

19. A slightly earlier use occurred in the federal prosecution of Alger Hiss [*U.S.* v. *Hiss*, 185 F.2d 152 (2d Cir. 1947); see, Collins, 1978:163].

20. Our information is derived from Associated Press wire stories dated February 11 and 12, 1952 (the first of which was clipped from the *Clearwater Sun*), a much longer newspaper account dated February 13, 1952, and a personal communication from Burns Roper (whose father took the poll in question). The quotation is from the third newspaper story, which unfortunately lacks any notation as to the newspaper from which it was clipped; the information, however, agrees with that contained in the Associated Press stories.

21. We note also another case in England involving similar issues. In *Powley* v. *ACAS et al.* (1977), the court overturned an ACAS ruling on the grounds that the opinion survey "was biased in that it would unfairly favour the referring union at the expense of the [unstated alternative of a] staff association" (cited in *National Life Assurance Company, Ltd.* v. *ACAS et al.*; High Court of Justice, Chancery Division, February 23, 1979; judgment of Mr. Justice Browne-Wilkinson).

22. If we take account of the caveat noted in Appendix C that our count covers only 10 to 20 percent of the polls that were actually published in the newspapers during our clipping period, the potential audience for the clippings is even larger. (The raw counts for Great Britain were deflated by 31/43 to adjust for the longer time period of clipping there.)

23. In addition to this exposure, our review of television news stories between 1972 and 1979 suggests that about 9 poll stories were broadcast each month on the three national television networks. Our review was conducted by reading abstracts of the national (evening) news broadcasts of the ABC, CBS, and NBC networks for every day of one month (June) in each of the years of 1972 through 1979. We used a single month because the organization of the transcripts made it cumbersome—given our limited resources—to undertake a probability sampling from all the days in each year.

The abstracts were compiled by the Vanderbilt Television News Archives and were supplied to us through the courtesy of the Library of the George Washington University.

24. As mentioned earlier, the weighted figures in Table 2.5 can be taken as indicators of the probability that the results of a poll, once printed, would be read by someone choosing randomly from the daily output of the nation's presses. Any marked difference between the weighted and unweighted percentages is an indication of the extent to which a topic is viewed as newsworthy by many newspapers or by papers with large circulation. For the most part, the differences between the weighted and unweighted figures were small; in some instances, however, they were not. From the weighted figures, the most newsworthy polls in America were on problems in the country and on party support. Among British newspaper stories, one topic—inflation—appeared to be particularly newsworthy: it has a weighted count that is three times as high as the unweighted one. Two other topics also appear much more frequently in print than their unweighted counts indicate. These topics are pay increases and monetary policy. Since polls on inflation, monetary policies, and pay increases were all related to the discussions at Great Britain's annual Trades Union Congress, this result is not surprising.

25. A direct comparison of figures in Table 2.5 and Table 2.6 is not possible because Table 2.6 assigns a single code to reflect the subject of the poll while Table 2.5 allowed multiple subject codes for each poll.

26. The opening paragraphs of Rosenberg, Verba, and Converse (1970:1–2) note:

> This book is about public opinion and the Southeast Asian war. It is frankly partisan in its guiding purpose. We view the war that the United States has waged, and is waging, in Southeast Asia as ruinous for the inhabitants and governments of that area and as a burden which our own political system and national life cannot much longer sustain. . . .
>
> So convinced, we aspire to do what we can in the continuing campaign to persuade the American people that they must reject this war and call for its rapid termination through political settlement. As university-based specialists in the study of political public opinion and the attitude change process, what we think we can do is illuminate some important aspects of the persuasion task to which thousands of Americans have now committed themselves. . . .

The book then continues with a remarkable and modest statement of the limits of the survey craft:

> In the sense that we aspire to advance such efforts, our interest is, then, unabashedly partisan. But we are convinced that a partisan commitment to these goals would be poorly served if we allowed it to deflect us toward unjustified optimism . . . Focusing on public opinion about the Vietnam war, we have striven to avoid an intellectual failing which has too often corrupted the analyses drawn by other critics of that war. Thus we do not seek to prove that poll data and other indexes of public opinion show a majority of American people to now be basically opposed to any continuing involvement in Vietnam. . . . The ambiguities inherent in poll data—and also rather endemic in the political thinking of many Americans—*can* be so interpreted. But such interpretation is more often wishful than accurate, more often propagandistic than analytic.

3

Standards of Practice and the Misuse of Surveys

The practice of survey research is constrained by the limits of knowledge, by the resources available for a given study, and by the norms of the research community as to what constitutes adequate practice. Several attempts have been made to codify the latter in terms of standards for the conduct and reporting of survey research. While these standards are not always met, they do establish a lower limit for what is considered to be acceptable practice and, in some cases, provide mechanisms for the public exposure and censure of deficient practices.

In this chapter we first briefly outline some relevant sections of four sets of standards. These include an early proposal made by Stuart Dodd (1947), the standards of the American Association for Public Opinion Research (AAPOR) (1979–80), the standards of the Council of American Survey Research Organizations (n.d., 1982), the principles of disclosure of the National Council on Public Polls (NCPP) (1979), and the standards for federal surveys (Gonzalez et al., 1975). In the second part of this chapter we use data from our studies of newspaper clippings and academic publications and other sources to assess the state of survey practice and survey reporting. In the final section we examine examples of misuses of polls and surveys.

3.1 Attempts at Setting Survey Standards

THE DODD PROPOSAL

In 1947 *Public Opinion Quarterly* published an article by Stuart Dodd that proposed a detailed code of standards to govern the certification of opinion survey organizations. The standards were never adopted, but they represent a view of what was seen as important 35 years ago—and still is.

Dodd proposed standards for the conduct and the reporting of opinion survey results in 41 separate areas (some with several subsections), covering such aspects of survey practice as sampling adequacy, interviewer recruitment and supervision, questionnaire design, public availability of survey results, the integrity and professional competence of the survey organization, and so forth. Dodd's proposal required that survey agencies have a reputation for and pledge to undertake "honest fact finding." To ensure this, he recommended full disclosure of survey methods and frequent validity studies. Survey organizations were also directed to have both an "interpartisan" or nonpartisan staff and backers (Standard I.2.c), and sponsorship by "public leaders or social scientists of high integrity." Survey organizations were also forbidden from becoming "public partisans or pressure groups on issues surveyed" (I.3). Indeed, Dodd's standards even envisioned "an international inspectorate . . . to investigate [the] standards, integrity and performance of any survey agency, when invited" (I.4).

While it's possible to complain that Dodd's view of the opinion pollster as a disinterested purveyor of "facts" is unrealistic, it did fit the prevailing ideology of the leading pollsters of that era. Gallup, for example, saw the polls as "a new instrument which may help bridge the gap between the people and those who are responsible for making decisions in their name" (Gallup and Rae, 1940:14). Such a perspective suggests that an appropriate code of standards for polling might be similar to that for a board of elections, that is nonpartisan, independent, and subject to external review. Dodd's proposal reflects precisely this orientation.

In other areas, Dodd's standards propose criteria of excellence which are as relevant today as they were in 1947. For example, on the issue of validity, Dodd wrote (I.5):

> Agencies should make studies of their survey results, correlating them against criteria (such as elections, censuses, government statistics, larger samples, friends' samples, and other independent measures of behavior or conditions which check on assertions, etc.) whenever possible and publish the correlations. Accurate predicting is the final test of validity.

Similarly, Dodd exhorted standardization of survey practices to improve across-survey comparability (I.6):

> Agencies should strive for comparability between different surveys, in different time periods, in different countries or languages, by progressive research on and standardization of terms and technics.

We note that if such a standard had been adopted on any scale, it might have prevented or reduced some of the inter-survey discrepancies that have haunted us in recent years and, indeed, as described in Chapter 1, that partly motivated the present project.

Other standards in Dodd's proposal counseled that pretesting of questions should be carried out on "expert" samples as well as samples of the target population and that "split ballots [should] also be used to test [the] effect of variant wording." They also urged "rigorous randomizing" in sampling (III.20.C); routine reliability studies for all measurements (III.23); full disclosure of question wordings, sample design, and sponsorship of the survey (V.32); and preservation of all records, preferably in a centralized archive so that they would "be preserved accessibly for further research or inspection" (V.35).

Dodd's standards for questionnaire design are frequently surprising in their approach to the difficult problem of question wording and question bias. For example, standard II.11 instructs that "the meaning of questions should be unambiguous, definite, and constant for all as tested by the percentage of respondents who, when they restate it in their own words, agree essentially on its interpretation." Dodd's standard is akin to the procedures discussed in Chapter 7 of this volume. Similarly, standard II.12 on question "fairness" dictates

> Questions should be so phrased as to analyze the issue fairly to all factions as testable by the percent of agreement on phrasing by a competent panel of persons holding all the diverse opinions on that issue.

This formulation appears to envision a preliminary design stage of questionnaire construction in which alternative forms of an item would be rated by panels of judges who vary in their own position on the issue. Presumably, a "fair question" would be one that maximizes some function of the proportions of judges designating it as "fair." Such a pretesting procedure, which is today almost never formally applied in polls and surveys, might have the virtue of airing disputes about question wording and question bias in advance of a survey.

THE AAPOR STANDARDS

In 1968, the American Association for Public Opinion Research responded to mounting public criticisms of polls and congressional proposals to regu-

late the industry by adopting a code of ethics and professional practice.[1] The AAPOR code is less detailed than that proposed by Dodd. In its present form it provides quite general prescriptions concerning standards of research practice and members' obligations in dealing with the public, clients, respondents, and the profession. These prescriptions include (American Association for Public Opinion Research 1979–80):

> [Members] shall not knowingly make interpretations of research results, nor shall [they] tacitly permit interpretations which are inconsistent with the data available [I.B.3].
>
> [Members] shall not knowingly imply that interpretations should be accorded greater confidence than the data actually warrant [I.B.4].

These standards extend not only to the behavior of AAPOR members, but also cover all public presentations of data collected by AAPOR members. Thus Standard II.A.2 states that

> When [members] become aware of the appearance of serious distortions of [their] research they shall publicly disclose what is required to correct the distortions.

Because each of the standards is stated generally, there is wide latitude for interpretation. Consideration of attempts to enforce these general guidelines provides some insight into their meaning. According to the chairperson of the AAPOR standards committee, only two complaints have been subject to intensive review; this is said to be due to the lack of complaints against AAPOR members. The organization, of course, has no power to impose its standards on nonmembers. (In Section 3.3 of this chapter, we review the application of AAPOR's standards in the case of one of these complaints.)

THE CASRO STANDARDS

In 1975 senior executives of several major survey research organizations founded the Council of American Survey Research Organizations (CASRO). CASRO describes itself as a trade association of companies specifically involved in the survey research industry, and its charter specifies that its major objectives include promoting the best interests of those companies engaged in the survey industry, promoting the expansion of the industry, acting as a spokesman on issues confronting the industry, and promoting the establishment, maintenance, and improvement of professional standards in survey research (see Council of American Survey Research Organizations, n.d.).[2]

In pursuit of the latter objective, CASRO adopted in 1979 a "Code of Standards for Survey Research." The code specifies a series of professional responsibilities to respondents (for example, protection of confidentiality and privacy and avoidance of harrassment), to clients, and to the public. Among the responsibilities of companies to clients is a general recognition of the need for rigorous research (Council of American Survey Research Organizations, n.d.):

> A survey organization must assist its client in the design of effective and efficient studies . . . If the survey research organization questions whether a study design will provide the information necessary to serve the client's purposes, it must make its reservations known.

In the reporting of survey results, the CASRO standards hold that "it is the obligation of the research organization to insure that findings they release are an accurate portrayal of the survey data (III.A)." The specifics of this obligation are described in considerable detail. The standards (Section III.B) require that a survey report should include ("or the research organization should be ready to supply on short notice") the following information:

1. The name of the sponsoring organization
2. The purpose of the study, including specific objectives
3. The dates of the field work
4. A definition of the universe which the survey is intended to represent and a description of the population frame(s) that was actually sampled
5. A description of the sample design, including cluster size, number of callbacks, method of interview, screening criteria, method of selecting sample elements, and other pertinent information
6. A description of results of sample implementation, including a full accounting of the final outcome of all interview assignments
7. Documentation and a full description for any completion rate that is cited
8. The exact wording of the questions used
9. A description of any weighting or estimating procedures used
10. A description of any special scoring, data adjustment or indexing procedures used
11. Estimates of the sampling error, when appropriate, including reference to other possible sources of error so that a misleading impression of accuracy or precision is not conveyed
12. Clear labeling of statistical tables including identification of questionnaire source, and the unweighted number of cases forming the base for each cross-tabulation
13. Copies of interviewer instructions, validation results, and other important working papers

The standards specify that all general public releases of survey findings will initially include "*at a minimum*": the names of the organization sponsoring and the agency conducting the survey, a description of the purpose of the survey, a sample description, fieldwork dates, the exact wording of questions, plus "any other information that a layperson would need to make a reasonable assessment of the reported findings."

The standards require that survey research organizations seek agreements from their clients for "review and clearance as to accuracy and proper interpretation" of any public release of survey findings made by the client. The CASRO standards also require that clients be informed that the survey organization reserves the right to make any necessary public clarifications if the client's public release of the survey findings is "incorrect, distorted, or incomplete."

The CASRO standards do not specify what enforcement mechanisms will be used when they are violated, but procedures were expected to be adopted in 1982. As of 1982, no complaints had been filed against any CASRO member for violation of these standards (Diane Bowers, CASRO officer, personal communication, April 26, 1982).

THE NCPP PRINCIPLES OF DISCLOSURE

The National Council on Public Polls was established in 1968 to represent the interests of the major public pollsters (Gallup, Harris, Roper, Yankelovich, and others). In 1979, the NCPP adopted "principles of disclosure" whose intent was "to insure that consumers of survey results that enter the public domain have an adequate basis for judging the reliability and validity of the results reported" (National Council on Public Polls, 1979). These principles require disclosure in press releases of survey sponsorship, dates and method of interview, descriptions of the population sampled and sample size, complete question wordings, and all percentages on which conclusions are based.

These principles do not require NCPP members to report such information for private surveys, but if the survey is made public by the sponsor the organization bears the burden of ensuring that the additional information is also made public. The NCPP also established a committee on disclosure that has the right to adjudicate complaints and, if appropriate, require the publication of further information concerning the sample design, sample composition, questionnaire content and question order, filter questions, sample weighting and sampling error of (published) surveys conducted by its members. The principles specify that failure to comply with such directives can result in a member organization being placed on probation or expelled from NCPP.[3]

The principles of disclosure specifically relate to the information contained in press releases rather than to that contained in published news accounts. But the NCPP also urges that "survey organizations reporting results will endeavor to have print and broadcast media include the above items in their news stories and make a report containing these items available to the public upon request" (National Council on Public Polls, 1979).

STANDARDS FOR FEDERAL SURVEYS

A different set of standards has been applied to surveys conducted by federal statistical agencies, in part because many of the issues are simpler. Information collected by the government is (supposed) to be made available to the public and so there is no motive for secrecy. In addition, as both a sponsor and performer of its own survey research, federal agencies have greater power to enforce standards and control the quality of their surveys.[4]

Almost all federal surveys must be reviewed and approved by the Office of Management and Budget (OMB).[5] Standards for this review are described in circular A-46, "Standards and Guidelines for Federal Statistics" (Executive Office of the President, 1976; see also Gonzalez et al., 1975; and see Brooks and Bailar, 1978 for a detailed example). The standards encourage diligent pretesting of surveys and allow for the possibility that the result of pretesting may be abandonment of the proposed survey. However, the pretest and final survey instruments must be separately reviewed by OMB, which, we believe, discourages conscientious pretesting since each OMB review can take several weeks or months to complete.

The OMB standards also prescribe that "potential nonsampling errors, including reporting errors, response variance, interviewer and respondent bias, nonresponse, imputation error and errors in processing the data should be carefully considered" (Executive Office of the President, 1976: 12). However, the standards do not provide guidance as to how this consideration should be given nor what is "carefully" enough. Moreover, the standards do not explicitly discuss question bias and similar issues that have traditionally been of concern in attitude and opinion studies. These omissions may reflect the traditional disinclination of federal statisticians to make subjective measurements. Indeed, this distaste for subjective measurements is reflected in the federal standards' observation that (Executive Office of the President, 1976:10):

> It may be necessary to modify the initial goal of measurement of a concept in order to define an objective statistic which can be measured, e.g.; 'health' is not objectively measurable, but 'bed-days' may be.

3.2 Indicators of the Quality of Survey Practice and Reporting

SURVEY PRACTICE

The existence of standards for the practice of survey research invites the inevitable question as to their effectiveness. In this section we review what evidence is available on the quality of common surveys and the reports of them. It is unfortunately true that there has been no wide-ranging evaluation of the entire domain of surveys. It is thus a necessarily sketchy portrait that we can draw, but we can determine the general contours of survey quality.

In 1975 the Survey Research Methods Subsection of the American Statistical Association (ASA) began work on a pilot evaluation of survey practice. This work was stimulated in part by the fact that "the public and users of [survey] data appeared critical of the quality of many surveys" (Bailar and Lanphier, 1978:vii). As part of a pilot project to design a large-scale study of survey practices, the ASA developed a rating procedure and tested it on a sample of 36 surveys: 26 sponsored by the federal government and 10 sponsored by state governments, academic institutions, or professional organizations. In evaluating the quality of survey procedures, the ASA pilot study considered, among other aspects, the adequacy of sampling procedures, questionnaire design, pretesting, nonresponse and estimation procedures, and the presentation of results.

On the basis of its review of the 36 surveys, the ASA concluded (Bailar and Lanphier, 1978:13):

1. Fifteen of the 26 Federal surveys did not meet their objectives, four because of poor design, four more because of failure to implement plans for probability sampling, and the remaining seven because of a combination of serious technical flaws. . . .
2. Seven of the ten non-Federal surveys did not meet their objectives, one because of poor design, four more because of failure to implement plans for probability sampling, and the remaining two because of a combination of serious technical flaws.
3. Samples were, for the most part, poorly designed. Very little documentation existed. . . .
4. Survey response rates were difficult to collect and compare . . .
5. Quality control over data-processing operations varied considerably. Neither verification of coding nor machine editing to resolve inconsistencies and resolve impossible codes was routinely performed.
6. Estimation procedures varied widely. If weighted data were used for estimation of means, totals and proportions, the weights were frequently not used in estimating the variances of these means, totals and proportions. Variances

were not even estimated in ten of the 24 Federal surveys requiring variances or for six of ten non-Federal surveys.

7. . . . Causal inferences were often made in cases where variables were correlated in such a way that causal relationships could not have been determined by a survey of the type used.

Hence, from this study, it appears that common practice frequently does not meet the standards elaborated in extant codes. This failing seems true even for federal surveys that are subject to an explicit external OMB review.

A further indication of the quality of ordinary nongovernmental surveys can be derived from data on the composition of survey samples. Figure 3.1 presents data on the proportions of college-educated respondents in national adult samples in the United States. It can be seen by comparison with the estimates of the Census Bureau that many well-known polls consistently overrepresent the well-educated segment of the American population in their samples. In spite of a very impressive body of statistical theory and a great deal of practical experience in sampling, survey practice apparently often falls far short of that envisioned in existing standards.

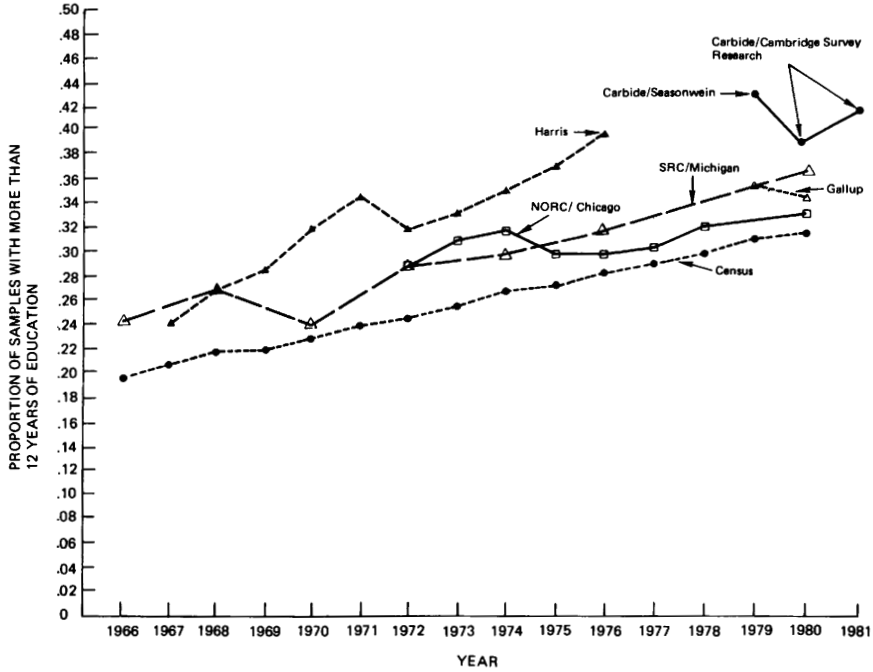
QUALITY OF REPORTING

The limited impact of standards for survey reporting is suggested by the data collected in our exploratory study of American and British news stories (see Table 3.1).[6] In general, these data indicate that little in the way of supplementary detail is made available to newspaper readers. For example, the wording of poll questions is fully reported in only 18 percent of newspaper stories in the United States and 9 percent in Great Britain. The failure to report complete question wordings (a requirement of the NCPP's standards of disclosure), makes it impossible to detect many dubious and misleading interpretations of results (see Section 3.3). Descriptions of sampling procedure, sampling error, nonsampling errors, nonresponse rates, method of interview, and response distributions were lacking in more than 70 percent of the American and British newspaper stories represented in our sample. This lack of information makes it impossible for readers to judge the quality of the data that is being offered as newsworthy evidence in the story. Reporting practices in this regard need to be improved. (See our Recommendation 3, Chapter 10, on adhering to appropriate standards of reporting and disclosure.)

One of the most distressing findings shown in Table 3.1, from the panel's point of view, is the neglect of nonsampling factors in published reports. Despite all the problems that inhere in measuring responses to subjective questions, less than 3 percent of the newspaper stories even vaguely refer to the possibility of nonsampling errors.

FIGURE 3.1

Composition of samples used in surveys by seven organizations: proportion of samples having 13 or more years of education.



SOURCES: Census estimates are derived from the Current Population Survey as reported in CPS Series P–20, Nos. 158 (Table 1, 1966), 169 (Table 1, 1967), 182 (Table 1, 1968), 194 (Table 1, 1969), 207 (Table 1, 1970), 229 (Table 1, 1971), 243 (Table 1, 1972), 274 (Table 1, 1973–74), 295 (Table 1, 1975), 314 (Table 1, 1976–77), 356 (Table 1, 1978; Table 2, 1979). Estimates for 1980 are derived from unpublished CPS data.

NORC estimates are for the General Social Survey; see NORC, *General Social Survey, 1972–1980: Cumulative Codebook*, Chicago: NORC, 1980.

Harris estimates are average of estimates from two surveys selected at random from the national surveys archived at the Louis Harris Data Center at the University of North Carolina. These were computed by Stanley Presser (cf. "The Harris Data Center , Harris National Surveys, and trend analysis, in E. Martin et al.(eds.), *Sourcebook of Harris National Surveys: repeated questions 1963–1976*. Chapel Hill, NC: IRSS—University of North Carolina, 1981: page 43). It should be noted that Presser found that the Harris weighting variable had very little effect on these estimates.

SRC estimates are from the SRC election studies. Estimates for years 1966-1976 are also from Presser (1981). Estimate for 1980 is from Inter-University Consortium for Political and Social Research (ICPSR, 1981) American National Election Study, 1980: September–October (C–3) Survey, November–December (C3PO) Survey. (ICPSR No. 7763), Ann Arbor, MI: ICPSR, page 236. (Note election study data exclude Washington D.C. from sampling frame.)

Gallup estimates are averages of estimates from three surveys randomly selected from among those archived at the Roper Center. These estimates were weighted in the manner prescribed by Gallup. (Source: Don DeLuca, personal communication, April 10, 1981).

Union Carbide/Seasonwein estimates are for survey previously discussed in this chapter. These estimates are from a tabulation presented in Union Carbide (1980a:59). The tabulation is described as giving "the composition of the . . . samples on which data in this report is based."

Union Carbide/Cambridge Survey Research (CSR) estimates are from subsequent surveys conducted for Union Carbide by CSR. The basic publications for these surveys are (Union Carbide, 1980b, 1981). It is not indicated how these estimates were derived from the survey samples (e.g., whether any post-hoc weighting scheme was used), but the reports' text says these numbers "show the percentage that the group represented in the overall population (1980:2; 1981:ii)".

Question wordings vary somewhat between organizations; estimates of question effects are not available.

Age composition of samples. All estimates from 1972 to 1981 are derived from samples of the population

Because the number of clippings that mention such factors is distressingly low, it is important to cite one example of such mention as a model for other reports of survey results. The following comes from an Associated Press wire story that was printed in the Montgomery, Alabama, *Journal* on July 23, 1980:

> As with all sample surveys, the results of the AP-NBC News polls can vary from the opinions of all Americans because of chance variations in the sample. For polls based on 1,949 interviews, the results should vary no more than 3 percentage points either way because of variations in the sample. Of course, the results could differ from other polls for a number of other reasons. Differences in the exact wording of the questions, the timing of the interviews and how the interviews were conducted could also cause variations.

Such an explicit recognition of the potential effects of nonsampling factors in survey research is necessary to alert readers to the existence of such sources of variation in measurement. Of course, it would be even better if the report provided some information on the range of variation that would be induced by varying question wording and other salient nonsampling factors.

We must hasten to note that the reporting of surveys in academic journals also leaves much to be desired. The major medium for disseminating the results of scientific research is the scholarly journal. The norms regulating the reporting of data in such journals are not the same as those governing publication in commercial media, where popularity and readability are all (or almost all). Scholarly journals are oriented toward the goals of replicability, originality, and excellence. Perhaps most important is the norm that the work be reported in sufficient detail to permit replication by another researcher. (Anyone who has refereed work for a journal knows how important it is to distinguish between inadequate reporting—which can be rectified—and inadequately performed research. It is certainly easier to give guidance on [or to rectify] the former.)

Publication of full technical details of a particular survey occurs more regularly in scholarly journals than elsewhere. Nonetheless, numerous

FIGURE 3.1 *(continued)*

aged 18 years of age and older. SRC data in earlier years reflect the voting aged population in each state; this was generally the population aged 21 and over (although a few states had lower ages; *source*: Ann Robinson, ISR, personal communication, March 29, 1981). Harris samples until July, 1969 included persons aged 21 and over. Between July 1969 and "early" 1971 Harris samples included persons aged 16 and over. In all subsequent years the Harris surveys sampled the population aged 18 and over (*source*: Presser, 1981:9).

All Census estimates shown in Figure 3.1 are for the population aged 18 and over. Use of an older population for the years 1966–1971 would not markedly change the comparison shown in this figure. While published Census tabulations (see citations above) do not permit isolation of the population aged 21 and over, estimates can be constructed for the population aged 20 and over. In the years prior to 1972, use of a population aged 20+ (rather than 18+) would have changed our Census estimates to: .201 (from .197 in 1966), .211 (from .207 in 1967), .220 (from .216 in 1968), .223 (from .228 in 1969), .233 (from .228 in 1970), and .243 (from .237 in 1971).

Use of a population aged 16+ in the years 1969 to 1971 (as done by Harris) would have changed our Census estimates to: .232 (from .228 in 1969), .242 (from .228 in 1970), and .252 (from .237 in 1971).

TABLE 3.1
Percentage of Poll Reports Providing Information Required by the Principles of Disclosure of the NCPP and Other Poll Details (using weighted data)

Information	Country of Report	
	U.S.A.	G.B.
Required by NCPP		
Sponsorship of the survey	92	66
Dates of interviewing*	50	33
Method of interview	25	5
Population that was sampled	74	76
Size of sample*	55	44
Size and description of the subsample	4	2
Complete wording of questions on which the release is based	18	9
Percentages on which conclusions are based[a]	98	89
Other Poll Details		
Agency conducting the fieldwork	87	82
Sample type*	1	17[b]
Sampling error	7	[c]
Nonsampling factors	1	2
Complete reporting of response categories	12	9
Sample described as representative or scientific	11	17
Nonresponse rate	2	[c]

[a]Coding the presence of all the percentages upon which conclusions were based (as specified in the NCPP principles) is virtually impossible because of problems in defining conclusions so we have coded whether any numeric percentages were reported.
[b]This information is usually restricted to noting that it was a quota sample.
[c]Less than 1 percent.
*Several British newspapers have entered into an agreement setting standards for the reporting of election polls. Asterisks indicate information which is required by those standards.

lapses do occur, and survey results are sometimes reported without adequate descriptions of sampling methods or of the questions asked. Presser, commenting on this in the June 1980 issue of the *American Political Science Review*, suggests that journals require authors to report the following information: (1) a description of the population sampled and details of how the sample was drawn sufficient to permit fairly exact replication; (2) response rates (for quota designs, the number of refusals); (3) interviewer characteristics; and (4) exact question wordings.

A review of the reporting of surveys in sociology, social psychology, economics, political science, and public opinion journals is presented in Table 3.2. This breakdown indicates that even when the authors themselves conduct the surveys they report on, they do not describe the surveys in sufficient detail to permit readers to make an independent assessment of the quality of the surveys. Even so fundamental a datum as question wording was missing from a majority of articles in all fields with the exception of Public Opinion. And even there, 45 percent of the reports lacked complete question wordings.

In summary, common practices in the reporting of surveys in both scholarly journals and the news media, and the quality of survey execution itself, leave much to be desired. The systematic evaluations that we have examined, though few, point to substantial deficiencies in the quality of the survey enterprise. Moreover, the deficiencies have been observed in areas for which there are clear standards and a wealth of theoretical and practical knowledge.

3.3 Misuses of Polls and Surveys

While it is possible to make systematic evaluations of some aspects of survey practice, it is not possible for other aspects. We have no sure way of gauging the frequency of misuses in other areas, but they are clearly of sufficient gravity to merit detailed illustration. Indeed, many misuses, such as bias in question wording, are particularly common in surveys that cover subjective questions, and the rest of this chapter discusses such misuses.

PSEUDO-SURVEYS

Solicitation and "Sugging". We commence by noting the frequent use of pseudo-surveys in the solicitation of funds and the selling of products. "Sugging," or *s*elling *u*nder the *g*uise (of a survey), involves approaching people as potential survey respondents, and using this pretense to gain their cooperation when the real intent is to sell something. Baxter (1964) found that 27 percent of respondents in a survey by the Roper organization reported experiencing such sales tactics.

It is for this reason that some major research organizations equip their field interviewers with notifications of clearance by the Better Business Bureau. Indeed, the Direct Mail/Marketing Association's guidelines of ethical business practices, which outlaw telephone "sugging," have recently

TABLE 3.2
Reporting of Information about Surveys in Basic Research Journals

Information	Percentage of Articles (reporting self-conducted surveys) that Include Information				
	Sociology Journals	Political Sci. Journals	Economics Journals	J. of Pers. & Soc. Psych.[a]	Public Opinion Quarterly
Sampling method	59.4%	49.2%	3.9%	57.4%	63.0%
Question wording	46.5	44.4	2.9	50.4	54.8
Mode of data collection	47.0	66.6	17.6	78.2	72.6
Response rate					
1949/50 & 1964/5	20.9	24.0	20.0	13.6	41.3
1979/80	51.3	36.8	4.2	22.5	63.0
Date of survey[b]					
1949/50 & I964/5	36.7	56.0	60.0	22.7	56.5
1979/80	75.6	81.6	54.2	19.7	59.3
(N)	(217)	(63)	(34)	(115)	(73)
1949/50 & 1964/5	(139)	(25)	(10)	(44)	(46)
1979/80	(78)	(38)	(24)	(71)	(27)

SOURCE: Presser, Chapter 3, Volume 2. See notes to Table 2.3 for list of sociology, political science, and economics journals.
[a]*Journal of Personality and Social Psychology*.
[b]Coded as being reported if, at least, year was mentioned.

been revised to prohibit (1) the use of the mails as well as the telephone, and (2) fund-raising activities as well as the sale of merchandise and services (see AAPOR, 1980). These additions have been necessitated by frequent attempts at direct mail solicitation, for example, enclosing a questionnaire in a solicitation appeal. This tactic is often designed to increase contributions rather than to collect a particular set of data (see *Changing Times*, 1980b:67).

One concrete example of such solicitation occurred during the last presidential election. In September, 1980, the Republican National Committee conducted a "Critical Issues Survey," using a 14-item questionnaire that included such questions as

> The Soviets now have a combat brigade in Cuba training Marxist revolutionaries for use in South America and Africa. Do you approve of Mr. Carter's decision to do nothing in response to this direct Soviet/Cuban challenge?

Accompanying the questionnaire was a four-page letter from former President Gerald Ford that expressed his dissatisfaction with various aspects of President Carter's administration and solicited contributions to the Republican National Committee in order to ". . . get the facts out about the dangerous and irresponsible programs of the Democrats to the voters in the next election." Such tendentious prose may arouse the reader to action and increase contributions, but it also may serve to give recipients a false image of survey research. Interest groups of varying political persuasions have used this technique of soliciting funds under the guise of a survey. Figure 3.2 presents another example of a similar survey from a group of scientists whose orientation is quite different from that of the Republican National Committee.

Unrepresentative Polls. Polls of this type typically provide instructions through the mass media for registering opinions by calling a particular telephone number to record one's vote. Local television and radio stations, and, on occasion, national television networks have used such surveys. However, the results of such surveys are not generalizable to any known population because of the self-selection bias of the sampling procedure, that is, the only persons included in the survey are those who choose to call in. Frequently, however, written summaries of the results of such polls neglect to include a disclaimer about self-selection bias.

An example of such misrepresentation occurred during the 1980 presidential election campaign in connection with the televised Carter-Reagan debate. Immediately following that debate, the ABC television network conducted a telephone call-in poll and announced on the basis of that un-

FIGURE 3.2

Example of a direct mail solicitation/survey.

Please return this questionnaire within 10 days to Union of Concerned Scientists in the envelope provided.

☐ I want to become a Sponsor of the Union of Concerned Scientists to help its work in controlling the dangers from nuclear power plants. Enclosed is my donation. Please keep me informed about the progress the Union of Concerned Scientists is making in its efforts.

$15 ☐ $25 ☐ $35 ☐ $50 ☐ $_____

☐ **I am already a Sponsor**, this contribution renews my support.

Your responses are important

Please take a moment to answer the questions on the back of this form and return it in the envelope provided for your convenience. Your responses are important. They will give the Union of Concerned Scientists an indication of how the American people feel about nuclear hazards and solar solutions.

By becoming a Sponsor of the Union of Concerned Scientists, you enable their scientists, engineers and lawyers to carry on desperately-needed work in •public advocacy programs •technical research on nuclear power problems and solar solutions • informational programs •legal actions to bring about enlightened energy policies in this country.

You can become a Sponsor of the Union of Concerned Scientists by checking the appropriate box on the left and returning your sponsorship donation when you return this questionnaire. It is not necessary to become a Sponsor in order to respond to the questionnaire.

Please make your check payable to the Union of Concerned Scientists. Your contribution is tax deductible.

May we have your answers to the following questions

1. Do you think that there are serious risks associated with nuclear power?

☐Yes ☐No ☐No Opinion

(If you answered "yes", which risk concerns you the most?)
☐major reactor accident
☐nuclear weapons proliferation
☐radioactive waste disposal ☐Not sure

2. Do you favor a stronger effort to develop solar energy alternatives to nuclear power?

☐Yes ☐No ☐Not sure

3. Do you think the government can be trusted to make this effort?

☐Yes ☐No ☐Not sure

4. Do you feel that the industrial establishment will turn on its own from nuclear to solar energy?

☐Yes ☐No ☐Not sure

5. Which national nuclear energy policy do you favor?

☐increased construction of nuclear power plants
☐continued construction with a slow investigation of solar options
☐a halt in nuclear construction and a significant solar effort
☐phase out of existing nuclear plants and major solar effort
☐immediate shutdown of all plants
☐not sure

6. Do you favor expanded government research to develop alternative energy sources (such as biomass, wind power, tidal power)?

☐Yes ☐No ☐Not sure

7. Do you believe that the work of the Union of Concerned Scientists, as described in the enclosed letter, is necessary?

☐Yes ☐No ☐No opinion

Please return this form in the enclosed envelope. Be sure to check the box on the front of this form if you want to become a Sponsor of the Union of Concerned Scientists.

UNION OF CONCERNED SCIENTISTS 1384 Massachusetts Avenue, Cambridge, Massachusetts 02238

representative survey that Reagan had won the debate.[7] While the National Council on Public Polls (1980a) condemned ABC's actions, the network did not make a clear retraction of its survey findings.

Many newspapers and magazines subscribe to a related version of this practice—instead of providing phone numbers to call, the newspaper (or magazine) includes a "coupon" (or questionnaire) soliciting views on one or more subjects. The reader fills in the coupon and mails it to the newspaper's office. The results of such exercises suffer from the same self-selection bias and consequent lack of generalizability as telephone call-in surveys. The QUBE polls described in Chapter 2 also have analogous deficiencies.

QUESTION WORDING AND QUESTION BIAS

Failure to properly account for the crucial factor of question wording can cause one to draw misleading inferences from survey data. And for the subjective domain, knowledge about the effects of question wording is often lacking. So, for example, criticisms of question "bias" or the "fairness" of question wording persistently arise in opinion polls, but there is no theory, accepted procedure, or standard for dealing with such criticisms (see Chapter 5).

The issue of question wording poses a particularly pernicious problem to the extent that the public considers polls to be unofficial elections. While this certainly was the prevailing ideology 40 years ago, it has become less dominant today. Currently, partisans in public disputes often commission "public opinion" polls (that is, advocacy polls) in which the wording of questions reflects the positions of their sponsors: for example, in the debates over the SALT treaty (Lanouette, 1980) and abortion (Marsh, 1982: 143–144). While such practices have sometimes been a subject of enforcement actions by professional organizations, these actions are fraught with problems. At a minimum, such actions presume both that there is an "unbiased" way of wording any question and that there are objective methods for finding the appropriate wording.

Question wording is an area in which few rules exist. One of the few generally recognized rules is that the results obtained in any survey depend on the exact questions that are asked: different questions on the same topic —even questions that are logically equivalent—may sometimes produce quite different results (see Chapter 5). It is therefore important for any serious program of subjective measurement to ask a sufficient number of different questions to permit an assessment of the effects of question wording on the measurements.

Some well-known problems arise in the reliance upon a single question. For example, if a question presents respondents with a problem and only one solution to it, there may be a heightened tendency for respondents to

subscribe to the one solution offered. Thus, responses to such questions are misleading in that they do not provide evidence about other options that might also have been endorsed. Another dubious practice is asking double-barrelled or two-part questions. Such questions obfuscate all inferences about the respondents' true opinions due to lack of knowledge about which part of the question is being answered.

Two examples of these abuses of the survey method are presented as illustrations.

Americans' Support for the Consumer Protection Agency. One of the two formal complaints under the AAPOR standards was made in the spring of 1975 when the executive director of the Consumer Federation of America complained to AAPOR that a widely publicized survey conducted by an AAPOR member organization contained questions that were "slanted" and whose results "verge[d] on fraud." Members of a Senate committee also attacked the survey as "loaded to produce a predetermined conclusion."[8] The surveys[9] was sponsored by the Business Roundtable, a lobbying group comprised of chief executives of major American corporations. It inquired about pending legislation to set up a federal Consumer Protection Agency (CPA); the legislation was widely opposed in the business community.

The key and widely reported question in Business Roundtable's survey read

> Those in favor of setting up an additional federal consumer protection agency on top of the other agencies say it is needed because the agencies we have are not getting the job done themselves. Those who oppose setting up the additional agency say that we already have plenty of government agencies to protect consumers and it's just a matter of making them work better. How do you feel? Do you favor setting up an additional consumer protection agency over all others, or do you favor doing what is necessary to make the agencies we now have more effective in protecting consumer interests?

This question was preceded by a long set of other questions asking about other consumer-oriented federal agencies, such as the Consumer Product Safety Commission, the Office of Consumer Affairs, and so on. Proponents of the Consumer Protection Agency complained about the bias of this question, but the results ("75% opposed") were widely cited as evidence that the public did not want to have an agency established. Thus the *Wall Street Journal* used the findings to argue that the "vast majority" of Americans do not want a consumer protection agency. Reacting to complaints of question bias, the *Wall Street Journal* (June 2, 1975:10) editorialized that

> . . . [this] question does precisely state in terms everyone can understand, the arguments for and against a CPA. What the consumerists object to is the fact that [the survey organization] did not load the question in favor of the CPA. . . .

In response to complaints, AAPOR established a committee to determine whether or not the survey organization had violated AAPOR's code of standards. The committee reported (American Association for Public Opinion Research, 1976) that "the questions used in the survey have some degree of bias in the direction of obtaining anti-consumer advocacy or anti-bureaucratic response," but it concluded that there was no evidence that the bias was deliberately intended. In its discussion of this finding, the committee noted that the use of the words "additional" and "on top of" in the question probably sensitized respondents to the question of enlarged government activities, that the question oversimplified the issue by implying that there was a simple binary choice between either making the other agencies more effective or setting up a new agency, and that the preceding questions concerning other federal agencies conveyed the impression that a number of agencies already existed to protect the consumer.[10] In its discussion, the AAPOR committee observed that bias can inadvertently enter into questions from (1) ingrained beliefs of the question writer, (2) a lack of consultation with individuals having a different viewpoint, or (3) failure to conduct adequate tests of questionnaire objectivity. This statement of the committee attests well to the present lack of knowledge about the survey measurement of subjective phenomena.

It is surely a truism that all people have their own particular frames of reference that they bring to any question or topic. Consultation and review may help people sharpen their understanding of the different viewpoints that may exist, but it is difficult to know when enough consultation has been done or when the consultants are sufficiently diverse in view. Nor is it obvious what the result of such consultation should be. Finally, it must be noted that "tests of objectivity" mentioned in various standards and reports are, to our knowledge, almost never used in contemporary survey practice —although they (or something akin to them) remain as feasible today as in 1947 when they were proposed by Dodd. (We would not, however, wish to endorse the accolade "tests of questionnaire objectivity" to describe such procedures.)

Indeed, it would seem from the foregoing example that the major difference between Dodd's 1947 proposal in this regard and present practice is that now professional discussions of question "bias" occur only when (and if) the public outcry is sufficient to generate a complaint to a professional organization. Such issues have not become, as Dodd proposed, a routine topic for study during pretesting. This is certainly a lamentable situation, and—given the growth in size and influence of the survey enterprise—it should be a cause for grave concern among those interested in the future of the enterprise (see Recommendations 2 through 6 and 9, Chapter 10).

Support for Wage Limitation in Britain. Our second example comes from Great Britain. In 1977 Margaret Thatcher, then leader of the British

Conservative party in opposition, said that if there was a strike threatened by a major union while she was prime minister she would treat it as a confrontation between that union and the people, and "let the people decide" what should happen through a referendum or poll on the subject (London *Times*, September 19, 1977). (The *Daily Express* soon announced that an opinion poll had shown that two-thirds of British adults thought that this suggestion was a "good idea," although such a question may only show that people support the idea of referenda in general [see Kornhauser, 1946–47].) Such a confrontation occurred in the autumn of 1980, when manual workers in the public sector challenged the government's ceiling of 6 percent for pay increases. As described in Chapter 2, the British press made wide use of polls on wage limitation during the part of the summer in which we were studying the British media.

The closest thing to a referendum on the subject was a poll conducted by Opinion Research and Communication (ORAC) in early November, which claimed to show that a large majority of the British people supported the government ceiling on wage increases (the fullest report of this poll can be found in the London *Times* of November 11, 1980). Several questions were asked to elicit views on the pay policy. The first question was

> The Government believes that in the interest of fighting inflation, local authority [municipal] workers should get no more than a 6% pay rise this time. Do you agree with this view or do you think they should get more?

The responses to this question are shown in Part A of Table 3.3.

This question is a classic example of Kornhauser's (1946–47) criticism of polls about union activities: a commonly agreed evil, inflation, is pitted against a single solution—limitations on workers' pay. The assumption that a 6 percent ceiling on pay increases will "fight inflation" is not the subject matter of this question; that premise is taken for granted. The question is one-sided inasmuch as only the government's point of view is presented. To see what this means, it is instructive to change the question so that it is one-sided in the opposite direction, that is, to present only the union viewpoint. Consider, for example, the question:

> Inflation last year increased the cost of living in Britain by almost 20 percent. The public sector unions believe that local authority workers should get a 20 percent pay rise so that they will be able to afford to buy the goods being produced by British industry and help stop the recession. Do you agree with this view or do you think the government is right to try to cut the living standard of these workers?

It is likely that the responses to this question, if it had been asked, would have been rather different.

TABLE 3.3
Published Report of Responses to Survey by Opinion Research and Communication (G.B.)

A. Question: "The Government believes that in the interest of fighting inflation, local authority [municipal] workers should get no more than a 6% pay rise this time. Do you agree with this view or do you think they should get more?"

	Trade Union Members	Non-trade Union Members	All
Should accept 6%	46%	62%	56%
Should not	49	31	38
Don't know	5	7	6

B. Question: "The Government has said that if local authorities [governments] do pay more than 6%, rate-payers [taxpayers] will have to foot the extra bill. Do you think local authorities should go to rate-payers to pay extra or should they stick to the 6% pay limit?"

	Trade Union Members	Non-trade Union Members	All
Ratepayers extra	27%	18%	21%
Stick to 6%	55	72	66
Don't know	18	11	13

NOTE: Base Ns for percentages were not given in the news report.

In addition to the initial question, respondents in the ORAC survey were also asked:

> The government has said that if Local Authorities [municipal governments] do pay more than 6%, rate-payers [taxpayers] will have to foot the extra bill. Do you think Local Authorities should go to rate-payers to pay extra or should they stick to the 6% pay limit?

The responses to this question are shown in Part B of Table 3.3. Every newspaper that reported the poll (except the *Guardian* and the London *Times*) quoted the findings from this second question as showing that two-thirds of the public supported a 6 percent limit on wage increases, failing to mention that tax increases were posed as the only alternative. Indeed, most newspapers merely repeated the verdict of the ORAC press release, which claimed that the results showed strong backing for the government's policy of wage limitation.

Poor reporting, however, is not the major point at issue. Both questions

are quite one-sided. The act of answering either of them commits the respondent (seemingly) to endorse the assumption that wage rises are a cause of inflation and that the choice is between public sector pay rises or tax increases. To counter the bias of these questions, alternative questions, either more balanced in their presentation or incorporating opposite biases, should have been asked and reported. The responses to such alternative formulations could have been compared to those obtained with the original formulations in order to assess the extent to which the survey results were affected by the slanted wording of the original questions. Unfortunately, this was not done, and the extent of the bias in the results of the poll is only a matter for speculation since no other evidence is available.

Exposing Question Bias. A related, though not exactly similar, example in the United States illustrates one mechanism for combatting the effects of one-sided question wording in a highly publicized survey. In 1978 the CBS/*New York Times* poll conducted a study on the "New Right" in which respondents were asked whether various activities were appropriate for the federal government, for example, helping people obtain medical care, providing jobs for unemployed workers, and so forth. The results of the poll were interpreted as showing that conservatives were moving to the left, that is, toward more liberal positions. The North American Newspaper Alliance (NANA) was unhappy about the one-sided solutions offered by the CBS/*New York Times (NYT)* questions and commissioned their own survey.

The NANA split ballot experiment included *both* the questions as presented by the CBS/NYT poll and analogous questions with the words "private enterprise" substituted for "federal government." The first half of the experiment confirmed the results of the CBS/*NYT* poll and the second half contradicted it (see Table 3.4). Using the logic of the original analysis, the NANA results for self-described "liberals" would have required the interpretation that liberals were moving to the right, toward more conservative positions.

We cite this example to illustrate the problem inherent in wording questions so that they provide only one solution to a problem and to commend this method of examining the effects of questions by split ballot experiments built into surveys.

At a more general level we note that while questions can often be criticized, it is misleading to imply that there is a single "proper way" to ask a question. Any question can be considered limited from some point of view, and single questions are inherently inadequate to convey a picture of any subjective phenomenon of importance. For this reason, a range of questions should be asked and the sensitivity of measurements to wording effects should be assessed (see our Recommendation 9, Chapter 10). Moreover, serious efforts using formal measurement models to assess attitudes should

TABLE 3.4

Published Results of Poll by North American Newspaper Alliance (1978)

Question Wordings and Percentage Agreeing

Medical Care

1. "Do you agree or disagree that the *federal government* ought to help people get medical care at low cost?"

	Overall	*Conservatives*
CBS/NYT	81%	79%
NANA–1	81%	74%

2. "Do you agree or disagree that *private enterprise* ought to help people get medical care at low cost?"

	Overall	*Liberals*
NANA–2	71%	73%

Employment

1. "Do you agree or disagree that the *federal government* ought to see to it that everybody that wants a job can get one?"

	Overall	*Conservatives*
CBS/NYT	74%	70%
NANA–1	70%	72%[a]

2. "Do you agree or disagree that *private enterprise* ought to see to it that everybody that wants a job can get one?"

	Overall	*Liberals*
NANA–2	70%	72%

Race Relations

1. "Do you agree or disagree that if blacks are not getting fair treatment in jobs and housing, the *federal government* should see to it that they do?"

	Overall	*Conservatives*
CBS/NYT	68%	64%
NANA–1	66%	60%

2. "Do you agree or disagree that if blacks are not getting fair treatment in jobs and housing, *private enterprise* should see to it that they do?"

	Overall	*Liberals*
NANA–2	56%	61%

Table 3.4 notes appear at foot of page 84.

be vigorously explored (see Chapter 6 and Recommendation 14, Chapter 10).

SELECTIVE REPORTING OF POLL RESULTS

Earlier in this chapter we noted that the proportion of commercial survey activity that is published is tiny. The vast majority of such surveys are market research studies, which might be of interest to one's competitors, but would hardly excite most people. However, one should be aware of the considerable power that the sponsor of such research has in deciding whether or not to release the information collected or in deciding which of the questions to include in a press release. Some codes of standards explicitly prohibit selective reporting when it is misleading, but the right to release the findings *at all* rests with the organization that commissions a survey. Moreover, the fact that there is a code that prohibits selective reporting of results does not prevent it from happening—particularly when the survey sponsor is not a member of AAPOR, CASRO, or NCPP. Let us consider one example.

During the last decade, corporate advertisements on general political and social topics have appeared with some frequency in the American media. Such advertisements sometimes report the results of opinion polls claiming to show that the public supports the particular views espoused by the ad's sponsor. For example, in 1980–81 Union Carbide published the results of a series of surveys that, the company claimed, showed that the public supported government budget cutbacks to aid economic growth (see Union Carbide, 1980a, 1980b, 1981). In several prominently published newspaper advertisements, Union Carbide reported survey results that purported, among other things, to show that:[11]

> . . . 69 percent of the public would favor tax cuts [for business] if business said that some of the money saved would be used to buy new equipment and create additional jobs.
>
> Sixty-nine percent also favored increased tax credits for new capital investments, and 57 percent favor larger tax deductions for prior investments by valuing equipment at replacement—rather than historical—cost.
>
> . . . by more than two to one, the public favored the reduction of restrictions

TABLE 3.4 *(continued)*

NOTE: Table shows results of Opinion Research Corporation personal interview survey (samples of 1,035 [NANA–1] and 1,075 [NANA–2] American adults) interviewed between May 11–30, 1978, and results of CBS/*New York Times* (NYT) survey (N = 1,599). (CBS/NYT survey was released in January 1978; interviewing dates were not supplied in NANA report.) Our figures were taken from text of NANA (1978); accompanying table contained apparent labelling errors. Base sample sizes for percentages were not given in NANA report.

[a]Figure is quoted by NANA (1978), but their text suggests there may be a typographical error; the text reads: "65 percent of those who called themselves 'very conservative' and 59 percent of those who called themselves 'moderately conservative' agreed. Overall, 70 percent of all conservatives agreed."

which prohibit companies from joint development of export markets and R&D projects.

These findings were used to support the advertisement's subsequent proposal for passage of a Capital Cost Recovery Act to liberalize businesses' depreciation schedules in order to reduce business taxes, for relaxation of government regulation of industry, and for the provision of government aid to business in order to spur exports. In its own publication, Union Carbide (1980a:IV) concluded that the results of its first survey showed

> The American public overwhelmingly attributes [the] dim economic future to a cluster of forces . . . . But while all these forces are blamed for lagging growth, it is clear that the government is seen as the main institutional cause for the decline in our economy.
>
> Government is seen as being wasteful, bloated, inefficient, and costly and as an inherently negative factor in economic growth. Americans believe that government programs should be cut back . . .

However, the survey results that were *not* reported in the advertisements reveal a somewhat different picture.[12] The reports of the surveys in the advertisements were selective in several important respects. First, they neglected to point out that 66 percent of the respondents in the first survey thought that "businessmen are more afraid to take risks than they used to be" and that 60 percent also thought that "a few big companies divide up business between them" (Union Carbide, 1980a:19). Both of these factors were believed to be responsible for slowing growth by a substantial majority of respondents. Furthermore, more than 66 percent of the respondents agreed with each of the following statements (Union Carbide, 1980a:26):

> When the economy is booming, most of the increase is kept by business in the form of higher profits, higher executive salaries, and fat expense accounts. The average person does not benefit much.
>
> When the economy goes down, business keeps its profits high and cuts jobs and wages.
>
> All companies need to make some profits to stay in business. But the profits that businesses are making today are much higher than they need to be.[13]

And, far from disapproving of government regulatory activity, the surveys seemed to evidence considerable public support for regulation by the government in many spheres. When asked, for example, whether "the federal government is doing more than it should, less than it should, or about the right amount to regulate major corporations in areas like product safety and other things that have to do with protecting the public," 70 percent of the

respondents replied either "about right" or "less than it should." When asked about ten specific types of government regulation, a majority of respondents supported present or even stricter government regulations. (See Appendix E for complete wordings and response distributions for these questions.)

There are other criticisms that might be made of the Union Carbide surveys: for example, it might be asked whether most people hold meaningful opinions about whether tax deductions for prior investments in equipment should be valued at replacement rather than historical cost. However, the important point to be made is that Union Carbide used its survey results as a petition (albeit a very expensive type of petition) to lobby for its point of view by presenting only those results that favored the company position. Obviously, supporters of increased government regulation could have written a very different advertisement on the basis of the same survey results.

It is not just the sponsor of a survey who acts as a gatekeeper in reporting survey results. The media which report the results must make a judgment about the newsworthiness of a survey. A systematic study of the way in which three major Swedish newspapers refer to public opinion has been conducted by Holmberg (1980). In an analysis of the editorial content of the papers over the last 50 years, he found, as one might expect, that an assertion about public opinion was more likely today than in the past to be grounded in objective evidence (that is, a poll or survey). But, disconcertingly, he also found that newspapers are much more likely to report poll results that showed a swing of opinion toward the candidate (or party) supported by the paper.

We conclude by noting that advocacy polling of the sort exemplified by the Union Carbide surveys is becoming more common (see O'Neill, 1980, for an advocate's descriptions of its use and problems). Mitchell (1980b: 88–92) cited a recent and particularly flagrant example that appeared as a two-page advertisement in the *Washington Post*. The ad by the American Nuclear Energy Council (an industry trade group), addressed to "the President & Congress," used poll data to argue against a moratorium in reactor licensing that was being considered in the wake of the Three Mile Island accident. Mitchell not only raises the issue of selective reporting in regard to this example, but he concludes that the use of public opinion data in the ad involved "misrepresentation and outright error." In light of the growth of advocacy polling, we suspect that abuses of this type may become an increasing problem in the future.

DEFICIENCIES IN EVALUATION AND POLICY RESEARCH

The abuse of survey methods is not restricted to the commercial or journalistic arenas, as it might seem up to this point in our narrative. There are

important instances where poorly conducted or even overtly biased research is used as input for important legal or policy decisions, even with nonsubjective survey measurements. For example, in the trial of a civil suit on the results of the 1980 census (*Carey et al.* v. *Klutznick et al.*, 80 Civ. 4550), the state and city of New York cited a poorly conducted survey as evidence of their claim that the Census Bureau had severely undercounted the population of New York City. Yet according to several expert witnesses, the cited survey itself was compromised by an incomplete sampling frame and an inadequate sampling plan, the questionnaire did not document the existence or extent of an undercount, and the report of the survey results did not include details (for example, sample size, response rate) that would be required to permit an accurate assessment of its quality (see Bradburn, 1980; Bureau of the Census, 1980d).

As we noted in Chapter 2, one currently burgeoning area of survey research involves program evaluation. The growth in popularity of evaluation research has been meteoric, with a flood of books about it, the creation of journals, and the establishment of college courses on the subject. Yet the quality of the surveys that are often part of such research is sometimes a cause for concern.

One such evaluation survey was conducted for the U.S. Department of Transportation in order to assess the impact of the landing of the supersonic Concorde at Dulles airport in Washington, D.C. This survey, conducted in communities surrounding the airport to measure public perception of the noise and vibration caused by the airplane, was seriously flawed in several respects. Most importantly, the samples were inadequate for the purpose intended. In the words of a panel of the National Research Council (1977: 13–14) that reviewed this survey:

> The primary design shortcoming of the Dulles survey was in the sampling. The survey respondents around Dulles are not a probability sample of the population surrounding the airport. Towns were selected purposively not randomly, and some towns and rural areas are not in the sample at all . . . .
>
> The design of the selection of respondents within the designated communities also creates ambiguity and makes it difficult to define the actual universe sampled.

A review by the General Accounting Office (1978:26–28) pointed out this flaw as well as inadequate training and supervision of interviewers, among other problems.

A Common Deficiency. The foregoing example illustrates a case where poor results arose from poor practice, *not* from a deficiency in knowledge. There are, however, more fundamental problems in the use of subjective

measurements in such surveys. For example, a survey used in the evaluation of the Emergency School Assistance Act (ESAA) in the early 1970s (National Opinion Research Center, 1973; Crain and York, 1976) suffers from none of the sampling and training problems that plagued the Concorde evaluation. The survey met contemporary standards of practice in these areas. Besides being technically sophisticated in sampling, the study undertook careful pretesting to ensure that respondents could read and understand the questionnaire items, and the data analyses were elaborate and often ingenious. The problems this survey suffered are rarely resolved in any survey, and they thus serve to illustrate validity problems commonly afflicting subjective surveys.

The essential question in this evaluation was whether or not ESAA-funded programs met their goal of assisting schools that were implementing court-ordered desegregation programs. A primary goal was to improve race relations in newly desegregated schools. The survey therefore included a large number of subjective measurements that were to be used as indicators of the program's impact. The evaluation compared responses of students in (randomly selected) schools that received ESAA funds to those of students in (randomly selected) schools that did not receive ESAA funding (see Gilbert, Light, and Mosteller, 1975, and Acland, 1979, for reviews of this research).

The fundamental problem with this evaluation survey was that there was no empirical evidence to substantiate the assumption that pupils' responses to the attitude questions provided valid measurements of the phenomenon being studied (race relations). Ideally, one would have liked to know whether pupils' responses to such questions as the following were empirically linked to independent (especially nonverbal) measures of the quality of race relations:[14]

> If you could choose the kind of school you would go to would you pick one with: (1) all white students, (2) all black students, [or] (3) a mixture of different kinds of students?
>
> In general, do you think that white people are smarter than black people, that black people are smarter than white people, or do you think that a person's color doesn't have anything to do with how smart he is?
>
> How uncomfortable do you feel around students of a different race? (1) generally very uncomfortable, (2) generally somewhat uncomfortable, (3) occasionally somewhat uncomfortable, [or] (4) not at all uncomfortable.

Naturalistic observations of such things as the patterns of seating in lunchrooms, invitations to after-school social functions, and so forth might have provided helpful auxiliary data in this case. These data would have assisted in the evaluation of the social climates of the schools, and, since they did not depend on self-reports, would have allowed some explicit validity tests

for the survey measurements. For example, one might expect to find an association between pupils' reports of discomfort in interracial social settings and the actual frequency of such interactions. Unfortunately, validity data of this sort were not gathered in the ESAA study.[15] Hence any conclusions based on pupils' responses to the survey questions are vulnerable to claims that the subjective measurements might be, for example, ephemeral responses of pupils to the situational demands of (personally inconsequential) questions posed by strangers in an environment saturated with cues as to the socially desirable answers. Lacking empirical evidence of the meaningfulness of the subjective measurements, one could do little in reply to such claims except protest to the contrary.[16]

In singling out the ESAA evaluation survey for a critique, we wish to emphasize that we are not criticizing a distinctive fault or a practice unique to that survey. In contrast to research and evaluation studies that use cognitive tests, survey research studies do not ordinarily provide external validation of their subjective measurements and hence the meaningfulness of their measurements (and thereby the resulting inferences) are open to challenge. This deficiency is to be lamented and ought to be remedied by all those who wish to use subjective survey measurements to evaluate social programs, or for any other substantive purpose. External validation of measurements of subjective phenomena is often possible and is always desirable. (See Recommendation 12, Chapter 10, on putting surveys in context with other data.)

DEFICIENCIES IN BASIC RESEARCH STUDIES

Readers should not infer from the foregoing examples that basic scientific research using surveys (or surveys conducted by academic researchers) are immune to the problems detailed in the foregoing examples. We noted in Chapter 1 a survey of university faculties that has been criticized for its shortcomings, and Table 3.2 details the deficiencies in reporting survey results in academic journals.

In the interest of balance, we conclude by citing the example of a survey published by the *American Psychologist*.[17] The July 1981 issue of this journal published an article entitled, "Survivors of the Sixties: Comparative psychosocial and political development of former Berkeley activists." The abstract of this study reads (in its entirety):

> Former Berkeley free speech movement activists were compared with student government contemporaries and a cross section of the 1964–1965 student body along the dimensions of self- and ideal-self-constructions, locus of control, political beliefs and behavior, occupational choice, and moral development. Though their personality and political activism profiles were not distinctive, former activists were more likely to endorse leftist convictions, to work in human service and

creative occupations, to earn lower annual incomes, and to demonstrate principled moral reasoning than were comparison groups. Though free speech movement activists appear to have made important life transitions 15 years following their arrests, they still remain a distinct social and political cohort.

The conclusions were based on a survey of former Berkeley students, and the author notes that these conclusions are constrained "by certain methodological weaknesses" among which is a "modest rate of return" for the survey. This caution, however, is something of an understatement. The population sampled in the survey consisted of all students arrested in a Berkeley sit-in on December 2, 1964. The report indicates that more than 800 people were arrested, a total that may include some nonstudents. Of those arrested, 111 were located by the researcher and only 30 completed the survey.[18] Even if one assumes that half of those arrested were nonstudents, the noninterview rate for this survey would be more than 90 percent. We can only wonder what standards the editors and referees of this influential journal were using when they decided to publish this survey report.

3.4 Conclusions

Existing standards for the conduct and reporting of surveys appear to have had a limited impact on survey practice. When systematic evaluations have been possible, we find that current practice leaves much to be desired. We also find numerous examples of inadequate practice in areas in which it has not been possible to undertake systematic reviews.

Inevitably one must ask how to improve this situation. One possible strategy for improving survey practice would be to enforce existing standards more rigorously. We note, for example, that the American Association for Public Opinion Research has undertaken only two major proceedings under its 1968 code of practice. Certainly, other actions might have been taken, and an activist approach to exposing misleading or inadequate practices might have had a salutary effect on survey practice.

There are, however, grave problems with such an approach. First, codes of professional practice bind only the members of a professional organization. Practitioners who do not belong to the professional organization—or who choose to leave—will not be subject to review or sanction. Hence, an activist approach in a situation in which membership in a professional organization is not required by law or custom may have the effect of policing those who are least in need of policing. It leaves the unscrupulous free to ply their trade unmolested.

As an alternative strategy, a market for quality survey research might be

created. In Chapter 10, we make several specific recommendations in this regard. We urge in Recommendation 2 that the survey research community encourage more competent and informed public discussion of the faults and merits of polls and surveys. Such discussion would, we hope, lead to the improvement of survey practice by creating a market for better surveys. When the consumers (and funders) of survey research appreciate the effects of question wording, for example, and begin to demand estimates of these effects in survey reports, substandard practices in this regard may go the way of the *Literary Digest* polls. (Which is to say that they will become less common but not disappear entirely).

In the same vein, establishment of a center for the study and evaluation of survey methodology might have a salutary effect (see Recommendations 13 and 7, Chapter 10). If such a center were to take an active role in evaluations of surveys such as that exemplified by the split-ballot rebuttal poll (shown in Figure 3.3) and the analysis of sample composition (shown in Figure 3.1), it could serve a useful role in informing the public of the effects of deficient survey practice and thereby create a more discriminating public. In the long run, such activities would help educate the public to discriminate among surveys according to the quality of their research methods. In such a milieu, regulation might be unnecessary since deficient practices would find few patrons.

Notes

1. AAPOR had had a brief and general code from 1950. The greatly expanded 1968 code was adopted when Congress was considering and held public hearings on legislation to regulate polls (see Wheeler, 1976:250).

2. This description of the CASRO charter is taken from an undated CASRO brochure entitled: "CASRO: An organization of and for the survey research business person." Quotations from the CASRO code of standards are from the amended 1982 edition; the original code (which bears no date) was adopted in 1979; it is identical to the 1982 version save for one clause.

3. In the wake of NCPP's adoption of these principles, two organizations (Cambridge Survey Research and NBC News) resigned their membership in NCPP. The resignations were accompanied by a statement from Cambridge Survey Research expressing concern over potential conflicts between contractual obligations to maintain in confidence information about its clients' surveys and concern over the wide powers granted to the Committee on Disclosure. NBC News issued a statement:

> These [NCPP] procedures involve a process by which anyone who objects to a public opinion poll may force the organization which conducted the poll into binding arbitration.

NBC News does not believe that the best interests of a free press are served by subjecting journalistic organizations to binding arbitrations concerning news judgments.

These resignations may raise some questions about the deterrent value of NCPP's primary sanction for enforcement of their principles—expulsion from the organization—since two organizations felt free to resign from the organization. (In contrast to NBC's position, the CBS network and the *New York Times* subscribed to the NCPP principles on disclosure.)

4. Of course, situations analogous to those in the private sector do arise when one government agency sponsors a survey conducted by another agency (often the Census Bureau) or when agencies use private contractors to conduct surveys.

5. The exceptions are surveys conducted by regulatory agencies, which are reviewed by the General Accounting Office, and those done for Congress or the judiciary, which are not subject to review.

6. See Chapter 2 and Appendix C for a description of the sample in our newspaper study.

7. Besides the self-selection bias, there were other deficiencies in that poll. For example, telephone callers from certain sections of the country were more likely to encounter busy signals.

8. The information and quote come from the report of an ad hoc AAPOR committee appointed to investigate these complaints (American Association for Public Opinion Research, 1976), and from *Advertising Age*, April 7, 1975.

9. Actually, the results were based upon a set of questions included in an omnibus survey; we refer to it as a survey for convenience.

10. The AAPOR committee also concluded that the survey organization "did make interpretations which do not appear to be supported by the data made public" and hence was in violation of the AAPOR code. The committee found that the survey showed that only 22 percent of the sample "claimed" that they had heard of or knew anything about the purposes of the CPA, and that "just 1 percent expressed a reasonable understanding of what the [CPA] was all about." Thus, most respondents were required to voice an opinion based solely upon the "faulty or incomplete" information provided in the text of the cited question (American Association for Public Opinion Research, 1976).

11. The quotations are from a full-page advertisement which appeared in the *Washington Post* on November 12, 1980 (page A9); other advertisements appeared on May 4, 1980 (page C4), April 30, 1981 (full page A9). Some advertisements were also placed in other newspapers; see, for example: the *New York Times* November 12, 1980 (full page A32), April 30, 1981 (full page B20), and the *Wall Street Journal* November 12, 1980 (full page 9).

Subsequent to the preparation of this section, Kinsley (1981) published a journalistic account and critique of Union Carbide's surveys.

12. The detailed reports (Union Carbide, 1980a, 1980b, 1981) were mentioned in the advertisements, and they were made available (free of charge) to any reader who wrote for a copy. The first survey was conducted by Roger Seasonwein Associates; two subsequent surveys were conducted by Cambridge Reports.

13. As Appendix E makes clear, the responses on this topic are not simply "anti-profit."

14. In particular, given that the measurements were made after school integration had taken place, one would like to assess the extent to which students' answers to the questions asked by the outside evaluation team might reflect a parroting of the "correct answers" to "test" questions rather than providing real indicators of students' personal feelings about integration (especially since the questionnaires had a space for recording pupils' names.)

15. Other data were collected from school counselors and principals and were an important part of the evaluation, but they were not used for the purpose of validation.

16. A further problem with the survey measurements in the ESAA study arose from the fact that response categories were arbitrarily assigned numeric scores that were then summed (across several items) into a composite index. Such ad hoc procedures are a less than optimal way of treating responses that have discrete distributions and that may have complex patterns of multivariate association, for example, response consistency effects across items (see Chapter 6 and also Clogg, Chapter 11, and Duncan, Chapter 12, in Volume 2). In making this comment, we emphasize that the use of multiple indicators is desirable practice although its execution was not optimal in this instance.

17. This journal is circulated to over 50,000 subscribers. According to a statement of editorial policy published December 1981 (pages 1,473–1,474), it is intended, among other things, to "present the status of substantive areas of psychological knowledge in a form and style suitable to the general membership" of the American Psychological Association.

18. Nonresponse was a severe problem in the other groups as well. In the "university cross section" group, responses were obtained from 25 of 253 persons in the original sample. And while the article does not indicate the original size of the student government sample, it does note that addresses were found for 94 persons in this group, and responses were received from 28.

PART III

MEASUREMENT ISSUES

All measurement is befuddled with error. About this the scientist can and does do something; he ascertains the possible extent of the error, determines whether it is constant (biasing) or variable, or both, and ever strives to improve his instruments and techniques.

QUINN McNEMAR (1946:294)

Primary responsibility for original drafting of this part of the panel's report was assumed by Abelson (4.3), Bailar (4.2), Duncan (6.1–6.4), Frankel (4.2), Kruskal (4.1), Schuman (5.1), Smith (5.2), and Turner (5.2, 5.3). See Note on Authorship, page vi.

4

Measurement and Error: An Introduction

4.1 Some Fundamentals

"What is truth? said jesting Pilate; and would not stay for an answer." Thus Francis Bacon opens his essay "Of Truth." Yet we cannot scoff at truth in trying to deal with measurement and error. We must begin with some sense or understanding or even faith that there is a true quantity, a true value under measurement, for error means discrepancy from truth.[1]

In the physical sciences, it is usually reasonable to think of real lengths, times, masses *out there*, so to speak, that are measured better and better as measuring equipment improves and as more resources are expended on the measurement process. Yet even in physics, especially at the boundary of research, there are ambiguities and difficulties with the concept of a true value. The Heisenberg uncertainty principle says that it is not possible to measure simultaneously with arbitrarily high precision some connected quantities. The writings of Youden, Deming, Eisenhart, and others give examples of measurements of the same quantity (for example, the speed of light) by different investigators or methods that differ far more than would be expected from intrastudy estimates of precision. There is no reason to

expect that *all* sources of systematic and random error can ever be known; the more refined our measurement method, the more fresh sources of error come to our attention.[2]

In the social sciences many variables are relatively objective in the sense that they appear at first to be measurable to an arbitrary degree of accuracy; for example, age, sex, marital status, income. Closer investigation, however, shows that all of them are more or less intrinsically blurred, and blurred in different ways, such as ambiguity of concept, inevitable measurement error, and the like. For example, age, in principle, can be verified by checking birth certificates, but some people do not have birth certificates, some certificates have been destroyed by fire or other loss, some contain slips of the pen, and there are bound to be identification problems in some cases. Other social and psychological characteristics correspond to concepts that are less clear even at the very first, and most of the so-called subjective characteristics fall under this heading, for example, attitudes, intelligence, confidence, opinions, and so forth.

Although all measurement is subject to ambiguity and error, especially as one nears current research frontiers, it is natural, most of the time, to act as if there were a true value to which our fallible observations provide more or less accurate approximations.[3] This approach gives little trouble for the weight of a potato, but for perceived happiness the notion of true value is certainly questionable. One can, however, compare two or more ways of trying to measure the same underlying characteristic, and insofar as different methods give the same results (but differ in their operations in nontrivial ways), that robustness suggests the measurement of the same underlying variable (see examples in Chapter 5).

There are several ways to approach the theoretical and practical difficulties of true value. One may simply state flatly that there is a true value out there, a value that one tries to get at via inevitably fallible measurement. At the other end of the scale, there is a cluster of relativistic or solipsistic positions that hold that *no* true value exists out there—indeed, there may be no out there to begin with. This report can hardly resolve the age-old question asked by Pilate about truth. (Perhaps in the *next* report . . . !) So we often proceed more or less in a pragmatic vein by comparing different modes of measurement. Sometimes one such mode is so much better (whatever that means) or so much more highly regarded than others that the first can serve as an operational substitute for true value, at least pro tempore. Such special modes have been called "preferred" or "exemplary" by some writers. (There are, of course, problems with this agnostic, workaday position; for example, since it is impossible to specify fully a method of measurement, any method will be at least slightly different from day to day and from person to person.)

OBSERVATION ERROR

When it makes sense to think—or to pretend to think—about true value, a good general way to regard observation error is to consider, at least abstractly, the joint behavior of true value and observed value.[4] Thus, for example, if the variable is age in years, one can imagine (for a given method of measurement) that attached to each person's name in the relevant population is a *true* age and also a distribution for *observed* age. For example, if John Brown is in truth 31 years old, the attachment to John Brown's name might read

Observed Age	21	. . .	29	30	31	32	33	. . .	41
Probability	.1		.075	.1	.5	.075	.05		.1

That is, the observation error in measuring John Brown's age is zero half the time, and it is usually less than 2 years, but there is a non-negligible chance of the error being 10 years exactly.

From this fundamental concept, one can proceed to the sampling of units. In terms of age and a specific probability sampling procedure, one would have a table with columns (say) for true age (in years), rows for observed age in years, and cell entries giving proportions. (Note that the sum of the probabilities for all cells in this table is 1.0.) A proportion 0.017 in column 16 and row 18 would mean that the probability is 17 out of 1,000 that a probabilistically chosen respondent will in fact be 16 years old and be measured as 18 years.[5]

When true value is a questionable concept—and even when it makes sense—the joint distributions for two or more modes of measurement come to mind. For example, age could be measured by (1) asking the respondent, (2) asking any member of the respondent's household, and (3) checking the respondent's birth certificate. Amount of sickness could be measured by (1) asking the respondent, (2) asking the respondent's physician, and (3) asking the relevant hospitals (see Andersen et al., 1979). There will nonetheless always be a nagging worry that all the modes of measurement considered are subject to the same systematic error relative to a better method, perhaps as yet undiscovered. Thus instead of a univariate distribution for John Brown's measured age, it's possible to begin with a (say) bivariate distribution in which rows and columns correspond to two modes of measurement; the entry in column 21 and row 25 would be the probability that, for John Brown specifically, method 1 gives 21 years and method 2 gives 25 years. In principle then, tables could be put together for all the individuals of a population (of course, using specific sampling and measurement procedures).

Age is a metric (numerical) variable and a relatively objective one. Other metric variables are further toward the subjective part of the spectrum,

such as respondents' judgments of what annual income is necessary today for a family of four to live in modest comfort in Chicago. Or the number of times they have been ill in the past year. Or, following Kinsey, their frequencies of sexual activity. Or, how fast a mermaid swims. All of these—except the last—are questions that have been used and that are thought useful. (Even questions like the last might sometimes make sense, for example, in studies of fantasy or myth.)

There are also nonmetric variables of several kinds, for which the same general framework may often be used. For example, consider ethnic affiliation, perhaps in the sense of the 1980 U.S. Census. If there is a true ethnic affiliation for John Brown, say Mexican, then we may consider setting down a distribution for measured affiliation something like the following:

Mexican	Puerto Rican	Cuban	Other Hispanic	Other
.6	.05	.05	.2	.1

(The distribution, as before, pertains to hypothetical repeated independent measurements.) Or, if true ethnic affiliation were deemed illusory, a bivariate distribution could be studied for two modes of measurement of ethnic affiliation.

A Simplification for Numerical Variables. For a metric or numerical variable, say income, it is often useful to simplify the joint distribution by looking at (diagonal) differences, that is, at the observed value *minus* the true value. (Here again, if a "true value" for the measurement is not assumed, it's possible to look at the observed value by method 1 *minus* the observed value by method 2.) This formulation gives the obvious and important linear decomposition

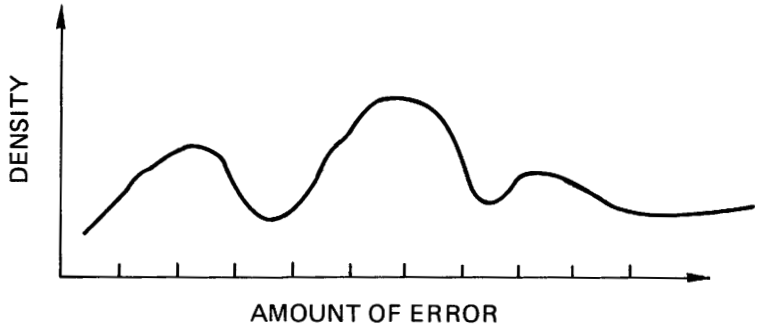
$$\text{observed value} = \text{true value} + \text{diagonal difference}$$

or its analog for two measurements.[6]

In short, we begin with a numerically valued variate and a measurement-plus-sampling method that generates a bivariate distribution of true value and measured value (or a bivariate distribution of two modes of measurement). We then look at differences (or ratios, etc.) which collapse the bivariate distribution of observed value by "true value" into a univariate distribution of (diagonal) differences (or ratios, etc). We may then think traditionally in terms of a hypothetical density function for amount of error (or amount of discrepancy between two methods), as shown in Figure 4.1. (Note that the neutral term "amount of error" is used, so that we do not prematurely or tacitly commit ourselves to differences. For specificity, however, we use differences from now on.)

FIGURE 4.1

Hypothetical density function for amount of error.



In studying error difference distributions, there are two traditional ideas that can be made precise in a variety of ways: systematic error and dispersion.

Systematic error refers to a general tendency for the differences to cluster or average around some value other than zero. It is frequently defined as the expected value (that is, average) of the difference of the observed values minus the true value, although other definitions can be found. The term "bias" is often used for that expected difference. *Dispersion* refers to the tightness or looseness with which the differences cluster around the systematic error. The most frequent way of defining dispersion is by the variance, the expected value of the squared difference around expectation, or by its square root, the standard deviation. Variance for the random variable X is the expectation (E) of the squared difference of X around its own expectation: $E(X - \mu)^2$, where $\mu = EX$. It may equivalently be defined as $\frac{1}{2} E(X_1 - X_2)^2$, where X_1 and X_2 are stochastically independent and distributed as X. Thus the variance concept turns on differences and expectation of the square. There are technical advantages (for example, additivity) in using variance or standard deviation,[7] but other ways of making the dispersion idea specific could also be used.

(Readers should be aware that terminology varies widely. For example, the term "validity" is often used to mean the absence of systematic error and the term "reliability" to mean low dispersion; see Section 4.3 for further discussion.)

It is often felt that a single number expressing error of an observational system is desirable. A technically convenient one is the *mean square error*, the expectation of the squared difference around the true value. Almost by definition,

$$\text{mean square error} = \text{variance} + \text{bias squared}.$$

In conventional symbols, if X is the observation, μ its expectation, EX and μ_0 the true value, then

$$\begin{aligned} \text{mean square error} &= E[X - \mu_0]^2 \\ &= E[(X - \mu) + (\mu - \mu_0)]^2 \\ &= E(X - \mu)^2 + (\mu - \mu_0)^2 \\ &= \text{variance} + \text{bias squared}, \end{aligned}$$

where $\mu - \mu_0$ is the bias or systematic error. The cross-product term is zero, since

$$E[(X - \mu)(\mu - \mu_0)] = (\mu - \mu_0)E(X - \mu) = 0.$$

Of course, one does not have to express observational error by an overall measure, a single number, or to use mean square error as that number. Yet the need is often felt for such a single number (perhaps from our common cultural background) as a means to compare, and choose among, different measurement modes. The term "accuracy" is often used in connection with overall measures of error.

It is often important to decompose variance into that part arising from sampling dispersion and that part arising from purely measurement error. Here our fundamental viewpoint is especially helpful. For a single observation, a sampling method provides a probabilistically chosen unit, call it U, which may be the name of a person, a sheet in a file, coded information on a magnetic tape, etc. Then associated with that U is a distribution for the actual random observation X. Of course X depends on U and the consequent measurement error.

Given U, X has an expectation written as $E(X|U)$. It is traditionally felt convenient to make the linear decomposition

$$X - \mu = [X - E(X|U)] + [E(X|U) - \mu],$$

where $\mu = EX$. The second term on the right is the purely sampling error; it depends on U only, not on subsequent measurement error. Now square and take expected values to obtain

$$E(X - \mu)^2 = E \operatorname{Var}(X|U) + \operatorname{Var}[E(X|U)],$$

or

$$\operatorname{Var} X = \begin{bmatrix} \text{expected measurement} \\ \text{variance given unit} \end{bmatrix} + \begin{bmatrix} \text{sampling} \\ \text{variance} \end{bmatrix}$$

The cross-product term disappears under the expectation operation if we first take expectation with respect to the measurement process. (Note the essential asymmetry of the preceding decomposition: measurement is done after the choice of a unit, or—in another argot—measurement is nested inside sampling units.) This decomposition can be described in various ways, and one traditional way is to consider hypothetical replicate measurements at different times, with subscripts for unit and time.

The foregoing exposition provides brevity and transparency at the expense of initial abstractness for readers unfamiliar with ideas of conditional expectation. The analysis may be extended to two or more measurements on the same unit, or to sampling with dependent units (as in cluster sampling). We can decompose covariances as well as variances. (This approach carries over readily to circumstances in which the idea of true value is ephemeral, but in which we are able to compare two or more measurement systems. We consider $\mathbf{X}$ and $\boldsymbol{\mu}$ as vectors and replace variance by the variance-covariance matrix.)

We should not be prematurely trapped, however, into use of conventional summaries. For example, age heaping would not be well studied by ordinary use of variances. In other cases it may be of primary importance whether an error exceeds or does not exceed some crucial value. For example, we may be examining a theory with a sharply defined parameter: the inverse *square* law, not the inverse 2.0001 exponential law; the hypothesis of exactly zero extra-sensory perception; zero mass for the neutrino (see Waldrop, 1981). Similar issues may also arise when the issue is eligibility for some government distributive program: household income, city population.

Measurements need not be numerical in the aforementioned sense, and many measurements are not. For example, a measurement may be ordinal (Rockwell hardness of a material, rank in the Army), or it may be purely nominal (religion, country of origin), or it may be circular (month of birth). Error structure for some of these cases has been examined. (See note 10.)

Sources of Error in Survey Measurements. Analysis of error structures for survey measurements requires study of such influences as survey interviewers, training of interviewers, details of forms, question wordings, and so on (as discussed in Chapters 5, 8 and 9). Knowledge of these influences is important in designing surveys to reduce error and in understanding stochastic dependence among errors of separate observations (either several characteristics of the same unit, or characteristics of separate units, for example, respondents interviewed by the same person).

One underlying difficulty is that some influences are clearly systematic or fixed, some are clearly probabilistic, and others may be thought of either way. The format of a questionnaire, for example, would generally be re-

garded as a fixed part of an investigation, and it might be desirable to compare the measurement variation associated with two different formats. Interviewers for large surveys would typically be regarded as chosen in some chance way, yet for a small survey with only one interviewer one would naturally be interested in that interviewer only. Question wording is usually thought of as systematic, but a few investigators regard questions as chosen stochastically from a population of questions.

When an influence is regarded as fixed, the approach outlined previously requires little or no change: there simply are two or more measurement methods to compare. When an influence is regarded as probabilistic, it is usually thought important to analyze or decompose variance (assuming that one is working with numerical variates and using variance) into components for (1) sampling error, (2) interviewer or other influences, and (3) measurement error. This may be done in a way similar to that in which variance was earlier split into sampling and measurement-within-sampled-unit components.

Reprise of the True Value Problem. Even when the quantity observed is nicely metric, if looked at closely enough, there are conceptual difficulties. The length of a bar of steel or the weight of a potato, appears at first to be clear in principle. But nearer the molecular or atomic level this is no longer the case. The very identity of the bar or the potato fluctuates constantly with time (see the writings of Steinhaus or Mandelbrodt).[8]

A familiar but difficult case is that of blood pressure taken in the physician's office. The measured pressure is a complex outcome of real pressure in the blood (varying over place and time) and of the structure of surrounding tissues in the body. Measured pressure fluctuates, sometimes wildly, with environment. What might be meant by "true value" for such a quantity? The interpretation of measured blood pressure depends partly on qualitative theory of blood flow and partly on empirical associations.[9] Similar examples of difficulties of true values are the temperature in Chicago on February 18, 1981, or a person's oral temperature at a particular time.

It may be argued that the true value idea can never stand up to sufficiently detailed examination—but then what idea can? It is always a question of being good enough for some current purpose.

For many kinds of measurement, however, it is hard even to make much sense of true value. Consider, for example, the questions: "Quickly, now, are you for or against the anti-solipsist act of 1981?" "Yes or no, saving money is a good thing."

If the dichotomous question is "male or female?" then nearly all of us know the answer and will respond correctly nearly all the time. For most practical purposes (except the study of physiological anomalies), there *is* a

true value for each person, and the only source of error is a (perhaps Freudian) slip of the tongue, a jest, or a downright lie. The basic error structure can be expressed in a 2×1 table for a given person.

But for forced opinion questions, like those mentioned previously, or for introspective questions ("On the whole did you or didn't you have a satisfactory January?"), or for some expert opinion questions ("Dr. Pathologist, does this slide show cancerous cells or not?"), the same stimulus, if it could be asked independently and repeatedly of the same person, would give rise to variability. This, in fact, has been demonstrated with a pathologist's slides by mixing in the same one—unidentified—with others. What is the true value in such a case? The Bernoulli probability? That is unsatisfactory. The consensus of all pathologists? There may be no consensus and, in any case, such an approach would not apply to the other dichotomous examples. Yet we all have a sense that there can be bias in such measurement processes. One can be asked whether last month was satisfactory in lugubrious or in merry tones. Subtle cues of voice or hand (recall the case of the ciphering horse, Clever Hans) can affect responses—even those of the pathologist.

In any case, *differences* are not relevant here, much less squared differences, or expectations of them. (However, we recognize that sometimes formal devices can be helpful: for example, letting 0 and 1 correspond to the response of a dichotomous question and looking at the meaning of the variance.) In general, error structure in these nonmetric cases needs to be looked at *ab initio*, and there is little literature about them.[10]

One hopes that there will be increased statistical research into modes of examining error structure for kinds of measurement other than those referred conventionally to the real line (or Euclidean space).

Substantive Theory and Measurement. Throughout science, the development of theory and the development of measurement methods are closely intertwined. There is fairly general agreement that a method of measurement is good if it provides numbers that fit in well with an interesting, productive scientific theory, so that measurement and theory go together much like a pair of roped mountain climbers. The history of our understanding of heat, electricity, blood flow, and so forth, show this clearly. Theory tells us what to look for and often how to look; measurements correct, refine, or overturn theory.

This interaction exists at all levels of theory, starting with the earliest stages in which a theory—if we may call it that—simply tells us what we should look at and what we should regard as units of observation or thought. Just as it is presumably a long way from the buzzing perceptual confusion of a newborn baby to the recognition of furniture, people, fin-

gers, and toes as worthwhile and coherent entities, so too it was a long way from early physical ideas to the abstractions of point masses, weightless cords, and the like. Surveys of subjective characteristics today are frequently based on theories at a somewhat vague level, yet far more specific than years ago.

Of course theory in the conventional sense of formulas and mathematics obviously affects measurement. Another kind of interaction between theory and measurement is that theory in one scientific field typically forms a basis for observations in another. Electric circuit theory is necessary for many psychological observations, optical theory is essential to instruments for observing wildlife, theories about human behavior affect the design of surveys of buying intentions. Theory also enters centrally in linking different modes of measuring the same thing. Length or distance may be measured by the electron microscope, precision calipers, yardsticks, radar, and so on. Theory helps in the interpretation of measurement, for example, in the recognition of gross error—a 15-year-old girl with 12 children (see also the following discussion of construct validity). Most basic, perhaps, from a philosophical viewpoint is the kind of theory expressed by the notions of concatenation or conjointness that underlie a measurement system. (See Section 4.3 and also the treatments of latent structures and in particular of the Rasch model in Chapter 6.) The central point, however, is that there is a seesaw or ratchet-like interaction between theory and measurement: both forever imperfect and each helping the other with its next step.[11]

In the following section we present examples of the empirical study of errors in survey measurement. The examples are drawn mainly from instances in which the phenomena might not ordinarily be considered subjective (for example, reports of actual behaviors, such as listening to a radio station or visiting a doctor, the languages spoken in a household, income). We subsequently turn, in Chapter 5, to analogous studies involving the survey measurement of subjective phenomena.

4.2 Measurement Variation in Surveys: Examples of the Effects of Respondents, Coders, and Interviewers

Variance is a measure of how closely measurements cluster around an expected value or how close independent measurements are one to another. There are many kinds of "closeness together" or variance in survey measurements. The major kinds of variation include sampling variation, response variation, interviewer variation, and coder variation.

By taking a large enough sample, one can reduce the sampling variation to any specified level. In thinking about response variation one might think about measurements obtained from a single respondent. If a person were asked the same question (dealing with a static phenomenon) many times, one might hope that the responses would be identical, that is, that there would be no *response variation.* Analogously, if several interviewers are separately collecting data from comparable sets of respondents, the hope would be that means (or other statistics) derived from each interviewer's work would be close together (which is to say we would hope to minimize the interviewer variance). A parallel situation occurs with the coding variation that is introduced when questionnaires are processed by several coders.

As we have seen in Chapter 3, survey reports seldom take account of the several sources of variation. Indeed, some writers (Wheeler, 1976; Roper, 1980) have complained that the commonly cited margins of error often encountered in published (and broadcast) reports promote overconfidence in survey estimates. These figures commonly account only for sampling variations and do not take into account the other sources of variation in survey estimates. (They also do not take into account the effects of the wording of questions and questionnaire context, which are discussed in Chapter 5.) An important result of the lack of precision introduced by these measurement factors is that it has the effect of increasing the total variance or, in a sense, of diminishing the effective sample size. Thus, a survey practitioner who makes inferences solely on the basis of sampling variance, without considering the effects of respondents, interviewers, coders, and so on,[12] is likely to make some inferential errors. As the magnitude of the variation introduced by these other factors increases, so does the likelihood of incorrect inference.

Similar concern arises about sources of variation in the measurements of physical scientists. Youden (1961) has discussed interlaboratory differences in measurements of the same quantity. He gave examples showing that a defined measurement procedure, used at different times and places, resulted in measurements that converged to different average values. Youden (1961:268) stated: "Every round robin [test of measurements made by different laboratories] results in a collection of laboratory averages that differ among themselves by more than can be reasonably accounted for by the within-laboratory precision." Youden advocated that each laboratory be judged by its closeness to the average over all laboratories rather than closeness to any true value.

Interlaboratory differences in physical measurements might be said to correspond to between-interviewer variability or the variability between sets of interviewers (or procedures) used by different survey organizations. Some authors have also referred to differences in physical measurements

done on different days—differences caused, perhaps, by changes in temperature, operators, instruments, and so forth. These kinds of differences might correspond to response variation (within-respondent variability) in sample surveys. External changes may cause measurement differences, but the within-respondent variability goes beyond external conditions and reflects internal changes in the respondent as well.

Survey models have been devised to help analyze some of these sources of variability. Hansen, Hurwitz, and Pritzker (1964) developed a model expressing within-respondent variability, and Hansen, Hurwitz, and Bershad (1961) developed a model to express interviewer and coder variability. These models depend on the decomposition of variance (discussed previously) in which variance is expressed as the expected measurement variance given the units plus the sampling variance. The models and experiments to estimate the parameters of the models have been used extensively at the Bureau of the Census and at Statistics Canada.

The total variance of a set of observations has been shown to be the sum of a simple response variance, a correlated response variance, and a sampling variance, denoted by σ_R^2, $\rho\sigma_R^2$, and σ_S^2. To see what proportion of the total variance is accounted for by variability in response of individual units, new questions being introduced to a survey or census are often studied with respect to simple response variance. If the response variance is large in relation to the total variance, then the question or questioning procedure might be reconsidered to see if it needs revision.

RESPONSE VARIATION

In order to estimate simple response variance, it is necessary to obtain at least two independent measurements for the same person. However, it must be recognized that it is (virtually) impossible to get truly independent observations. If a person was visited by an interviewer one week and asked a series of questions and is then approached the following week for another interview, it is unlikely that the person forgot entirely the results of the first interview at the time of the second. However, the effect of the lack of independence is to understate the response variance, so that at least a lower bound may be obtained. Another statistic, the index of inconsistency, has also been used in studying response variation. The index of inconsistency is the ratio of the simple response variance to the sum of the simple response variance and the sampling variance. Since both numerator and denominator are estimated, and the numerator is often subject to a downward bias, the index is frequently understated. Its range is between 0 and 1, with larger values indicating that greater proportions of the total variance of an item are due to response variance than to other variances. (Response variation might be reduced by designing a better question, by training interviewers to un-

derstand and ask the question uniformly, or by some other way of reducing measurement variability.)

Empirical estimates of simple response variances and indices of inconsistency vary greatly by type of item. In the 1960 and 1970 censuses, the indices for "year built" for rented units were much higher than indices for the same item for owner-occupied dwellings. This makes sense when one thinks about how much more difficult it would be for renters to know when their housing units were built (especially for older units built long before renters moved in). For this kind of question, it might make more sense to ask the owners of the buildings. Costs of utilities for renters also evidenced high indices of inconsistency in the 1960 and 1970 censuses (though in this case renters who paid their own utilities and who kept and would use records in responding provided more consistent estimates). Questions on the amount of rent paid, however, were reported more consistently.

Among population measurements, those pertaining to marital status and number of children in the household had comparatively low indices of inconsistency. Higher indices of inconsistency were found for educational attainment, occupation, industry, weeks worked, and wage and salary income, and even higher ones for self-employment income. Respondents' estimates of the value of their homes (for units that were owned) had the highest index of inconsistency for all measurements in the 1970 census. Again, this latter item may be one for which an owner has very little information if the unit has not been for sale at a time recent to the census.

Certain factors seem to play a large part in generating inconsistency. Some information is not known or it is not easily accessible to respondents. Other information needs thought. Some information is very susceptible to being forgotten so that long recall periods may cause less consistency in response.

INTERVIEWER AND CODER VARIATION

There are two main models that have been used to measure between-interviewer variability, one based on analysis of variance and one based on variance decomposition. They are alike in that they use the same estimators, based on identical experimental designs, although one uses fewer assumptions than the other. The variance decomposition model formulated by Hansen, Hurwitz, and Bershad (1961) at the Bureau of the Census is the one from which most results are available. Though most of the results pertain to means or proportions, the model has been extended by Koch (1973) to apply to other population parameters.

In the variance decomposition models, the variance of a mean, for example, is decomposed into terms that reflect the sampling variance, the simple response variance, the correlated component of the response variance, and

an interaction term that reflects the covariance of response and sampling deviations. In the case of a complete census, all terms are zero except the two response variance terms.

The correlated component of the response variance expresses the correlation of response deviations of different persons in the same interviewer's assignment. This term reflects the idiosyncratic influences of each interviewer who may have slightly different understandings of survey instructions, who may differ in their handling of nonresponse, or who may otherwise affect the responses given.

To estimate the correlated component of the response variance, an interpenetrating interviewer design is necessary, in which each interviewer's assignment is spread over the same area as all other interviewers' assignments. This type of design, introduced by Mahalanobis (1946), has been used successfully by the Bureau of the Census, Statistics Canada, and the Survey Research Center at the University of Michigan. In general, the use of this design in censuses has shown that items such as age, sex, race, marital status, and number of children show little between-interviewer variability. But questions on educational attainment, income, and place of birth show relatively high between-interviewer variability. In 1970, one of the new items on the census questionnaire was a question about language spoken in the home. The between-interviewer variability on this item was substantial (see Bureau of the Census 1968, 1979). For the most part, it was found that questions that had high levels of response variability also had high levels of between-interviewer variability.

An interviewer variance study was also conducted as part of the National Crime Survey in eight central cities. In this study (Bailey, Moore, and Bailar, 1978), relatively high levels of between-interviewer variability were found for questions asking whether the respondent had been the victim of a violent assault, especially one that did not involve theft. The results suggested that interviewers were applying the concepts or definitions related to the determination of an incident of assaultive violence in different ways. The responses to questions on household crimes, especially larceny, also showed considerable variability.

We note three other recent studies of interviewer variation as a factor in survey response. Kish (1962) has reported on interviewer variability in a study of blue-collar workers in an industrial plant who were asked factual and attitude questions about their jobs and related matters. He reported interviewer variation for some "factual" items such as education, income, and migration was no less than that for the attitude items included in that survey. Freeman and Butler (1976) examined between-interviewer variation in a study of mental retardation in a small California city. Both "factual" and subjective items were asked. The range of between-interviewer varia-

bility in this study was much greater than in most other studies, and these authors found especially high levels of interviewer variation for their attitude measurements. Freeman and Butler related the interviewer effects to the role of interviewers as interpreters of information: for example, interviewers are asked for clarification when questions are vague. Glasser, Metzger, and Miaoulis (1970) studied the effects of interviewers in telephone surveys designed to determine the activity going on in the households at the time of the telephone call. The authors found interviewer variability increased as supervision decreased, but nonetheless, there was measurable between-interviewer variation in survey response even among more strictly supervised interviewers.

Variation on survey measurements is also introduced by coding and other processing procedures. In a study carried out in the context of the 1960 census, the levels of coder variability were found to be less than the levels of interviewer variability. However, for items that needed coding and for which high levels of interviewer variability existed, relatively high levels of coder variability were also found. Perhaps the most thorough study of coder variability was carried out by Kalton and Stowell (1979). Estimates of both within-coder consistency and between-coder variability were made. Unfortunately, the experiment was based on only six coders so that some of the estimates were subject to large sampling errors. Kalton and Stowell found that agreement among coders for factual items was very high, but that when the coding required some judgment, considerable variability occurred. One overall finding supported by other studies is that catch-all codes and general codes exhibit great variability.

Many other examples could be provided of the effects of different parts of the "measuring instrument," including interviewers, coders, and respondents. The fundamental point is that, in addition to the variability introduced by sampling, the respondents, interviewers, and coders affect the precision of survey results. Factual questions that do not require much effort by respondents, such as questions on age, sex, marital status, and race, seem to have high precision. Questions that demand more effort by respondents, such as remembering past events, checking records, and so forth, show more variability in respondents' answers and in interviewers' and coders' reports. That is, questions that yield more variable responses from respondents also seem to show higher levels of interviewer and coder variance.

The lack of precision introduced by respondents and survey personnel has the effect of increasing the total variance of a survey estimate (in a sense, diminishing the *effective* sample size). Thus, practitioners who make inferences based solely on sampling variances (without considering the effects of respondents, interviewers, and coders), are likely to be in error

more often than they expect. As the magnitude of the variation introduced by the other factors increases, so does the likelihood of erroneous inference.

SYSTEMATIC ERRORS OR BIAS

Systematic errors are more difficult to estimate than variance because of the need for comparison with some (hypothetical) true value. Among survey practitioners, philosophical debates about the existence of "true" values for survey measurements have been carried on while, at the same time, a more pragmatic approach (Eisenhart, 1963; Deming, 1950) of using an exemplar or preferred measurement procedure has been followed. The true value has been defined in the latter approach as the value that would result from the application of a preferred or exemplary measurement procedure. The true value is thus defined by a series of operations required in a very careful measuring process. For example, to measure age, one may state that age will be determined by a specific calculation resulting from the month and year of birth *actually seen on a birth certificate*. Usually, the preferred procedure is very expensive or very difficult to carry out, and it may be impossible to specify fully an exemplary procedure. Survey practitioners have thus tried to get at accuracy in many and varied ways: sometimes an outside record source defines the true value; sometimes a more probing interview defines the true value; sometimes a report from another person provides the true value.

In most studies of the accuracy of survey response, accuracy has generally meant the extent to which there is absence of bias, that is, the absence of systematic errors. Sources of bias are numerous in surveys. Biases can arise from the questionnaire, instructions, field personnel, definitions, and respondents. We do not know of any studies that attempt to measure *all* of these biases individually for a particular survey. There are studies of particular sources of bias, and often the combined effect of biases may be studied, though the individual components are not estimated. Thus, it is not known how the different biases may affect each other. For example, it is possible that certain biases cancel out other biases. In any case, we emphasize that while it is important for survey planners to know the individual components of bias, the combined effect may be most relevant information for users of survey data.

In the following pages, we will cite some interesting examples of the ways in which investigators have tried to assess the accuracy of their measurements. It is not our intent thereby to dispense a nihil obstat for wholesale adoption of any specific procedure, but rather to titillate the imagination of those who must confront similar problems in their own work.

An Example of Verification. Dutka and Frankel (1976) describe a telephone survey designed to measure the number of people listening to specific radio stations. Telephone calls were made at random times during each quarter-hour period of the day to a random sample of persons in households with listed telephone numbers. Each person was asked if the radio was on, and if so, the station tuned in. In a test in the New York metropolitan area, the accuracy of these reports was evaluated. Each interviewer had an electronic device that would transmit over the telephone the broadcast coming from any one of the 20 leading AM stations in the New York area. In the interview, when a respondent said a certain station was tuned, the respondent was asked if he or she could hear the radio playing while at the telephone. If not, the respondent was asked to turn the radio louder. When the respondent could hear the radio while at the telephone, the interviewer pressed the test button of the electronic device that corresponded to the station reported. If the respondent reported that what was coming over the telephone was the same as what was on the radio, the original response was considered correct. In every *n*th case, the interviewer pressed an incorrect station. Of a total of 854 cases in which this validation approach was used, 9 percent of the original responses were found to be incorrect. (Other interesting examples of the use of verification or validity checks can be found in Bradburn et al. [1979:Chapter 1].)

An Example of the Use of a More Probing Interview. Another way to define a true value is in terms of information collected in a more probing interview than the original survey interview. Obviously, this method has its limitations because it will not provide any information about inaccuracies caused by deliberate lies, but the method may provide useful information about concepts that are ambiguous or difficult to understand. After the 1960, 1970, and 1980 censuses, reinterview studies were conducted to measure the accuracy of responses to selected census questions. An example is a question asked in 1970 about the language spoken in the home when the respondent was a child:

> What language, other than English, was spoken in this person's home when he was a child? (Fill one circle.)
>
> ○ Spanish
> ○ French
> ○ German
> ○ Other—specify
> _____
> ○ None, English only

Because of space and time constraints, it was decided that no instructions, definitions, or guidelines would be given to indicate how often or under

what circumstances a foreign language should have been spoken to be reported. Thus, there was a wide range of interpretation by respondents. Subsequently, however, supplementary data were collected (in a separate reinterview of a sample of respondents) about foreign language use in the respondent's childhood home (Bureau of the Census, 1974).

The detailed data included who used the language, under what circumstances, how often, and what was the primary language spoken. These detailed questions were used to stratify persons reporting foreign language use into seven levels:

Level 1—Foreign language spoken as primary language; English *not* spoken.

Level 2—Foreign language spoken as primary language; English *also* spoken.

Level 3—Foreign language spoken about *equally* with English.

Level 4—Foreign language spoken frequently in early childhood, spoken when relatives or visitors came, daily or several times a week; spoken daily or almost daily at home; spoken under other similar *frequent* circumstances.

Level 5—Foreign language spoken when relatives or visitors came, less than several times a week; spoken to prevent child or others from understanding; spoken under other similar *occasional* circumstances.

Level 6—Foreign language spoken very seldom for slang, phrases, expressions, etc.; spoken in conjunction with a language class taken in school; spoken under other similar *seldom* occurring circumstances.

Level 7—Foreign language was not spoken by the sample person, but was used by others in the home from time to time.

By varying the definition of "foreign mother tongue" on the basis of these levels of foreign language use, comparisons were made between the reinterview and the census response. The results are shown in Table 4.1 (a) and (b). It was found that census respondents tended to report *any* foreign language use in their childhood homes, regardless of intensity (although 25 percent of those who said in the reinterview that they spoke a foreign language had been reported as speaking only English in their childhood homes). Thus, almost all (96.8 percent) of the persons in level 1 reported foreign language use in the census while over half of the persons in level 7 reported foreign language use. If one considers levels 1 through 6 as defining foreign mother tongue, the proportion of persons reporting a foreign mother tongue in the census was overstated by about 2 1/2 percentage points. If one uses this definition, then the net difference rates shown in Table 4.1(b) are a measure of the inaccuracy of response or the amount of bias. A positive number means that the census respondents overstated the number of persons who spoke a foreign language in their childhood home; a negative number means that the census understated the number. (Alter-

TABLE 4.1(a)
Comparison of Census and Reinterview Responses for Foreign Language Use in Childhood Home, 1970 Census

Reinterview Classification	Census Responses					
	Number			Percentage Distribution		
	Total Sample Persons	Foreign Language Spoken in Childhood Home	English Only Spoken in Childhood Home	Total Sample Persons	Foreign Language Spoken in Childhood Home	English Only Spoken in Childhood Home
Total sample persons	11,102	1,655	9,447	100.0	14.9	85.1
Foreign language spoken	2,170	1,627	543	100.0	75.0	25.0
Spoken by person himself	1,383	1,216	167	100.0	87.9	12.1
Level 1	412	399	13	100.0	96.8	3.2
Level 2	445	410	35	100.0	92.1	7.9
Level 3	41	38	3	100.0	92.7	7.3
Level 4	350	276	74	100.0	78.9	21.1
Level 5	95	72	23	100.0	75.8	24.2
Level 6	40	21	19	100.0	52.5	47.5
Level 7	787	411	376	100.0	52.2	47.8
English only spoken	8,932	28	8,904	100.0	0.3	99.7

SOURCE: Bureau of the Census (1974).

TABLE 4.1(b)
Net Difference Rates for Foreign Language Use Estimates

Level	Percentage in Class According to Reinterview	Net Difference Rate (Estimated Bias)
1 through 3	8.1	6.8
1 through 4	11.2	3.7
1 through 5	12.1	2.8
1 through 6	12.5	2.5
1 through 7	19.6	−4.6

natively, one can use the cross-tabulation shown in Table 4.1(a) to recalibrate the results obtained with the cruder measuring instrument used in the decennial census.)

Examples of Cross-Checking Other Records. Another approach to the definition of a true value is possible when a record exists that is less error prone than a respondent. With appropriate safeguards to protect the confidentiality of responses, reports to the Internal Revenue Service have been used to measure the accuracy of income reporting in the census and the Current Population Survey (CPS), and aggregate educational enrollment figures collected by the Office of Education from individual school systems have been used to measure the accuracy of school enrollment figures collected at individual households in the CPS. However, such uses of record systems inevitably raise questions about the accuracy of the record system that provides the nominally true value. The record system may also be subject to error, and this possibility modifies the interpretation of the difference between the hypothetical true value and the observed value.

Other kinds of records may also be used. Ferber (1966) gives an excellent example comparing actual amounts in savings accounts with reported amounts. In a pilot study, a sample of accounts was selected from the files of a savings institution. A sample of people with accounts were questioned about their savings accounts. In the comparison of the responses with the actual account balances, bias was found: small accounts tended to be overstated and large accounts tended to be substantially understated. (Other interesting examples of the use of record cross-checks can be found in Bradburn and Sudman et al., 1979.)

Sometimes, another kind of true value is used on the assumption that a report by another person is more accurate than the response of the respondent. This type of standard was used in the 1960 and 1970 censuses when industry and occupation responses supplied by persons answering the cen-

sus questionnaire were checked against employer reports. Employers were asked to furnish job titles and job descriptions for a sample of persons. In addition, employers were asked to furnish a description of the major activity and principal product or service of the establishment at which employees worked. In 1970, detectable biases were found in some broad occupational categories. Assuming employers' reports are the true value, respondents overreported professional and technical occupations in the census while managerial, administrative, and laboring were underreported.

A similar kind of report furnished by persons who are believed to have more accurate information than respondents was illustrated in the measurement of the accuracy of health data (Andersen et al., 1979). The goal was to examine the error associated with respondents' reporting of illness conditions for which they saw a doctor, conditions for which they were hospitalized, and operations they had during hospital stays. For illness conditions for which respondents saw a doctor, the doctors were asked about patient conditions. The information from the doctors was accepted as a true value. Similarly, when data on hospitalizations or medical costs were verified, reports from hospitals and insurance companies were accepted as true values.

In summary, many different kinds of procedures have been used to define true values. All of them seem to have provided useful information. For subjective questions, however, objective verification through record cross-checking and similar procedures is not possible. As we discussed in Chapter 1, the very definition of "subjective" requires that the respondent alone have direct access to the object of the measurement process. Self-report is, thus, an inevitable component of subjective measurements, and it raises some unique questions. In turning our focus in the next section to survey measurements of subjective phenomena, we begin by briefly reviewing theories and perspectives that have been developed in the field of psychological measurement.

4.3 Psychological Measurement: An Introduction to the Subjective Domain

Survey measurement can be compared to many other types of measurement. However, since measurement in the physical sciences—although certainly not free of error and controversy—does not partake of many of the vexing problems arising with "reactive measurements" on human beings (Campbell, 1957), the most appropriate comparison point is with other kinds of psychological measurement, especially those that entail assessments of verbal responses.

As we noted in Section 4.1 of this chapter, the process of measurement involves the assignment of numbers to objects in an orderly way so that selected empirical relationships among the objects correspond to relationships among the numbers assigned to them. For a very simple example, when object *A* exceeds object *B* on some specified property (for example, *A* is preferred to *B* by a decision maker), then *A* should be assigned a higher number than *B* on the scale of measurement (for example, of preference or "utility"). Measurement thus represents a mapping of selected aspects of real objects into the abstract relational structures of number systems.

In an important sense, the quality of measurement depends on the constraints imposed by the object relations on which the mapping is based. Weak relational constraints give rise to weak measurement. For example, the simple relation, *A* exceeds *B* (in some property), is really not sufficient to generate a fully numerical scale of measurement, but merely a rank ordering or ordinal scale of a set of objects. This classification of various scale types was popularized by Stevens (1946) and continues to be discussed in modern treatments of scaling (for example, Carroll and Arabie, 1980).[13]

The axiom systems underlying measurement models are not always easy to test empirically. What is involved is a search for postulated patterns of association among responses to a set of objects. This search is made difficult by the presence of error in typical data sets. Despite the accelerated development of formal measurement models in psychology, applications to survey data require articulation in part because survey analysis typically treats categorical responses rather than quantitative scales. Nevertheless, the search for principled patterns of association in rich data sets is still the best basis for understanding, and many examples of this strategy are set forth in Chapter 6. Here, we turn to a review of other concepts from psychological measurement that may be applicable to surveys.

MENTAL TESTING

Within the branch of psychology known as psychometrics, there is a long history of attempts to measure (hypothesized) psychological characteristics of individuals by eliciting (usually verbal) responses from them. Psychometrics began with the development of tests of *mental abilities*, starting with the work of Alfred Binet in the early 1900s. The measurement of general intelligence and a host of specialized aptitudes and abilities has grown rapidly since then, given the many societal pressures toward objective identification of individuals of extremely high or low mental ability. Even so, the measurement of mental abilities remains controversial in many respects, not least of which is the still nagging question: What is intelligence?

The history is interesting in the context of the present report because survey measurement of subjective phenomena is today a mushrooming applied enterprise, showing much the same explosive and often undisciplined

growth the mental test movement has undergone. It is not appropriate here to review the enormous field of mental test measurement in any depth of technical detail; that is done elsewhere (see Gulliksen, 1950; Cronbach and Gleser, 1965; Anastasi, 1976; Wigdor and Garner, 1982). But we wish to draw upon the valuable experience of the psychometric tradition in pinpointing some conceptual issues in measurement.

Mental test experience is instructive because the measurement of mental ability seems to be objective: every test question presumably has a right answer, and thus a good score seems to imply the presence of the ability underlying the test. Yet the interpreter of test scores is faced with a number of questions about their meaning. Ability is inferred from performance on a test; it is not measured directly, and thus the whole process of mental test measurement has a strong subjective flavor. (In the related domain of personality measurement, the subjective aspect is even more explicit.) The main (but not the only) reason that there is an inferential leap from test scores to the imputation of underlying ability is that mental tests are composed of very particular questions or items—specific vocabulary terms, for example, or problems in arithmetic reasoning. Every test item has unique properties and cannot perfectly represent the underlying characteristic that one wishes to measure. Items vary greatly in their difficulty and in many other ways. Because of imperfect representativeness of any given item, it is necessary to consider the measurement in terms of samples of stimulus items from some relevant domain. Often, however, there can be debate about the exact boundaries of the relevant domain.[14]

Use of Multiple Items. Surveys often rely on only one or two questions to tap particular content domains. To understand the interpretive risk involved in this strategy, consider the following two questions from the 1977 General Social Survey (National Opinion Research Center, 1978):

Q.143: Do you approve or disapprove of a married woman earning money in business or industry if she has a husband capable of supporting her?

Q.189: If there is a limited number of jobs, do you approve or disapprove of a married woman earning money in business or industry when her husband is able to support her?

For the first item, 65 percent of a national sample approved, while for the second item, only 36 percent approved. When the two items are placed side by side, it is obvious that the second one is more stringent in the requirements needed to say "approve." This is essentially analogous to a group of individuals having more success with a mental test question when its framing makes it an easier question. Yet if only one of the two preceding questions was used (and the other not even considered), an investigator might be tempted to conclude that pro-feminist sentiment in the United

States was running at 65 percent—or 36 percent—depending on which item was used. Clearly either conclusion alone would be an erroneous interpretation. (See Chapter 6, Section 6.4, for a further analysis of these questions.)

Why, then, do survey investigators expose themselves to such an interpretive risk? The answer is, in part, because of limitations on total interview time in the face of a great many domains of intended survey coverage. But it can also be due to accumulated experience over time with a single item, or to puzzlement as to how to present results when multiple items are used. In mental testing, the typical interest is in the identification of differences among individuals. If, say, vocabulary knowledge were measured by asking only two vocabulary questions, then the possible scores would range only from zero correct to two correct, and it would not be possible to identify some individuals as notably better or worse than average. This consideration is not so vital for survey measures because rarely would one want to identify specific extreme individuals per se.

One further, important reason for using single items should be emphasized, namely that certain items have more or less fixed meaning. A good example is the typical "trial-heat" question asking for candidate preference in a well-publicized political race. It does not much matter in which order the candidates' names are given, or whether one asks, "Who would you like to see win?" The voting decision refers to a consensual political process with clearly entailed consequences. The definition of what is being asked about suffers no ambiguity, and there is no necessity for a sampling of questions. (This consideration is further elaborated in Chapter 7.)

RELIABILITY AND VALIDITY IN PSYCHOMETRICS

The general concept of reliability is applicable to any form of measurement and is even used colloquially when referring to data arising from informal judgments or assessments. For example, when we say that witnesses' reports of the height of a criminal suspect are unreliable, we mean that such reports typically vary from one witness to another, or for the same witness under different hypothetical conditions of distance, lighting, and so on. Measurements are unreliable, then, to the extent that they come out differently while purporting to be measurements of the same thing(s).

In psychometrics, the major source of unreliability comes from the item sampling process and the probabilistic relation between a particular mental ability and the tendency to be correct on any given item. Psychometricians try to characterize the reliability of particular tests with reliability coefficients, numbers ranging from 0 (for a completely unreliable test whose score assignments are totally random) to 1 (a completely reliable test whose score assignments are perfectly reproducible). There are three different

methods for estimating the reliability coefficients of mental ability tests: test-retest, parallel forms, and internal consistency.

In the test-retest method, the same test is given after a lapse of time to the same group of individuals (for example, an IQ test readministered after a year), and scores on the two administrations are compared. In the parallel forms method, a second test whose items are conceptually parallel to those of a first test is administered to the same group that took the first test. In internal consistency methods, "parallel" items are located within a single test, and scores on parts of the test are compared with scores on other, comparable, parts.

If one looks carefully at the conceptual basis behind these methods, it becomes clear that the definition of reliability is relative to whatever types of variation in test construction and procedure are acknowledged. This is true for all three methods, but can be illustrated most clearly for the test-retest method. When scores on a test and retest are compared, presumptions must be made about whether the retest is the "same" test as the original. If the temperature in the testing room varied, but this is an irrelevant circumstantial variation, the test would still be considered the same. However, to the extent that temperature made a difference for many of the individuals tested, the reliability coefficient would decline. But is the test still the same if administered in a different room, or by a different tester, or with oral instead of written instructions, or in a different language? Some of these variations might be considered substantive alterations in the test, implying that it is unfair to count the score variations they introduce against the reliability of the test. Judgments, then, of what circumstances constitute the test proper make a difference to the reliability coefficient.

Many such judgments seem reasonable and unarguable (say, that the test be given individually in a quiet room by a person with certified training) and constitute what psychometric experience would regard as standardized test conditions. But standardization can never be complete. In any case, factors that make test scores variable even when the test procedure is standardized contribute to errors of measurement. Mathematical specifications for reliability coefficients in psychometrics are usually given in terms of hypothetical true scores that distinguish individuals one from another and of errors of measurement that vary around those true scores.

It is entirely possible for a measurement procedure to be relatively stable (that is, reliable), yet still not produce meaningful results. The term *validity* refers to whether measurements produce results consistent with conceptual intent. In the mental test domain, several aspects of validity have traditionally been identified, but the major distinctions are among *criterion validity*, *content validity*, and *construct validity*. We consider each of these in turn (Wigdor and Garner, 1982).

Criterion (or predictive) validity refers to the degree to which the mea-

surement is predictive of the properties it purports to measure as manifested outside the measurement situation. Thus the Scholastic Aptitude Test (SAT) is supposed to measure the aptitude of college applicants for study in college. The SAT has criterion validity to the extent that high scorers on the test indeed tend to perform better in college. Prediction in this type of case is relative: that is, the better the test score, the better the grades, without there being any a priori idea of what absolute grade level is implied by a given test score. Validity in such cases is assessed by the correlation coefficient between test scores and grades.

In survey measurement, there is no interest in making predictions for specified individuals, although there can be concerns for predicting aggregate outcomes. For example, in trying to predict voter turnout in an election, it would be desirable to have a question on intention to vote that had high predictive validity. This has traditionally been a difficult matter since stated intentions often do not predict future behavior. Note that in this type of case one wishes ideally to achieve absolute prediction, whereby every stated intention is exactly predictive of behavior, so that the percentage of intenders will predict the percentage voting. Absolute prediction is obviously difficult to achieve, although it is possible to achieve the illusion of absolute prediction through compensating errors.[15]

Content validity (sometimes called face validity) is not based on data but on whether the questions asked are appropriate to the content area claimed for the measure. To test knowledge of the meanings of words, for example, it would not be appropriate to ask questions about how to spell them. The concept of content validity has ready application to survey practice. In assessing, say, attitudes toward minorities, questions concerning equal employment, integrated housing, affirmative action, school busing, intermarriage, and social welfare programs might represent a good sample of real issues on which attitudes are actually manifest amd consequential. Such issues are socially and politically well defined, making it easier to write questions that have content validity.

There are many domains, however, in which issues lack concrete definition. These are the domains that often engender controversy about survey results. For example, whether people have "confidence in institutions" may seem to be an interesting issue, but it lacks concreteness. The various ways in which people express confidence or lack of confidence in institutions are diffuse and indirect. Thus it is hard to compose relevant questions that have high content validity.

Content validity usually arises from social consensus on what defines a domain and what is appropriate to ask. An interesting example in the mental testing field is the National Assessment of Educational Progress, a testing program to ascertain the level of mastery by school children of basic knowledge in major areas of learning. The test items were selected by a long, nationwide process of consultation among educators as to what topics

were nearly universally represented in school curricula and thus should have been learned by most school children.

Despite the obvious virtues of content validity, high content validity is sometimes accompanied by low predictive validity. The most straightforward questions may evoke pat answers that fail to predict behavior. Asking people about their intentions (say, to vote, or to quit smoking, or their endorsement of conventional standards of moral conduct) may be of this character. Measurements with low content validity and high predictive validity can be imagined: subtle item sets on mental tests may predict later performances well or the possession of one ability (say, spelling ability) may predict possession of another (say, vocabulary usage), even if they have little apparent direct connection. Using measures with low content validity for predictive purposes is very likely to provoke controversy because people may consider it unfair to reach a judgment about X on the basis of a measurement of Z. The survey field is exposed to such difficulties when subtle or indirect measures of, say, racial prejudice are used, even if such measures have high predictive power, both individually and collectively. Ideally, one would like to have measures with both predictive and content validity, but this may not be easy to achieve.

It is even more difficult to realize some degree of construct validity. This third type of validity invokes the network of conceptual relationships between the measured variable and other variables. In mental testing, an ability considered a theoretical construct in a theory of mental functioning might be related to other constructs in a particular way. Speed of making mental comparisons, for example, might be important in a theory of performance in solving verbal analogies. Construct validity is the degree to which sets of measurements relate to other measurements in a coherent way as specified by some theory. There is no single coefficient or judgment for this degree: it depends on the nature of the relevant theory or theories and the ingenuity of investigators in bringing multiple measurements to bear upon each other, especially measurements of different types from different disciplines.

In survey measurement, there are a few good examples of constructs that are part of a coherent nexus of relationships. The prestige associated with occupations is one such example (see Chapter 6). Other examples can be more controversial, particularly if the theory associated with the main measure is complex or incomplete and itself controversial. This often arises with personality or attitude measures of an abstract or symbolic nature. Concepts such as alienation or anomie or malaise imply many things, but often these implications are vague or moot.

A well-known example of a personality measure embedded in an extensive theoretical apparatus is the F-scale for assessing the "authoritarian personality" (Adorno et al., 1950). The F-scale contains a broad variety of statements with which the respondent is asked to agree or disagree. Two of

them are "Obedience and respect for authority are two of the most important virtues a child should learn" and "Sex criminals should be publicly whipped." The theoretical construct that is postulated to relate the responses to these disparate questions is "authoritarianism," which is hypothesized to involve a punitive upbringing leading to a personality syndrome incorporating an obsession with authority relationships, the repression of sexual impulses, ethnic prejudice, a striving for power, conservatism in morality and politics, and so on.

With such a theory, construct validity can be quite different from content validity. The items on the F-scale, taken as a single set, have very low content validity. They don't seem to be about any one thing, variously covering obedience, sexual attitudes, politics, morality, and so on. Yet, if the theory of the authoritarian personality were correct and people high on the F-scale also proved to be highly politically conservative, racially prejudiced, sexually repressed, morally conventional, and so on, then the scale could be considered "valid" (in the construct sense) even though its item set is not, on the surface, coherent. In fact, the relationships postulated by the theory have proven to be weaker empirically than had been anticipated and capable of alternative explanation.

One value of the attempt to establish construct validity is the possibility that one will fail and in the process will learn why. An important part of the search for construct validity is the check for empirical relationships that the theory claims will not hold, that is, the check for artifactual relationships. In the case of the F-scale, there is some evidence that scores are related to intelligence, to social class, to education, and also to a general tendency to acquiesce to homilies. All of these relationships detract from an interpretation of the F-scale as a pure personality variable.[16]

The concept of construct validity, and indeed other concepts of validity and reliability, all direct us to look extensively in survey data for patterns going well beyond the examination of a single item on a single occasion. It is only in a wider context of coherent relationships that survey data have substantive meaning.

In Volume 2 we present three empirical studies that attempt to provide evidence of the validity of some common survey measurements of subjective phenomena. In each of these studies, the validity (or meaningfulness) of a subjective survey measurement is empirically demonstrated by establishing the existence of systematic relationships between time series of subjective measurements and independent measurements of other theoretically relevant (nonsubjective) phenomena. MacKuen in Chapter 14, Volume 2, for example, demonstrates systematic relationships between the public's assessment of the problems facing the nation—as measured in a series of surveys conducted over the last two decades—and patterns of newspaper coverage and objective events during the same period. Similarly, Beniger in Chapter 15 demonstrates that there are systematic relationships between

actual contraceptive behaviors, media reporting of the health hazards of specific contraceptive devices, and survey measurements of women's attitudes toward contraception. Finally, in Chapter 16, MacKuen and Turner show that there is both a considerable (although imperfect) agreement between five independent time series of presidential popularity measurements made between 1962 and 1980, and also a meaningful relationship between those measurements and independent measurements of political and economic activity in the United States during the same period.

Clearly, similar demonstrations might be attempted in the future for other subjective measurements. The demonstration of a meaningful relationship between subjective survey measurements and other (independently measured) variables is a fundamental task of scientific validation. While we do not subscribe to the notion that survey measurements of subjective phenomena are inherently meaningless, neither should the meaning of such measurements be taken as self-evident. Scientific measurement in surveys requires systematic exploration of the patterns of empirical relation between such survey measurements and other, independent measurements (see Recommendations 1 and 16, Chapter 10).

Notes

1. This is a popular but not universally held viewpoint; see subsequent discussion in text.

2. Some physical quantities clearly present difficulties of both concept and operational measurement, for example, the wind speed at 30,000 feet over Chicago.

3. This position may be called naive realism, but it is a working position held by many researchers, and it may, perhaps, be necessary for vigorous scholarly work—not to mention everyday life. "Truth is absolute," writes the historian Oscar Handlin, "it is as absolute as the world is real . . . the historian's vocation depends on this minimal operational article of faith" (Handlin, 1979:405). For further excellent discussion of this point, see the review of Handlin's work by Wood, 1981.

4. The ideas of probability and distribution refer traditionally to idealized independent repetitions of the measurement process on the same person. (Some readers might be more comfortable with a subjective probability interpretation—although we are not.)

5. A full treatment, allowing for stochastic dependence of measurement errors, would be in terms of a joint distribution for an entire sample of true and observed ages. This leads too far afield for our purposes, and indeed it is tempting to simplify instead. One is often interested in more compact derived distributions, for example, the conditional distribution in a column of observed ages for a given true age, or the conditional distribution in a row. Again, one is often interested in diagonal probabilities: the total probability on the main diagonal (column number = row number) is the probability of no age error; the total probabilities in the neighboring diagonals are those of 1-year error; and so on.

Such a joint distribution of true and observed age is of course specific to a population of people, households, or other units, and specific to a particular way of choosing units and questioning or otherwise measuring them. The joint distributions for different populations might will be of interest. In the case of age, for example, those distributions would doubtless be different for unmarried and married people. So there is really no end to the possible adumbration of error structures. Of course there is a hope—often realized—that some important aspects of the joint distributions are much the same for many populations and modes of measurement.

6. Note that this particular mode of simplification may not be desirable. For example, in census studies of age heaping, strange humps or modes in age distributions at 10 year points are typically noticeable, and we may be interested in spreading those out reasonably. Again, there is no law that we must look at differences; it may be more useful to work with ratios (say in the case of prices) or still other functions. Or it may be useful to work with differences of logarithms or with some other data transformations (see J. Kruskal, 1978).

7. A fundamental mathematical rationale is the way variance enters crucially in the central limit theorem and thus in the factotum use of normal approximations.

8. There may be other types of conceptual ambiguity in a measurement. Consider, for example, the concept of the size of a person. Even in a specific, operational setting that concept may have many explications. Size for aviation travel could mean weight (for fuel consumption planning) or width (for seat design) or girth (for seat belt design) or height (again for seating).

9. We also note that blood pressure and other physiological measurements (like electrical conductivity of the skin) are sometimes used to determine emotional states. We do not wish to comment here on the validity of such connections between relatively objective measurements and relatively subjective states, but rather to point out a way of getting at some subjective characteristics that, in principle, totally differs from asking survey questions.

10. For one attack in the case of nominal scales, unstructured polytomies, see Sutcliffe (1965). Not only dichotomous or polytomous responses call for special treatment. Consider observations on a circle (date of year, time of day). Use of differences as if this were an ordinary metric would in general be misleading, since December 29 (day 363) and January 2 (day 2) are really close together. Observations on a sphere, as in geography, again call for special treatment. Observations on a torus, such as bivariate observations with both coordinates separately cyclic, also arise from time to time. Observations on the simplex are fairly common (nonnegative numbers summing to 1 or 100: for example, percentages of components of rocks, blood, working forces, how you spend your time). Ordinary variances and covariances may be relevant for measurements on a simplex but the built-in dependence cannot help but affect analysis of error. Finally, we note that this sketch omits at least one other important kind of measurement, ordinal scales.

We also wish to note a more prosaic problem with regard to describing error structures from first principles: namely, missing observations. Suppose that an ordinary numerical observation is simply missing. Then the corresponding set of possible values is the line plus a disconnected point, and we might build up our theoretical structure by thinking of the probabilities for all pairs of numerical true value/numerical observed value *together* with pairs of numerical true value/fact of missing observation. It may be useful to think of several varieties of missing observation: sampled unit not found, found but refused to answer, answer illegible, answer impossible (for example, age of 226), answer lost in processing, and so forth.

These categories typically form an unstructured polytomy (or possibly an ordered polytomy). Working with the many questions of a real survey, one would presumably distinguish between the case of no answers at all from the case of a potential respondent and some answers. Much more might be done with this (see the report of Panel on Incomplete Data, 1983), but for present purposes it is enough to recognize the possibilities.

There is a substantial, although puzzlingly incomplete, literature on fundamental measurement issues: see, for example, Coleman (1964:Ch. 2); Pfanzagl (1968); and Roberts (1979). The discussions range from anecdotal descriptions to abstract axiomatizations. We found little that dealt with observational error beyond the standard statistical survey expositions. (An important exception is Adams, 1965, 1966.)

11. As pointed out earlier, measurement modes only become accepted insofar as they give results coherent with scientifically interesting and useful theories. For a stimulating and somewhat iconoclastic treatment see Adams (1966:148–153).

We may legitimately wonder whether social science measurements are fundamentally different from those in the physical and biological sciences. Two sorts of differences are frequently mentioned. First, in the social sciences, the measured unit (person, household, and so on) typically knows a measurement program is going on, which affects responses. (While this is often the case, it is by no means invariably true. In biological science, the same sort of feedback may occur. So the differences might be characterized as ones of degree.) Second, the social sciences have far less developed theoretical structures, and so dealing with measurement problems may be different. (By and large this is true, although with important exceptions. We return briefly to this topic in Chapter 6.)

12. Of course, measuring the effects of these other sources of variation will usually increase the cost of the survey.

13. It used to be thought that only weak levels of measurement were possible in psychology. So-called fundamental measurement (Campbell, 1920) seemed to obtain only in the physical sciences. Fundamental measurement is based on the object relation of concatenation. If objects A and B from a potentially infinite set of objects can be concatenated to define a new object $(A.B)$, and a measure of a physical property $(A.B)$ is always equal to the sum of the corresponding measures of A and of B, then this measurement is fundamental, yielding a scale with well-specified interval properties. A classic realization of this idea is for the measurement of length. If the objects are sticks, for example, and the concatenation consists in laying sticks end to end, then the length of the concatenate $(A.B)$ is in principle always equal to the length of A plus the length of B. It is thus possible to define a constant scale unit in terms of a stick of unit length that can concatenate with any other stick. Since it is hard to imagine the concatenation of psychological properties (for example, of intelligence), psychology was thought doomed to weak levels of measurement.

Fairly recently, however, it has been realized that more subtle relationships than concatenation can be defined for psychological entities, and under appropriate conditions they may lead to measurement on interval scales. The major such development is "conjoint measurement" (Luce and Tukey, 1964; Krantz and Tversky, 1971). The operation of conjoining arises when the levels on two distinct independent variables are simultaneously applicable to a set of observations. If certain plausible axiomatic properties apply to the (potentially infinite) set of possible conjoinings, then strong measurement is achievable. In practice, conjoint measurement depends on the absence of statistical interaction among the independent variables.

It would take us too far afield to give more details of the axiomatic foundations of psychological measurement. The interested reader will find an extensive literature on this topic elsewhere (for example, Krantz et al., 1971; Coombs, Dawes, and Tversky, 1970:Ch. 2; Pfanzagl,

1968; Roberts, 1979).

14. There are several mathematical models of the relationship of underlying ability to success on a given item (see Rasch, 1960; Lord, 1975; Hambleton and Cook, 1977). They share the assumption that the greater an individual's ability, the higher the probability of getting the item correct. This assumption is usually plausible, as far as it goes, but because of its probabilistic nature there is some inferential uncertainty. The test designer hopes that the imperfections of different test items are different and that errors introduced by individual items will tend to cancel out in large batches of items.

15. These points can be seen concretely by supposing that the frequencies in a sample are as follows:

	Actual	
Predicted	Vote	Not Vote
Vote	F_{11}	F_{12}
Not Vote	F_{21}	F_{22}

Perfect absolute prediction obtains if $F_{12} = F_{21} = 0$, that is, no individual prediction is mistaken. Apparent absolute prediction through compensating error occurs if $F_{12} = F_{21} \neq 0$, because the aggregate *predicted* vote ($F_{11} + F_{12}$) is numerically equal to the aggregate *actual* vote ($F_{11} + F_{21}$). Good relative prediction occurs if the odds ratio $F_{11}F_{22}/F_{12}F_{21}$ is large, regardless of the relationship between F_{12} and F_{21}.

16. A broader view might seek to embed authoritarianism in a general cultural framework, explaining the relationships to social class, intelligence, and so forth not as artifacts, but as intrinsic parts of the theory of authoritarianism. A relatively successful recent example of a broad construct bridging psychological and sociocultural factors is the "new urban black" identity explored by Sears and McConahay (1973). They postulated a constellation of pride in black identity, hostility toward white authorities, and racial militance would more readily explain participation in urban riots than would factors of economic or educational deprivation. Their construct of a new urban black identity, socially evolved in the younger generation in urban black neighborhoods, implied that participation in militant protests should correlate with youth and with having been born in the city (rather than being older, or being a rootless new arrival from the South). It also implied that there should be little or no correlation with income and education. These hypothesized relationships were validated in survey data gathered in Watts in Los Angeles.

5

Variability in Survey Measurements of Subjective Phenomena: Empirical Results

5.1 Response Effects With Subjective Survey Questions

The fundamental unit of meaning in a survey is the single question. There are smaller units to be sure, individual words being the most obvious, and in this chapter we will give examples showing how their variation can affect meaning. There are also larger units, primarily the scales and indices frequently employed in academic research, though much less often in poll reports by commercial organizations. But the discrete question plays a role in survey research analogous to that of atoms in chemistry; it is the distinctive unit of interaction in the course of administering a survey and the necessary building block in any analysis of data. To carry the analogy fur-

ther, the words, phrases, and other parts of questions can be thought of as the subatomic particles of the survey process, while sets of questions and responses that are turned into indices can be thought of as the molecular level.

Despite the importance of individual questions to the whole survey process, there is not much scientific knowledge about questions. Psychometric texts (for example, Nunnally, 1978) skip entirely or treat awkwardly the single question, preferring to move as quickly as possible to the technically more tractable process of combining questions. Survey texts (for example, Babbie, 1973) are apt to spend some time on how to write good questions and identify bad ones, but with primary emphasis on common sense and practical experience. In neither place are the effects of different forms of questions treated systematically, no doubt in large part because the information does not exist.

What experienced survey users do recognize is that although the individual question is the unit of meaning within a survey, the set of answers to such a question normally lacks clear interpretation when treated alone. This is because a small change in the wording of almost any question can alter the distribution of responses to it—sometimes substantially.[1] For this reason, experts rightly advise that the exact wording of a survey question should always be given when results are presented so that a user can make some judgment as to what the results mean. Unfortunately, even where the exact wording of a question is provided, it is seldom possible for even an experienced user to know how the responses might have been different if the wording had been different. Only empirical testing of such variations in particular populations can address this issue directly.

SUBSTANTIVE AND NONSUBSTANTIVE CHANGES IN WORDING

The assertion that "a small change in wording of almost any question can alter the distribution of responses to it" deserves close examination. Two kinds of wording changes can be distinguished: substantive and nonsubstantive. Substantive changes occur because of the fact that almost all social issues and internal feeling states are complex, yet individual survey questions must necessarily be kept simple. Thus a single question can usually tap only one facet of an issue. For example, the following two questions both deal with gun control, yet they would yield very different distributions of responses because they pose the issue in very different terms:

1. Would you favor or oppose a law that would require a person to obtain a police permit before he or she could buy a gun?
2. Would you favor or oppose a law that would ban possession of guns, except by police and military authorities?

Just as legislators might vote differently on these two questions if posed as possible laws, so respondents also vote differently when they are asked one or the other in a survey. Yet the answers to either of the questions might be cited as bearing on the general issue of gun control. Based on past data (see Smith, 1980a), we can be reasonably sure that the first would yield a clear majority supporting gun control, the second a clear majority opposing gun control.

It should also be evident that the "smallness" of a change in question wording is a matter of judgment about shifts in meaning, not a matter of the number of letters, words, or phrases altered. If question 1 were revised to 1′ by changing the word "buy" to "fire," then 1 and 1′ would differ in only one word, and yet one would expect a large difference in results if the two versions were administered to comparable samples of the American population. In many cases, therefore, a "small change in wording" involves an unmistakable substantive alteration in meaning and points to the sensitivity, rather than to the lability, of survey questions. In other words, the problem of question wording is not one rightly attributable to either survey questions or survey respondents, but rather to a strong tendency by survey users to be tempted by overly simple summaries of public attitudes and beliefs. What is called an issue is often really a constellation of issues, and the separable issues (for example, gun permits versus banning of guns) may be of at least as much interest as the overall rubric.

Most so-called small changes in question wording are probably substantive. However, there is also compelling evidence that what appear to be nonsubstantive alterations in questions can lead to important shifts in response distribution. One of the most striking examples was reported by Rugg (1941) using split-ballot data collected by Elmo Roper in 1940.[2] As shown in Table 5.1, respondents were considerably more willing to "not allow" speeches against democracy than they were to "forbid" such speeches. Although it may be true that the connotations of the terms "not allow" and "forbid" differ, their operational consequences for making speeches are the same—hence, logically the response distributions should not differ. Furthermore, this particular response effect has been replicated experimentally quite recently, indicating that it was not limited to the 1940 Roper Poll in which it was first discovered (Schuman and Presser, 1981). Its exact cause and its scope are not fully understood—the forbid-allow difference occurs with some but not all other issues—but the findings suggest that nonsubstantive changes in question wording can substantially influence responses.

Effects due to nonsubstantive changes in questions are not restricted to substitutions of one word for another. The order in which alternatives are read to respondents can also affect answers noticeably, as indicated by an example reported by Payne (1951) and shown in Table 5.2. Respondents are

TABLE 5.1
Forbid and Allow Experiment

"Do you think the United States should forbid public speeches against democracy?"		"Do you think the United States should allow public speeches against democracy?"	
Forbid ("Yes")	54%	Not Allow ("No")	75%
Not Forbid ("No")	46%	Allow ("Yes")	25%
	100%		100%

SOURCE: Rugg (1941).

NOTE: Exact Ns are not provided by Rugg, but information is given in Cantril (1940). Each version of the question was asked of about 1,300 respondents; hence, the difference is highly reliable. "Don't know" responses did not differ significantly by the form of the question and are omitted here.

more likely to choose whichever of the two alternatives is read last. Again, this instance of a "recency effect" has been replicated in the past several years, and there can be little doubt of its reliability over the three decade period (Schuman and Presser, 1981). While there is some evidence that such recency effects occur more commonly with complicated questions (Payne, 1951), there are also important exceptions that make generalization (or prevention) difficult.

The examples cited thus far indicate why the reported distribution of

TABLE 5.2
Results of Experiment on Response Variation Induced by Ordering of Response Categories

Ordering A		**Ordering B**	
"Which of these statements comes closest to your own idea of how gasoline and oil prices are decided: Each company sets its own prices to meet the competition, or the oil companies get together and set prices for their product?"		"Which of these statements comes closest to your own idea of how gasoline and oil prices are decided: The oil companies get together and set prices for their product, or each company sets its own prices to meet competition?"	
Competition	32.4%	Collusion	59.8%
Collusion	67.6	Competition	40.2
	100.0		100.0
	(445)		(214)

NOTE: The experiment was first reported by Payne (1951), but Payne simply reported a "significant" recency effect, without giving figures. The data above are from a replication by Schuman and Presser (1981) in April 1979; for these data, the likelihood-ratio chi-square statistic is: $L^2 = 3.86$, d.f. = 1, and p is less than .05.

responses to a poll or survey question does not necessarily have clear meaning when considered in isolation. Any one question quite likely taps only one facet of a multifaceted issue and if it were rewritten to tap another facet, the results might be very different. There may also be nonsubstantive aspects of the question that affect responses, that would be difficult to anticipate, and that are hard to explain.

Three qualifications to these conclusions are required. First, it is not the case that *any* change in wording will alter survey response distributions. For example, an anti-abortion advertisement in the *New York Times* in 1974 claimed that support for legalized abortion was lower when the word "abortion" was used in a question than when the words "end pregnancy" were substituted. An experimental test of this proposition, however, failed to show any evidence for such an effect (Schuman and Presser, 1981). There are a number of other instances in which rather noticeable changes in wording (or in response order) do not seem to change response distributions. Thus the problem is not that *any* change in wording affects responses, but that many do and that it is often impossible to predict in advance which changes will have effects and which will not.[3]

A second qualification is that, while the size of effects due to substantive modifications is probably unlimited—one could change percentages radically with appropriate modifications—most effects from nonsubstantive changes are less than 20 percent and often less than 10 percent. (The forbid/allow difference reported in Table 5.1 is probably the largest nonsubstantive wording effect ever discovered.) That is, if X percent of respondents endorse a particular position, it is extremely unlikely to vary by more than ± 20 percent because of a nonsubstantive variation in question form or wording. This hardly allows for much precision when stating a result in percentage terms, but it does tell us that, say, 75 percent "pro" is almost certainly not going to become 75 percent "anti" as a result of a nonsubstantive change in question wording.

The third important qualification concerns the distinction between substantive and nonsubstantive changes in wording. Although the distinction is often clear, there are instances in which it is not. Consider the following question asked by Gallup during the Vietnam War:

> If a situation like Vietnam were to develop in another part of the world, do you think the United States should or should not send troops?

Mueller (1973) noted that when the "same" question was asked with the addition of five words (". . . to stop a Communist take-over?"), the percentage agreeing to American intervention rose by some 15 percent, a difference later replicated several times experimentally (Schuman and Presser, 1981). It is not clear whether this effect on responses should be regarded as

a substantive one, in which mention of a Communist take-over specifies the situation more precisely, or is rather an emotional coloring to the question that is best considered nonsubstantive. It is possible that both occur; in any case, this example cautions us against assuming that alterations in questions can always be neatly classified as substantive or nonsubstantive. Our further discussion focuses mainly on what appear to be nonsubstantive alterations, but this caution should be kept in mind.

QUESTION CONSTRAINT AND QUESTION CONTEXT

Two other sources of response effects also cross the line between substantive and nonsubstantive alterations in questions. One reflects a fundamental fact about all survey questions: survey questions should be seen as implying rules according to which a respondent is constrained to choose only among the alternatives offered. Thus if an attitude question does not include a "don't know" alternative, in most cases respondents will not attempt to assert don't know, even though many would do so if the option were offered (Schuman and Presser, 1981). To some extent, the same thing is true of other alternatives that respondents demonstrate a liking for when they are available: few persons violate what they take to be the rules of the game by insisting on a choice that has not been presented explicitly to them. Even when a category of "other" is included among a list of alternatives, it is seldom used as a way of giving responses not otherwise offered (Lindzey and Guest, 1951). Thus the constraint imposed by the question is a very powerful one, and the investigator's decision to offer or omit an alternative has considerable influence in shaping the resulting response distribution (see Chapter 7 for further discussion).

It might seem that this problem could be solved by greater use of "open questions"—that is, questions that do not provide fixed alternatives but instead invite free answers from respondents, the latter being coded at a later point into summary categories. Such an approach has much to recommend it as a way of understanding more adequately the frames of reference of persons different in situation and outlook from the investigators, and it probably should be used more often in surveys than it is today. But open questions also entail very severe problems. The frame of reference is often so wide that responses are not comparable to one another and contain much that is irrelevant or ambiguous (Campbell, 1945), and these problems are accentuated by frequent failures in probing and recording by interviewers and misinterpretations on the part of coders. Moreover, it is not always true that open questions are less constraining than closed questions, for the provision of alternatives can widen as well as narrow the meaning of a question (Schuman and Presser, 1981). Given these and other problems, the primary value of open questions is at the early developmental stage of

a survey (McKennell, 1974), and it is unfortunate that survey investigations (and their sponsors) do not more often invest resources in this type of pilot work—and provide documentation of it in final reports.

A second source of response effects that is even less obvious than that already discussed arises from the context in which a question is placed. Such questionnaire context effects are not as well documented as is often assumed, nor probably as pervasive, but there are a few striking examples, one of which is shown in Table 5.3. In this case, a standard item on attitudes toward abortion from the NORC's General Social Survey (National Opinion Research Center, 1978) was borrowed for use in a Survey Research Center (SRC) survey. Routine comparison of the marginal results from the two surveys yielded the difference shown in Table 5.3—a difference that seemed too large to be explained by time of administration or other variations between the two surveys. The hypothesis of context effects was suggested by the fact that in the NORC survey the item followed a more specific question on attitudes toward abortion in the case of knowledge that the unborn child was defective, while in the SRC survey the item had been the only one about abortion. Subsequent split-ballot experiments provided strong evidence in support of the context hypothesis, since the difference shown in Table 5.3 was reproduced when the variation in context (but only in context) was created (Schuman and Presser, 1981).

There is some further, much more tentative evidence that the abortion item was particularly vulnerable to context effects because of its relatively general nature. It attempts to summarize a complex issue by means of a single general question and thereby creates an ambiguous frame of reference that is especially open to influence from previous questions and responses. In contrast, the more specific item about abortion in the case of a defective child did not show response variation by context in these experiments. (For other examples of general but not specific questions showing

TABLE 5.3
Discrepancy between NORC and SRC Results on General Abortion Question

		NORC Spring 1978	SRC March 1979
"Do you think it should be possible for a pregnant woman to obtain a legal abortion if she is married and does not want any more children?"	Yes	40.3%	58.4%
	No	59.7	41.6
		100.0	100.0
		(1,483)	(490)

NOTE: For this table, the likelihood-ratio chi-square statistic is: $L^2 = 48.34$, d.f. = 1, $p < .001$.

context effects, see section 5.3 of this chapter and McFarland, 1981; Sigelman, 1981; Smith, 1979; and Turner, Chapter 7, Volume 2.)

Yet, other instances of context effects seem to involve other underlying processes (as in the "reporter" example given later in this chapter), and the whole problem of question order is still in need of considerable research and systematization. Furthermore, there are numerous instances for which context might be expected to have an effect, but for which none appears when question order is manipulated experimentally. (For example, see the discussion of Table 5.4 in this chapter.) There are also cases where a nonexperimental difference between surveys seems to fit nicely an explanation in terms of context, but where experimental test fails to confirm the explanation (see Turner and Krauss, 1978 and Turner, Chapter 7, Volume 2).

INTERVIEWER AND OTHER SOURCES OF RESPONSE EFFECTS

Although we have now covered the major sources of systematic response effects due to questions themselves, there are other variations in survey procedures that influence responses in ways likely to go unnoticed without experimental manipulation. For example, there is convincing evidence that the race of survey interviewers influences answers to race-related items (Hyman et al., 1954; Williams, 1964; Hatchett and Schuman, 1975–76). Questions dealing directly with racial antipathies are the most sensitive in this respect: black respondents are less likely to express explicit distrust of whites when the interviewer is white, and white respondents mute negative sentiments about blacks when the interviewer is black.

A somewhat different type of example is shown in Table 5.4: a question that did not explicitly deal with racial issues revealed a substantial implicit effect due to race of interviewer. The example is of particular interest because it involved a 2 by 2 factorial design with question context as well as race of interviewer varied experimentally. Black respondents were asked: "Could you tell me who two or three of your favorite actors or entertainers are?" In half the interviews the question was located at the end of a long series of racial attitude questions, while in the other half it was placed earlier in the interview, before any racial questions were asked. The investigators expected the former context to stimulate more mentions of black entertainers, either as a main effect or in interaction with the race of the interviewer. However, the results failed to show any reliable evidence of a context effect, while the race of the interviewer did exert a powerful influence on answers. The special importance of the example is to show that a question can be quite sensitive to one type of response effect and yet quite insensitive to another. This result suggests that the problem of response effects can be more complex than that of merely identifying especially vul-

TABLE 5.4
Race of Mentioned Entertainers by Survey Context and Race of Interviewer

"Could you tell me who two or three of your favorite actors or entertainers are?"

Race of Entertainers Mentioned	Black Interviewers		White Interviewers	
	Form A[a]	Form B[b]	Form A[a]	Form B[b]
Blacks only	45.5%	39.1%	14.8%	19.2%
Mainly black	22.8	29.9	22.2	25.0
Black and white equally	8.9	6.9	11.1	7.7
Mainly white	14.9	18.4	29.6	25.0
Whites only	7.9	5.7	22.2	23.1
Total[c]	100%	100%	100%	100%
(N)	(101)	(87)	(54)	(52)

NOTE: The question was administered in the 1971 Detroit Area Study to black respondents only. Fit of alternative models to these data was

1. Response by Race of Interviewer: $L^2 = 31.52$, d.f. = 4, $p < .001$.
2. Response by Form (context): $L^2 = 1.68$, d.f. = 4, *ns*.
3. Response by Form by Race of Interviewer: $L^2 = 1.55$, d.f. = 4, *ns*.

(If Response (race of entertainer) is tested for a linear trend, with 1 d.f., the above three likelihood-ratio chi-square statistics become, respectively: 26.52, 0.01, and 0.03.)

[a]The entertainer question followed nonracial questions.

[b]The entertainer question followed a set of racial questions (see Duncan Schuman, and Duncan, 1973, for the wording of these questions).

[c]Columns do not always sum to 100.0 percent due to rounding.

nerable questions or particularly susceptible respondents; instead, there may be important theoretical differences between the various sources of response effects. (We return to this general problem later in this chapter in our consideration of factors that distinguish the sources of random and systematic effects.)

Color is the only visible interviewer characteristic that has been shown repeatedly to produce large systematic effects on survey responses, but this may be partly because there are fewer experimental investigations of other social attributes, such as age and sex. There is also some evidence that mode of survey administration (for example, telephone versus face-to-face interviewing) has some systematic effects on response, though these appear to be less common than might have been expected in advance. (See Bradburn et al. [1979] and Groves and Kahn [1979] for experimental data and Sudman and Bradburn [1974] and Singer [1979] for general discussions.) All told, any set of answers to a single administration of a survey question might best be regarded as a single treatment in an incomplete experiment. Other treatments—variations in the question itself or in the circumstances of its administration—might produce different results.

QUASI-FACTUAL QUESTIONS

Our discussion thus far has dealt largely with survey questions intended to measure attitudes and beliefs. It might be thought that problems of response effects could be avoided by restricting surveys to factual (objective) questions. However, as we discussed in Chapter 1, the concept of fact in this sense is an ideal; in practice many survey questions labeled "factual" require answers that have an important subjective element. This is true to a small extent even for such an apparently straightforward objective variable as sex of respondent, as controversies over participants in women's sports have indicated. When we turn to completely social concepts like ethnicity, unemployment, and housing quality, we are dealing with variables that are better labeled quasi-factual or even semi-attitudinal. Chapters 4, 5, and 6 in Volume 2 discuss these three topics, illustrating the difficult issues of conceptualization and measurement in such cases. When concepts are inherently imprecise, or when there is a considerable gap between an event and its report (as in retrospective surveys), it is likely that there are also some of the same problems of question form, wording, and context that we have considered.[4]

A striking example of an apparent context effect on seemingly factual items is reported by Gibson et al. (1978) and Cowan, Murphy, and Wiener (1978). The National Crime Survey, conducted by the Census Bureau in 26 cities from 1972 through 1975, obtained respondent reports of victimization experience during the preceding 12-month period. On a random half of the interviews the reports followed 16 attitude questions about crime; on the other half the attitude questions were omitted. Victimization reports, especially of less serious crimes, increased significantly and appreciably for the subsample of respondents who were asked the attitude questions. The most plausible interpretation of the difference is that the attitude questions stimulated memory for and willingness to report more experiences, though some degree of misreporting cannot be ruled out. It is interesting to note that the attitude questions had been placed first in the survey on the common assumption that they, not the factual questions, would be sensitive to context.

COPING WITH SYSTEMATIC RESPONSE EFFECTS

The kinds of problems discussed in the first part of this section make it clear that interpretation of a response distribution for a single question is uncertain at best. Although popular presentations of survey results often ignore this fundamental truth and report that X percent of a national sample feel such-and-such, experienced survey analysts attempt to deal with the problem of response effects by abandoning absolute interpretations and relying on comparative strategies of one kind or another. We will now consider

several of the main strategies, what is known about their success in limiting response effects, and certain necessary precautions in their use.

Comparisons Over Time. For an investigation wishing to say something that captures a little of the absolute level with which an attitude or belief is held by a total population, the best approach is probably to look at trends over time within that population. A 20 percent increase over, say, 10 years in choice of a particular alternative to an attitude question provides direction and a rough sense of the amount of change from the baseline.[5] Such a comparison, of course, requires that exactly the same question be repeated. If the "forbid" version from Table 5.1 had been asked at time 1, a comparison with the "allow" version at time 2 would obviously be misleading, since the time effects would be confounded with wording effects. What an investigator does hope is that both the "forbid" and "allow" versions show the same trends, so that replication of one will lead to the same conclusions as replication of the other, at least when nonsubstantive question variations are involved. In fact, this is exactly what happened for the forbid-allow experiment when it was repeated in the mid-1970s: support for the civil libertarian position (allow/not forbid) had increased appreciably over the 35-year period, but the magnitude of the increase was essentially the same for the two versions of the question (that is, there was not a statistically significant interaction of response, time, and question version). Thus, in this unusual instance where we could test directly the frequent assumption that different question forms will yield equivalent estimates of change over time, the assumption was sustained.

Occasionally an analyst feels it necessary or desirable to compare response distributions for questions that differ in wording in what are presumed to be nonsubstantive ways. From the examples already discussed, it is evident that this is a dangerous practice. One safeguard is for an investigator to include two forms of a question as part of a current split-ballot experiment in order to be sure that the wording variation makes no difference, though this in turn assumes that the wording effect itself does not interact with time. In any case, all departures from perfect constancy of wording must be regarded as possibly confounding and should be made only with clear recognition of that possibility.

Holding the wording of questions constant is necessary, but not always sufficient, to prevent incorrect inferences about change. Had the differences in response to the abortion item shown in Table 5.3 appeared in surveys administered 10 years apart, it would have been natural to attribute them to time-related shifts in attitudes rather than to questionnaire context. In order to guard against this possible misinterpretation, either the items that are used must be invulnerable to context or the context must be held constant. Such advice is easy to give, but difficult to follow, since our

understanding of context is too slight to identify context-free questions (if such exist) or to specify relevant contexts in any firm a priori way. What does seem to be true is that context effects almost invariably involve questions that are closely related in terms of substantive content. Investigators should therefore be especially wary of moving individual items out of, or into, sets of related items when attempting a replication. Having constructed the General Social Survey series on abortion, for example, NORC and others interested in measuring change should administer the total set of (six) items in the future, or at least drop off only those items that appear at the end of the series. Thus, constructing an initial set of social indicators involves long-term decisions and commitments about sequencing of items, not merely about individual questions.

Other context-like features of surveys may also compromise trend analyses. Trend studies of racial attitudes have been compromised somewhat by wide variations in the racial composition of interviewing staffs that have occurred over the past 20 years. And recent shifts from face-to-face to telephone interviewing at least raise the possibility that discontinuities in trends may be due to method rather than to attitude change. In all these cases, the best insurance against misinterpretation is the split-ballot type of experiment that, at least for one point in time, allows a direct test of the effect of the method variation on response.

Although it is usually plausible to assume that by holding question, context, and other circumstances constant one can draw unambiguous conclusions about change, it is useful to present an example that shows that even this strong assumption can be wrong. The example is available because of a split-ballot experiment carried out in 1948 by Hyman and Sheatsley (1950) and replicated in 1980 by Schuman and Presser (1981). (The latter reference contains data for the full experiment, only part of which is presented here.) The following question about "Communist reporters" was posed to a national sample:

> Do you think the United States should let Communist newspaper reporters from other countries come in here and send back to their papers the news as they see it?

Half the questionnaires in each year asked this question after another question on whether the Soviet Union should allow in American newspaper reporters; the other half asked the two questions in the reverse order. Table 5.5 shows the results of the experiment in 1948 and again in 1981.

In both years there is a strong context effect: respondents are more willing to allow Communist reporters into this country after having been asked the American reporter question, presumably because (as Hyman and Sheatsley noted at the time) a norm of reciprocity is made salient by this ordering. What is of greater interest for our present discussion, however, is

TABLE 5.5
Percentage Responding Yes to Communist Reporter Question in Split-Ballot Experiments over Time

	Order of Questions	
Year	**Communist/American[a]**	**American/Communist[a]**
1948	36.5% (581)	73.1% (635)
1980	54.7% (342)	74.6% (335)

NOTES: Base Ns are presented in parentheses.
[a]"Communist/American" order refers to subsample receiving Communist reporter question first. "American/Communist" order refers to subsample receiving Communist reporter question after having been asked a question about allowing American reporters into the Soviet Union.

that the context effect itself interacts with time, so that even with context held constant one would draw different conclusions about the direction of change from the two contexts (the likelihood-ratio chi-square for 3-way interaction of Response by Context by Year is: $L^2 = 10.07$, d.f. $= 1$, $p = .01$). When the Communist item appears first in the sequence, there is a decrease of 18 percentage points over the 32-year period in the negative response toward Communist reporters. But when the Communist item appears after the American item, there is essentially no change in response over the same period (in 1948, 73.1 percent respond "yes" versus 74.6 percent in 1980).[6] The reasons for this difference in trends are perhaps not hard to understand, given knowledge of the experimental variation at both points in time. But without the replicated experiment, a user would be unlikely to imagine one set of results if only the other were available. This example may well be the exception to the simpler rule about similar conclusions occurring when context is held constant, but one must be warned not to bank on it.

Form-Resistant Correlations. Although trend studies have been singled out as a separate type of analysis because of their important and frequent use in presentation of poll or survey results, they are in fact simply one example of standard survey analysis. Monthly Gallup Reports, for instance, regularly present breakdowns of data by demographic and social categories such as age and religion. In all such analyses, the user ordinarily assumes that the associations would not change as a result of nonsubstantive alterations in the question, and indeed it has sometimes been assumed that even substantive alterations of the kind discussed at the beginning of this section will not lead to important changes in associations. This assumption of "form-resistant correlations" is occasionally stated explicitly by survey specialists, as in the following quotation from Stouffer and DeVinney (1949:168):

> . . . one must be careful to focus attention on differences in percentages among different categories of men with favorable attitudes on a given item, not on absolute percentages. The fact that the percentages saying the army was run pretty well or very well are large does not mean, necessarily, that so many men were actually favorable to the army—such percentages are artifacts of question wording and of the checklist categories arbitrarily selected as "favorable". . . . But when we focus on *differences* in percentages responding favorably to the same questions, among men in different categories, the differences can be meaningful in a sense in which the absolute values cannot. [emphasis in original]

Although the assumption has seldom been tested explicitly, a series of recent experiments provides a fair degree of support for it (Schuman and Presser, 1981). Formal changes in question wording, such as omission or inclusion of "don't know" alternatives to questions did not usually lead to changes in associations with other items or with background variables. There are some very interesting exceptions to this generalization, and one certainly cannot trust blindly in it, but much more often than not it was possible to alter wording in a way that affected item marginals without significantly changing the association of the item with other items or with background factors. When associations did vary significantly, the variation was sometimes one of magnitude but not of direction. For example, the "allow" version from Table 5.1 shows a significantly stronger relation to education than does the "forbid" version, but in both cases the association is statistically significant and in the same direction (support for the civil libertarian position increases with increasing education).

The most likely systematic exception to the preceding generalization has to do with items in the form of statements. There is some evidence that such statements, to which respondents are asked to agree or disagree, may create their own nonsubstantive association with education (Campbell et al., 1960). Data directly on this point are sparse and often difficult to interpret, but it may be that for such types of question the assumption of form-resistant correlations does not hold at all. Whether such an acquiescent tendency also occurs with interrogative forms is even more uncertain at this point. (For a more complete review, see Schuman and Presser, 1981.)

Multi-Item Strategies. Many of the problems we have discussed thus far can be seen as inherent in reliance on individual items. But in all such cases it is difficult to know what aspect of the item determines the reponse. By considering combinations of items together, users can hope to separate out those features in which they are particularly interested. Yet the use of combinations of items has its own difficulties, especially in large-scale surveys for which costs and respondent impatience make extensive psychometric development virtually impossible.

If items are moderately intercorrelated and conceptually related, a standard practice of academic analysts is to add the items together in some fashion in order to construct an overall index. In addition to the practical convenience of working with a single attitude variable rather than a large set of separate items, such an index creates a dimension based on whatever the component items have in common that led to their intercorrelations. Idiosyncratic features of the individual items tend to be eliminated in this process. Thus responses determined by a particular word ("allow" versus "forbid") or a particular context no longer play an important role. For example, if the increase in "yes" responses to the Communist reporter item shown in one context in Table 5.5 reflects a general drop in fear of Communist espionage, then a set of similar items is quite apt to pick this up, even if one or two of them happens to show no change because of contextual factors. This is a considerable advantage, and thus an index is able to capture broad trends that are lost in the peculiarities of separate items. Associated with this gain, moreover, is an increase in reliability of measurement, since individual people are more accurately located on the larger dimension than they can be on the individual items for which both specific factors and chance are free to play greater roles.

Yet substitution of an index for separate items involves some sacrifice as well, especially in surveys for which methods of construction and types of subject matter are rarely close to textbook ideals. For one thing, it is not always desirable to substitute a larger concept for the more specific issues dealt with by individual questions. An example of two questions on intermarriage is presented in Table 5.6. While it may be desirable for some

TABLE 5.6
Two Questions on Racial Intermarriage

"Do you approve or disapprove of marriage between whites and nonwhites?" (July 1978)		"Do you think there should be laws against marriage between (Negroes/blacks) and whites?" (Spring 1977)	
Approve	32%	Yes	38.0%
Disapprove	58	No	58.8
Don't know	10	Don't know	3.2
	100		100.0
	(1,335)		(1,352)

NOTE: The questions were asked of white respondents only. The left-hand column percentages are taken from the Gallup Opinion Index, November 1978. The right-hand column percentages are calculated from the NORC 1977 General Social Survey (National Opinion Research Center, 1978). The difference in time and in terminology for color are unlikely to have affected the percentages by more than a few points. The difference in "don't know" percentages mainly reflects a general Gallup-NORC difference in this type of response, rather than being peculiar to these questions (see Smith, 1979).

purposes to combine the two items into a more general index on attitudes toward intermarriage (or to put the items together with others on racial issues into a still more general index of, say, interracial attitudes), reliance on such an index alone loses the unique substantive information contained in the original items. Eliminated along with possibly unwanted wording and context effects are substantive aspects of issues that are of real importance. Interesting as the social scientist's larger concepts ("prejudice," "authoritarianism," and so forth) may be, the more concrete level at which issues are defined in the society also has important meaning.

Furthermore, although a multi-item index tends to eliminate unwanted idiosyncrasies, it does not eliminate—but rather cumulates—unwanted systematic sources of response effects. We noted earlier that the use of the agree/disagree format for questions may tend to create correlations with education, and if so, an index constructed from a set of such items is no less vulnerable to the same problem. Indeed, the increased reliability of such an index could stem from just this cumulation.

In sum, the typical attitude index functions as a kind of rough averaging procedure that eliminates *both* desirable and undesirable features of individual questions. Ideally one would like the benefit without the loss, and there have been techniques that claimed to offer this (for example, Guttman's unidimensional scaling); however, they have not proved very successful in practice in part because sets of items rarely meet the requirements imposed by the theoretical models.

The situation is further complicated in attitude survey analysis by the strong interest in tracing change over time. Even when items form a seemingly coherent index at one time, social change may lead to trends that not only differ among items, but do so in a way that is of paramount importance to understand (Converse [1972]; see also Wendt [1979] and Piazza [1980] for a discussion of similar problems with other independent variables). If individual items are buried within an index, such crucial information will be missed.

Perhaps the main difficulty lies in the assumption that a set of attitude items is like a set of words in a spelling or vocabulary test. The latter can be thought of as a sample of more or less interchangeable elements from a population, even if the population cannot readily be enumerated. But attitude questions, especially those dealing with public issues, are much less interchangeable. Each one is apt to have unique meaning, so that its distinctive determinants are of importance in themselves and not merely because they point to a single concept of interest to an investigator. (Theoretically, one might like to multiply items around each single issue, but in practice a survey can't do this, again because of high cost to investigators and limited patience of respondents. Such a goal also assumes, probably wrongly, that elemental issues can be defined in a permanent way.)

Emphasis on the limitations of index construction is not meant to argue against the desirability of working with sets of items rather than with single item measures. At the most elementary level, comparison of the percentage distributions for two related items (for example, registering guns versus banning guns) is frequently useful. Likewise, comparisons of the correlates (including time) of separate items is typically illuminating, pointing to differences in meaning and implication that may not be apparent from wording or marginal differences.

For most purposes, analysis of separate items, of their relations to one another, and of summary combinations should probably go together in survey work, with judgments made as to the best mix given a set of items and a set of analytic aims. After all, the ultimate goal of analyzing survey data is to discover their meaning, and for this purpose an investigator would be foolish to throw away any source of information. It is also likely that in the future methods will be developed that allow analysis of item combinations to be pursued in a more regular and rigorous fashion than is now possible (see Chapter 6; Duncan, Chapter 12, Volume 2; and Duncan, Sloane, and Brody, 1982).

Systematic and Random Response Effects. It is often assumed that the kinds of systematic response effects we have been discussing occur primarily when attitudes or beliefs on a subject are not well crystallized. Rugg and Cantril (1944:48–49) operated on this assumption in their pioneering studies of question wording:

> The extent to which the wording of questions affects the answers obtained depends almost entirely on the degree to which the respondent's mental context is solidly structured. Where people have standards of judgment resulting in stable frames of reference, the same answer is likely to be obtained irrespective of the way questions are asked. On the other hand, where people lack reliable standards of judgment and consistent frames of reference, they are highly suggestible to the implications of the phrases, statements, innuendoes or symbols of any kind that may serve as clues to help them make up their minds.

It is interesting that much the same assumption has also been made about random response error, for example, by Converse (1970) in his argument that most unreliability over time in survey results is due to respondents who lack attitudes on an issue but who nevertheless give answers. Thus, although the connection is seldom made explicitly, both types of effects are being attributed to the same cause, namely, the absence of crystallized attitudes by some people on some issues that are asked about in surveys. Moreover, following this reasoning, by omitting either certain questions or certain respondents or both, one should be able to reduce substantially the

occurrence of both systematic and random response effects in attitude surveys. Unfortunately, Rugg and Cantril did not provide an independent operational definition of "crystallization," and therefore their examples are circular. When response effects appear, the absence of "opinion crystallization" provides the explanation. But the only indicators of lack of crystallization are the response effects themselves. Thus, although their emphasis on opinion crystallization is intuitively appealing, it cannot constitute a useful approach in its original form. Converse (1970) provides a more convincing model that points to evidence of lack of crystallization ("nonattitudes"), but also does not identify individual respondents directly in these terms.

One suggested way to assess lack of crystallization is to assume that it is reflected in the "don't know" level for items. Thus Gallup (1941:261) writes:

> . . . the "suggestive" effect of a . . . difference in wording is largely a function of the degree of crystallization of opinion on an issue, and tends to decrease to zero as the no opinion vote tends to zero.

Unfortunately this does not appear to be borne out in practice. Thus Payne (1951:135) did not find response order effects of the kind shown in Table 5.2 to be directly related to "don't know" levels, and there is only a little evidence of such an association across the various experiments on the effects of question form carried out by Schuman and Presser (1981). No similar analysis seems to have been done systematically for the relation of "don't know" levels to random response error, though Converse (1970) suggests that such a relation does hold for measures of reliability over time in panel data (see review by Smith, Chapter 9, Volume 2).

Another possible approach to providing a separate indicator of attitude crystallization is to use measures of self-reported attitude strength. Respondents who say that they do not think an issue is important, or who do not feel strongly about it, can be hypothesized to be especially susceptible to both random and systematic response effects when questions are asked about the issue. (Note that our attention here shifts from identifying susceptible questions to identifying susceptible respondents, but the same approach might be adapted to distinguishing among questions.) Moreover, there is evidence that simple response consistency (reliability over time) can be predicted well from initial attitude strength, that is, that random response is linked to low attitude strength (Schuman and Presser, 1981).

However, despite the theoretical and empirical reasons for expecting a negative correlation between attitude strength and systematic response effects, such correlations did not appear regularly in several recent experiments (for a more complete review, see Schuman and Presser 1981). The evidence is certainly not conclusive, but at present it points toward a sharp

distinction between sources of unreliability (random error) and systematic response effects. The former does seem to be a function of low attitude strength; the latter does not, but may instead often result from genuine changes in the perceived meaning of questions, as exemplified by the questions about Communist and American reporters, discussed previously. Further investigation of the similarities and differences between random response error and systematic response effects is needed.

Conclusions. In this section we have not employed the words "bias," "true value," or "systematic error," but rather the more neutral term "systematic response effect." We did so because a difference in results produced by a nonsubstantive question variation seldom allows the researcher to feel confident that one version of the item is closer to truth than another. Construct validation may provide evidence in certain cases, but for now it is probably better to concentrate on determining the types of form, wording, and context variations that have effects, the kinds of effects that occur, and the sources of the effects.[7] When effects do *not* occur, of course, we can feel somewhat reassured about the robustness of a particular result in the face of nonsubstantive alteration of the question.

It is unlikely that a single, general theory of response effects can be developed since the effects themselves seem to be of several quite different types. Some involve unintended changes in meaning (for example, the aforementioned reporter items), and in such cases the effect itself becomes a substantive finding that can increase an investigator's awareness of real social phenomena. From this standpoint, surveys are a richer, as well as a more complicated form of measurement than is popularly thought, and split-ballot experimentation allows one to exploit the survey method to improve our understanding of the subtleties of verbal response. Some effects, on the other hand, appear to be more mechanical in nature (for example, those due simply to the order of response alternatives within a question), and the goal in these cases must be to recognize when effects will occur and how they can be handled. Still other response effects, such as those due to omission or inclusion of a "don't know" response, seem to involve a constraint on a particular type of answer that respondents would give if it were offered, but the constraint affects only that response and rarely has wider implications for other uses of the item.

The theoretical distinctions just noted are among those suggested by the research now available. But such methodological research is still quite limited. Compared to the large amount of money spent on applied surveys in this country, the investment in methodological research on response effects has been relatively slight. Until such efforts are made on a larger scale, we cannot say how much our understanding can be improved in ways

that will lead to better construction and interpretation of surveys using subjective measures (see Chapter 10, particularly Recommendations 9 through 13).

We end this section by noting that the type of examples we have discussed can sometimes lead to despair on the part of those who tend to regard surveys as quite simple measuring instruments. But the despair is just as unwarranted as is the assumption of simplicity. The basic approach to learning about large populations by means of posing questions to relatively small probability samples is a sound one. It becomes complicated not because of artifacts unique to surveys, but because human responses are inherently both subtle and complex. Analysts need skill and experience in order to design surveys and interpret the resulting data in ways that will maximize insight and minimize misunderstanding: "Errors are not in the art but in the artificers" (Newton, 1687).

To provide some appreciation for the magnitude of the errors commonly found in survey measurements of subjective phenomena, the following section examines instances in which either the same survey was carried out in tandem by different survey organizations or in which putatively equivalent survey measurements were coincidentally made by different survey organizations. We then conclude this chapter by offering a demonstration of some strategies that may provide a basis for the future diagnosis and control of some common sources of measurement variation in surveys of subjective phenomena.

5.2 Reproducibility of Survey Measurements of Subjective Phenomena

An acid test of the reliability of a measurement is its reproducibility. If a measurement of a particular phenomenon can be reproduced by independent investigation under like conditions, then there is presumptive evidence that the measurement is, within a given degree of exactitude, reliable. Of course mere replication does not ensure perfection. Measurements reliable enough for one purpose may be too crude for other purposes. And of course, as we discussed in Chapter 4, finding that a measurement is reliable says nothing about its validity since, for example, constant error may be distorting all measurements. Reproducibility under like conditions is, however, a minimum criterion of scientific measurement.

In survey research, as elsewhere, it is difficult to carry out replications under identical conditions. Test/retest experiments (in which the same person is interviewed twice) suffer from the intervention of time and therefore

from the possibility of true change, memory lapses, interview effects, and sample attrition. Replication studies (using different samples) also suffer from the intervention of time, and they are frequently unable to duplicate precisely previous survey procedures. Perhaps closest to a strict replication are surveys conducted in tandem, when two data collection organizations simultaneously carry out identical surveys. Tandem designs control for all factors except intrinsic differences between the organizations if the design is strictly followed—which is frequently not the case. Thus the situation is similar to having independent observations by two scientists or laboratories using the same measurement procedure. If the results obtained by the two survey organizations are similar (no variation beyond sampling), then there is evidence of the reliability of the compared items in particular and the survey procedures in general. If there is unexplained variation, then either the particular measures or the survey procedures are suspect.

It is not, of course, an all-or-nothing situation. Houses (survey organizations) may disagree significantly, but much of the variation may be explained by discernible and controllable factors. If so, the problem is a failure to control for all relevant conditions. Also, if the variations are statistically significant and of undetermined origin but are nonetheless small in magnitude, then it can be said that the item and the survey procedures were reliable at a given level of precision.

In tandem designs, however, the two houses typically know that the results are subject to comparison with each other so one expects a special effort may be made to execute the survey carefully, paying special attention to procedures and perhaps using the best available personnel. Hence, it might be expected (assuming care to be an important variable in determining the quality of survey results) that tandem surveys would tend to be more reliable than most single-house surveys. So, tandem surveys probably provide an upper bound estimate of reliability.

In addition to examining examples of such tandem surveys, we will also review nonexperimental survey comparisons. Such comparisons are possible when two or more houses ask identical questions at approximately the same time without prior consultation, which happens occasionally. These serendipitous coincidences have the advantage of permitting comparisons free of the best-foot-forward effect discussed in the previous paragraph, but they are not exact replications since variations occur in timing, question order, and various other factors. Still it is of interest to see how closely survey measurements are reproduced under less than optimum conditions.

TANDEM SURVEYS

We are aware of five instances in which two survey organizations cooperated in the fielding of a survey: the Stouffer survey in 1954 conducted by

the American Institute of Public Opinion (AIPO, the Gallup organization) and NORC; the *Academic Mind* survey in 1955 conducted by Roper and NORC (in cooperation with the Bureau of Applied Social Research, Columbia University); the Steiner television survey in 1960 conducted by Roper and NORC; the privacy and confidentiality survey in 1976 conducted by the Survey Research Center (SRC) and the Bureau of the Census; and the National Health Care Expenditure survey in 1977 conducted by Research Triangle Institute (RTI) and NORC. In the following sections, we briefly review the comparability of the measurements made in these tandem surveys.

Stouffer Survey. The Stouffer survey on civil liberties (Stouffer, 1963) consisted of two distinct samples, a national cross-section and a sample of "community leaders." For the first, AIPO and NORC each independently drew full probability national samples of the adult population. They administered identically worded and ordered questions, used identical interviewer instructions, and coded open-ended questions according to standardized instructions. The community leadership survey was conducted in a similar manner with each house interviewing leaders in communities in their samples.

Stouffer reported no general comparison between the results of the two houses, but he did break down some of his tabulations by house. In twelve dichotomized comparisons between the national cross-sections, the absolute percentage difference averaged 1.9 percent.[8] A third of the total difference came from a single item that involved ambiguous codes (Stouffer, 1963:158). Stouffer's occasional references to house effects indicate that there was little or no difference in the conclusions by house (see Stouffer, 1963:28, 84, 111, 103, 113, 158).

A more complete comparison of house differences was carried out by Stember (1955). He looked primarily for three suspected house effects—differences in the level of "don't know" (DK) responses, differences in the volume of open-ended responses, and differences within the black subpopulation. Stember expected to find fewer DK responses on the NORC survey because NORC's general training instructs interviewers to probe initial DK responses to try to convert them to substantive responses while AIPO's general training puts more emphasis on preserving DK responses. (It was thought that past differences in the general training and experience of the two sets of interviewers would override the standardized instructions that interviewers were given for this particular survey.) Stember found no statistically significant differences on open-ended questions, but on closed-ended questions he found small yet consistent differences in the expected direction (Stember, 1955:31, 34). In nine of ten comparisons, AIPO had more DKs than NORC and five of these were statistically significant (Stember, 34). This difference was confirmed by recent analyses of the tolerance

scales; Lockwood (n.d.) found that almost all interhouse variance was attributable to differences in the DK categories.

Stember also expected to find more detailed and voluminous responses to open-ended questions because NORC's training emphasized complete verbatim answers while AIPO's training permitted capsulized answers and because NORC's questionnaire allowed five and a half times as much blank space to record answers. Contrary to his expectations, Stember found no notable difference in the quantity of coded responses. However, when he recoded a small subsample of the questionnaires using a more detailed coding scheme, he found that the NORC interviewers did, in fact, produce more responses. He concluded that the simple coding scheme used in the original analysis was too gross to capture the effects of these house differences.

Stember next compared differences in the distribution of open-ended responses and found that in eight of thirteen comparisons there were statistically significant differences. Differences were greatest on items dealing with vague, remote, or abstract topics.[9] Finally, Stember expected differences among black respondents since NORC tried to match the race of respondents and interviewers while AIPO did not. In eight of twelve comparisons Stember found that Southern blacks gave significantly more "frank" responses (that is, less moderate, more concerned about racial matters) to NORC's black interviewers.

In summary, Stember found some evidence for each of his hypothesized house effects. On DK responses, the differences were limited to closed questions and small in magnitude. On open-ended questions, the differences in quantity of response were imperceptible in the data as originally coded, but they were detected through recoding and reanalysis. On race, differences in organizational policies regarding the racial matching of interviewers and respondents produced some significant differences in the results obtained on race-related questions. Most of these effects were small, but the number of statistically significant associations was greater than would be expected to occur from sampling variation.

The Academic Mind Survey. Lazarsfeld and Thielens (1958) report the results of a survey of college teachers in the social sciences; this survey was administered jointly by NORC and Roper. Details of this collaboration are sketchy. Apparently, identical instruments and procedures were used but the two samples were not fully comparable since the availability of interviewers for selected colleges (rather than randomization) was used to assign interviews to a particular house.

No comparisons are reported by house although considerable attention is devoted to differences in the typical NORC and Roper interviewers (Lazarsfeld and Thielens, 1958:343–370). The only reference to house differences states (p. 344):

> . . . preliminary tables indicate that overall differences between the agencies are slight and do not operate in any consistent fashion, although, as we would expect, the N.O.R.C. interviewers obtained somewhat larger volumes of free-answer commentary. But this finding needs to be qualified by noting that, as things turned out, the agencies did not have entirely comparable assignments. . . . Roper interviewers . . . were given more assignments to colleges outside metropolitan areas . . . while N.O.R.C. interviewers got comparably more metropolitan institutions, where the respondents were more likely in any event to have more to say.

Steiner Television Survey. The Steiner television survey (Steiner, 1963) inquired about public use of and attitudes towards television. NORC and Roper drew probability samples (with quotas at the block level) of the national adult population, and administered the survey following their usual procedures. Each house administered eight distinct versions of the questionnaire for a total of sixteen distinct versions. Since every version was different, there are no completely comparable questionnaires across houses. Large sections of the questionnaires did have identical wordings and orders, however, so many cross-house comparisons are possible. Interviewers from both houses identified the survey as being jointly conducted by the Bureau of Applied Social Research, Columbia University; the National Opinion Research Center, University of Chicago; and Elmo Roper and Associates.

Steiner gives complete house breakdowns for all questions in an appendix. In the body he uses only combined totals since he finds (Steiner, 1963:7):

> Comparisons revealed small and mostly mechanical differences in the two sets of results. The substantive findings were almost always the same, so the findings throughout the book are presented for the total sample. . . .

An analysis of house differences reported in the appendix (Steiner, 1963: 260–320 showed an average absolute difference of 2.96 percentage points.[10] Given the 1,200-case size of the two samples, the differences generally are somewhat larger than would be expected due to sampling variation.[11] Once again, however, a clear house effect was spotted on the DKs. On 52 comparisons between houses, NORC had fewer DKs on 42 items, Roper fewer on 4 items, and there were no differences on 5 items. On average, Roper had 2.25 percentage points more DKs than NORC.

Privacy and Confidentiality Survey. The Privacy and Confidentiality Survey inquired about public attitudes and knowledge concerning survey practices. The sample was drawn by SRC and the cases randomly assigned

to SRC and Census Bureau interviewers. All coding was done by SRC. Interviewers identified their organizational affiliation to all respondents.

A panel of the Committee on National Statistics (National Research Council, 1979) reported the results of the survey. Its examination of house differences found (p. 22):

> In general, the results were very similar. Some differences appeared where they might be expected, in responses having to do specifically with identification of the collecting organization or in some cases with perceptions of it.

The panel did find some notable differences in both case and item nonresponse. The Census Bureau's refusal rate (6 percent) was half that of SRC (13 percent). It is presumed that this resulted from a greater perceived compulsion or actual willingness of respondents to cooperate with a governmental survey than with a nongovernmental survey. SRC, on the other hand, got lower item nonresponse than the Census Bureau.[12]

The panel reports some house differences but includes (mainly) instances in which noteworthy differences occurred (NRC, 1979, p. 22). Because of this, the published comparisons give an exaggerated view, that is, the average percentage difference of 6.3 for published comparisons is probably high.[13] In addition, most of the larger differences came from questions that either explicitly referred to the Census Bureau, asked about types of survey agencies, or involved interviewer rankings of respondents. Questions that did not directly or indirectly refer to the collection agency produced much smaller differences.

The 1977 National Medical Care Expenditure Survey. The National Medical Care Expenditure Survey was a large, full probability survey on medical conditions, treatments, and expenses. RTI and NORC each drew independent national household samples. Instruments, specifications, training, field procedures, quality control, and supervisory training were all standardized.

Analysis of these data has not been completed, but numerous comparisons have been made (Meyers and Oliver, 1978; Cohen and Horvitz, 1979; Cohen, 1980). In a detailed analysis of differences on age, sex, race, region, education, city size, marital status, and subjective health status (including several multivariate comparisons), Cohen and Horvitz (1979) found few statistically significant differences. In 53 comparisons involving subjective health, only 3 significant differences ($p < .05$) were found. The largest was in the DK category, which was twice as high for NORC (6.5 percent) as for RTI (3.0 percent). On four medical care measures (length of time waiting to see physician, proportion of dissatisfied visits, average number of visits, and proportion ever on Medicaid) that were tabulated by race, residence, and

family income and compared across houses, Cohen (1980:17) noted: "our findings supported the notion of compatibility as we consistently observed nonsignificant survey organization effects when testing for differences in domain estimates of the alternative health indices."

A third, more limited analysis (Meyers and Oliver, 1978) compared reasons for refusals. NORC and RTI had almost identical response rates (93 percent), but differences were found on the reports filled out by interviewers as to the reason for refusals. NORC had significantly higher levels of unspecified refusals than did RTI. This difference resulted partly from the fact that NORC's data were coded by the National Center for Health Services Research while RTI's were coded by the National Center for Health Statistics. The former agency tried to maximize coded reasons; part of this coding difference shows up in NORC's lower share of "other" codes. The authors concluded, however, that even if stricter coding rules had been applied, house differences would have remained.[14]

NONEXPERIMENTAL COMPARISONS

Early Work. In 1945, Hadley Cantril published a research note in *Public Opinion Quarterly* entitled, "Do different polls get the same result?" This note reviewed 20 instances in which the same (subjective) survey question was asked by two different survey houses during the period from 1940 to 1944. Unlike the tandem surveys just reviewed, these instances involved serendipitous coincidences. Cantril's review was apparently stimulated by the suggestion of S.S. Wilks[15] (Cantril, 1945:62), and it shows great concern with evaluating (or perhaps vindicating) the practice of quota sampling—a practice that is not held in high regard by mathematical statisticians and others because of its use of nonrandom selection criteria. Thus, Cantril (1945:45) hoped for "the accumulation of data for an exchange of information and, when later comparisons are made of different methods of sampling, a study of the adequacies and inadequacies of different [sampling] methods for different problems."

The age of these comparisons and their focus on sampling methods (which have changed over the last 35 years as more organizations have adopted probability sampling)[16] diminish their practical importance for the evaluation of the contemporary state of survey practice. Indeed, two of the four survey organizations included in Cantril's study no longer exist (Office of Public Opinion Research [OPOR] at Princeton University and the *Fortune* Poll). We thus review this evidence separately from that derived from more recent studies.

Cantril presented his 20 nonexperimental comparisons and he quickly rested his case after displaying the average absolute difference (3.24 percentage points) obtained for each response category of the 20 questions for which interhouse comparisons were possible. He did, however, note that

the average discrepancy was smaller when the two measurements were made during the same 10-day period (average: 3.05 percentage points) than when the period was longer (for 11–70 days: 3.41 percentage points). And Cantril briefly noted (1945:62) that "the more concrete political questions result in more agreement than the non-political" (average discrepancies of 3.15 and 3.43 percentage points). Cantril used "political" to denote candidate preference questions, that is, questions comparing preferences for the Democrats and Roosevelt versus the Republicans and Wilkie. The non-political questions asked about a variety of topics, such as: Who do you expect to win the war? Has the government gone far enough in asking people to make sacrifices for the war? Should every man be required to serve in the armed forces?

The wording and results for the two questions from Cantril's study that produced the most discrepant results are presented in Table 5.7. (The full text and results for all 20 questions are presented in Appendix F.) It will be seen from Table 5.7 that the two most discrepant measurements involved quite general judgments about national policies.

TABLE 5.7
The Most Discrepant Questions from Cantril's (1945) Compilation

"Which of these two things do you think the United States should try to do when the war is over: Stay out of world affairs as much as we can, or take an active part in world affairs?"

	AIPO (April 25, 1944)	OPOR (April 5, 1944)
Stay out	18%	28%
Take active part	74	66
No opinion	8	6
	100	100

Index of dissimilarity = 3.9

"In general, do you think the government has gone too far or not far enough in asking people to make sacrifices for the war?"

	AIPO (April 25, 1944)	OPOR (April 20, 1944)
Too far	12%	8%
About right	42	37
Not far enough	37	49
No opinion	9	6
	100	100

Index of dissimilarity = 4.4

NOTE: Exact sample sizes are not given, but Cantril indicates that sample sizes were approximately 3,500 (AIPO) and 1,200 (OPOR). See Appendix F for the full results.

This finding is of particular interest since it has been suggested elsewhere (see Section 4.2) that factual questions that are vague, general, or ambiguous or that ask about distant events produce greater measurement variation (that is, greater respondent, interviewer, and coder variance). Such wide variation also occurs when respondents are questioned about matters about which they know little (for example, Census Bureau measurements of the cost of utilities when asked of persons who do not pay their own utility bills or measurements of the date of construction of one's residence when asked of apartment dwellers). There is also some indication that interviewer and coder variance increases as a function of the degree of interpretation required in selecting a response category. Thus, sex, age, and marital status evidence less variation than questions about income or the language spoken in one's childhood home.

Recent Work. In recent reviews of discrepancies in the measurement of subjective phenomena, similar conclusions have been suggested (Turner and Krauss, 1978:468; Turner, 1980, 1981a). (This evidence is reviewed in greater detail in Turner, Chapter 7, Volume 2, and potential mediating processes are described in Chapters 7 through 9 in this volume.) The characteristics of survey questions that have been hypothesized to lead to increased levels of measurement variation[17] are similar to those discussed previously. They include questions:

1. That are vague (or overly general) in their meaning, for example:

   In general, how happy would you say you are these days? Would you say that you are very happy, pretty happy or not too happy?

2. That ask for judgments on topics to which respondents have given little thought, for example:

   Science and technology can be directed toward solving problems in many different areas. In which of the areas listed on this card would you *most* like to have your taxes spent for science and technology? [A list of 13 areas was presented, including weather control, reducing and controlling pollution, finding better birth control methods, space exploration, discovering new basic knowledge, and so forth.]

3. That require a (seemingly) arbitrary choice of a response category, for example:

> As far as *people* running these institutions are concerned, would you say you have a great deal of confidence, only some confidence or hardly any confidence at all in them?

4. Whose answers have no explicit behavioral implications in the everyday lives of the respondents, for example:

> [Would you favor or oppose] allowing businesses to take larger tax deductions on equipment by valuing the equipment at what it would actually cost today—rather than at what it actually cost to buy the equipment several years ago . . . ?

In this typology a question—such as the first about happiness—would be expected to evidence relatively higher levels of measurement variation than the following two-part question (asked of women of childbearing age):

> "Looking ahead, do you expect to have any (more) children?"
> "How many more?"

These expectations derive from the fact that the first question asks about a vague concept (happiness) that people can interpret in various ways; it also requires a seemingly arbitrary choice between the categories of very and pretty happy; and the answer (for example, saying you are "very" versus "pretty" happy) would seem to have few immediate implications for the respondent's behavior in real life.

In contrast, the second set of questions asks about a topic that has a relatively unambiguous meaning. The questions require a choice between response categories that have a clear meaning (0, 1, 2 . . . children), and they ask about a topic which respondents (married women of childbearing age) probably have thought about. Finally, the responses have some direct implications in the everyday lives of the respondents (for example, contraceptive practices).

There are several instances of independent measurements made at roughly the same time using these two questions. The happiness question has been shown to produce rather discrepant estimates in a number of instances—even when the measurements were made using *identical* questions asked by the *same* survey organization at almost the same time (see Figure 1.1 in Chapter 1). In contrast, the questions on childbearing expectations have shown insubstantial discrepancies in four instances in which data were collected by different survey organizations using different wordings of these questions. (These data are reported fully in Turner, Chapter 7, Volume 2.)

Our review of both the past and the more recent literature (Cantril, 1945;

Smith, 1978, 1979, 1981; Turner 1981a) yielded a total of 126 instances in which the same questions were asked by different survey measurement programs at about the same time.[18] These instances are not by any means representative of all possible house comparisons and overrepresent both certain houses (NORC appears in the vast majority of comparisons) and certain questions (confidence questions account for 50 comparisons). Therefore, the results are not representative of survey research in general, but the comparisons do cover a wide variety of questions and houses (NORC, SRC, AIPO, Roper, Harris, and so forth).

A summary tabulation of our findings (see Appendixes F and G for details) is presented in Table 5.8. This tabulation evidences a dispersion in the results of different surveys which would be difficult to attribute to the effects of sampling. Approximately 45 of these differences were outside of the (approximate) intervals typically allowed for sampling error (for example, for simple random samples of size 1,500, differences of 4 percentage points or more would be expected to occur less frequently than 1 in 20 times). (We also note that there is an apparent—and disturbing—trend in the comparisons shown in Table 5.8: the more recent survey measurements appear more discrepant than those compiled by Cantril over 30 years ago.)

The nonsampling variation that is evident in this tabulation undoubtedly comes from many sources: questionnaire context, interviewer effects, item specifications, training and staff differences, and so forth. Two important sources were house variations in item nonresponse and change over time. As is the case for the tandem surveys described previously, there is substantial evidence of house differences in the handling of DKs. For one comparison between SRC and NORC, SRC had more DKs on all 17 items (Smith, 1980b). Similarly, for 11 items on national spending, the average difference was 4.6 percent with DKs included but only 3.3 percent with DKs excluded from analysis (Smith, 1978). We also note that while in tandem surveys there is no problem of change over time contributing to the variation, in nonexperimental comparisons it is always possible that true change in the population contributed to the observed differences.

Because of the great range of discrepancies that have been observed in survey measurements of subjective phenomena, it has sometimes been suggested that measurements of some phenomena may be more vulnerable to artifactual variation (and hence intersurvey discrepancies). For exploratory purposes, Table 5.9 presents the topics and response categories for a selection of our nonexperimental comparisons along with those for some experimental comparisons in which only the context of the measurement was altered. (A more comprehensive listing of nonexperimental comparisons is presented in Appendix G; and see MacKuen and Turner, Chapter 16, Volume 2, for a review of 513 survey measurements of presidential popularity made between 1963 and 1980.)

TABLE 5.8
Summary Tabulation of Discrepancies

Discrepancy	Date of Survey: 1944 and Before	Date of Survey: 1972 and Later	Total
0.0 – 0.9	6	17	23
1.0 – 1.9	6	14	20
2.0 – 2.9	2	23	25
3.0 – 3.9	5	8	13
4.0 – 4.9	1	9	10
5.0 – 5.9	—	8	8
6.0 – 6.9	—	5	5
7.0 – 7.9	—	8	8
8.0 – 8.9	—	3	3
9.0 – 9.9	—	4	4
10.0 – 10.9	—	2	2
11.0 – 11.9	—	—	—
12.0 – 12.9	—	2	2
13.0 – 13.9	—	2	2
14.0 – 14.9	—	—	—
15.0 – 15.9	—	—	—
16.0 – 16.9	—	1	1
Totals	20	106	126

NOTE: Tabulation summarizes comparisons presented in Appendixes F and G. For dichotomous (or dichotomized) variables, discrepancies are percentage-point differences (e.g., if survey *X* found that 45 percent said "yes" to a question and survey *Y* found that 47 percent said "yes" to the same question, the discrepancy would be 2 percentage points). For polytomous variables, discrepancies are index of dissimilarity values: the index indicates the percentage of respondents who would have to switch response categories in order to bring the two surveys into complete agreement. (For dichotomous variables, the two measures of discrepancy will give the same numerical result when the sample sizes in the two surveys are the same.)

We note that Table 5.9 contains relatively more significant discrepancies for the measurements at the top of the table (Nos. 1–15), which involved rather amorphous phenomena such as confidence in the people running institutions, evaluations of national spending, and contemporary driving standards, and which required a choice among relatively imprecise response categories. In contrast, questions (Nos. 16–21) about the legalization of marijuana, gun control, the death penalty, political party affiliation, and

fertility show relatively few discrepancies. The latter questions involve somewhat less amorphous phenomena and response categories, such as support or nonsupport of legislation, preference for a political party, or actual number of children expected.

The questions shown in Table 5.9 may also differ in their relevance to the lives of respondents. Many of the questions at the bottom of the table solicit opinions about topics that have been the subject of considerable public discussion (for example, capital punishment and the legalization of marijuana), or that should find concrete behavioral expression in the everyday lives of many respondents (for example, voting, party registration, contraceptive practice). Among this group of questions, only two significant discrepancies were observed in 19 comparisons. In contrast, many questions in the first group are about topics that are not usually subject to considerable public discussion (at least as phrased in these questions), for example: Do you have confidence in the people running major companies? Are we spending too much on science? More than one-half of the comparisons (42 of 82) in this group of questions produced significant discrepancies.

While such a typology may provide a reasonable place to begin future research on vulnerable measurements, it should be noted that these question characteristics are very general and subjective (and might themselves be criticized as amorphous or vague). In addition, when applied to many common measurements, they are redundant. For example, questions about "amorphous" phenomena frequently, and perhaps inevitably, require arbitrary choices between response categories. Thus it would be difficult to specify which question characteristics (or combinations of characteristics) might be crucial in inducing vulnerability. Finally, we note that results—such as those for the Communist reporter item (see Table 5.5)—suggest that other factors not included in this typology also operate.

There is, however, some other evidence that supports such a typology of vulnerable measurements. Studies of Census Bureau estimates of crime victimization (Cowan, Murphy, and Wiener, 1978; Gibson et al., 1978) have revealed that substantial variations have been induced by differences in survey contexts. Gibson et al. (1978) report that this measurement artifact produced a relative increase of 12 percent in the reported rate of property crime and 21 percent in the reported rate of personal crime. Examination of the National Crime Survey questionnaire suggests that these results may fit within the foregoing typology. The victimization questionnaire includes, for example, the following screening question:

> Q. 42 Did anyone try to attack you in some other way? [other than attacks already mentioned]

The concept of *attempted* attack in this question is somewhat amorphous. (When, for example, does an accidental bump become a hostile blow?) Co-

wan, Murphy, and Wiener (1978:282) find that reported rates of "attempted assaults (wihout weapon)" showed significant variation across measurement contexts (Z of difference = 2.76) and that "attempted assault (with weapon)" showed a difference of borderline significance (Z = 1.83); however, Cowan, Murphy, and Wiener find that the reported rates for assaults "with injury" showed no significant variations across measurement contexts (Z = 0.17 and 0.27). This pattern of results suggests that the vulnerability of these reports to context effects decreased as the events in question became more unmistakably hostile and, we assume, more salient to the victims.[19]

CONCLUSION

Overall we find a high degree of comparability between the results obtained when two survey organizations conduct the same survey in tandem. Most differences are small and typically within sampling variation. It is also clear, however, that there are differences that exceed sampling variation.[20] The method of handling nonresponse appears to produce both the most common and usually the largest differences in tandem surveys. When DKs and other nonresponses are eliminated from the analyses, differences usually decrease substantially and frequently become statistically insignificant. Other differences appear to be related to the coding of open-ended questions, but even with standardized coding schemes, house differences can still occur. (This kind of difference has also been found within organizations when coding identical questions from different surveys; see Schuman [1974].) Finally, some differences that are clearly tied to the survey organization do occur in tandem surveys.

Our overall conclusion regarding tandem surveys is subject to some qualifications in that the analysis of the replications in no case was as intensive and sophisticated as some of the work on split-ballot replications within a single survey (see Section 5.1). Hence, there may be unsuspected failures to replicate, particularly in regard to multivariate aspects of these data. We cannot be sure, for example, that the higher DK rate for AIPO than for NORC is wholly innocuous. Perhaps the NORC habit of probing leads some respondents who do not really have opinions to answer in an artifactual way or, perhaps, AIPO's nondirective approach misses the less vocal respondents. But such effects, even if they exist, could be small and difficult to detect in multivariate analysis of samples as small as the ones commonly studied.

With this caveat, we conclude that the tandem surveys demonstrate that quite satisfactory replication of survey results *can* be obtained when sufficient care and attention are given to all the details of a survey. When independent surveys are made in the same population at the same time with comparable instrumentation, the results are the same within the range permitted by calculable sampling error. The importance of this conclusion

TABLE 5.9
Examples of Discrepancies in Survey Measurements of Subjective Phenomena

Topic	Response Categories	Type of Comparison[a]	Number of Comparisons	Significant Discrepancies[b]
1. Confidence in people running national institutions	Great deal of confidence, only some confidence, hardly any confidence	Different houses[c]	27	18
		Context experiment[i]	13	3
2. Evaluation of amount of federal income tax	Too high, about right, too low	Context experiment[c]	1	1
3. General happiness	Very happy, pretty happy, not too happy	Different houses[d,i]	2	1
		Same house/different surveys[i,j]	2	1
4. Evaluation of prestige of occupations	Excellent, good, average, below average, poor	Context experiment[i]	10	7
5. Evaluation of contemporary driving standards	Lower than they used to be, higher than they used to be, about the same	Context experiment[f]	2	1
6. Evaluation of traffic noise	Becoming noisier, less noisy, or about the same	Context experiment[f]	2	1
7. Evaluation of need for nighttime truck deliveries to retail stores	Yes, no[g]	Context experiment[f]	2	1

8. Evaluation of spending on national problems	Too much, about right, too little	Different houses[h]	10	5
9. Misanthropy I: people try to take advantage of you	Try to take advantage if they got a chance, try to be fair	Different houses[h]	2	0
10. Misanthropy II: most of time people try to be helpful	Try to be helpful, mostly just looking out for themselves	Different houses[h]	2	1
11. Misanthropy III: most people can be trusted	Can be trusted, can't be too careful in dealing with people	Different houses[h]	2	1
12. Courts' treatment of criminals	Deal too harshly or not harshly enough with criminals	Different houses[h]	1	0
13. Men or women better suited for politics	Men better suited, equally suited, women better suited	Different houses[h]	1	0
14. Vote for woman for president	Yes, no	Different houses[h]	1	1

TABLE 5.9 *(continued)*

Topic	Response Categories	Type of Comparison[a]	Number of Comparisons	Significant Discrepancies[b]
15. Children safe bicycling in local area	Yes, no	Context experiment[f]	2	0
16. Law requiring police permit to purchase gun	Favor, oppose	Different houses[h]	3	0
17. Support for legalization of marijuana	Yes, no	Different houses[h]	1	0
18. Political party with which one identifies self	Republican, Democrat, Independent, other	Different houses[h]	3	1
19. Favor death penalty for murder	Yes, no	Different houses[h]	3	0
20. Ideal number of children for family	0,1,2,3 . . .	Different houses[h]	1	0
21. Expected number of children	0,1,2,3 . . .	Different houses[i,e]	8	1

NOTES: These examples are reproduced from Turner (1981a:Table 9); they were derived from comparisons cited in Kalton, Collins, and Brook, 1978; Smith, 1978; Turner and Krauss, 1978; and Turner, 1981a. Only survey measurements of subjective phenomena are included; reports of behaviors (e.g., voting) and demographic estimates included in the foregoing sources were excluded. Comparisons involving variant question wordings have been excluded—except where noted otherwise.

Appendixes F and G in this volume present a more comprehensive list of (nonexperimental) intersurvey comparisons. MacKuen and Turner (Chapter 16, Volume 2) present a seperate assessment of the reliability and validity of 513 survey measurements of presidential popularity made during 1963–80 by the Gallup Organization, Louis Harris and Associates, CBS/*New York Times*, NBC/AP, and the Roper Organization.

[a]All comparisons involve either (1) independent estimates made by different survey research organizations (different houses); (2) experiments involving the manipulation of the context in which survey questions were presented (context experiment); or (3) estimates derived from seperate surveys conducted by the same survey research organization (same house/different surveys).

[b]Discrepancies are absolute differences between estimated proportions of population giving a particular response in two surveys (or two experimental conditions). Discrepancies have been termed "significant" if the observed discrepancy exceeded twice the sampling error for the difference between the proportions. Published statistics and/or standard errors from the source publications were used, if available. When data were derived from clustered samples and no published statistics were available, estimates were based on the formula for simple random samples, but the sample size was deflated ($N' = 0.66N$) to allow for sample design inefficiencies.

[c]Turner and Krauss, 1978.

[d]Comparison involves a modest difference in question wording.

[e]Note that all comparisons involve a subset of GSS sample that numbered less than 250 each year.

[f]Kalton, Collins, and Brook, 1978.

[g]Source does not quote explicit response categories; these are implied by text of question.

[h]Smith, 1978.

[i]Turner, 1981a, and Chapter 7, Volume 2.

[j]Comparisons were made by pooling data from 2 months of NORC Continuous National Survey closest to dates of General Social Surveys.

is great, since the sine qua non of science is that it produces findings that are reproducible under clearly defined conditions. Whatever it is that surveys measure, they *can* measure it consistently, and in that sense, surveys are reliable.

The fact that replication is theoretically possible does not, of course, guarantee that it will be achieved. Obviously, it is possible to do substandard work, and no doubt a great deal of such work gets done. But there is a standard, and work to that standard can be performed, given the resources and the mobilization of technical competence that are required. Therefore, when an ostensible failure of replication is encountered (as in some of the unplanned between-house comparisons we reviewed), it should not be regarded as a kind of inevitable hazard of the enterprise. On the contrary, causes for such failures can reasonably be sought in the departure of the survey designs from comparability (whether of sampling frame, questionnaire design, field work, or whatever).

What we do not know, and what will take a continuing program of research to determine, is how far one can depart from strict replication of all aspects of survey design and still achieve satisfactory agreement of results. Indeed, since it is likely that the robustness of survey measurements is subject-matter specific and interacts with fluctuations in the social milieu, the job of estimating the robustness of our results is a continuing one that must remain a permanent aspect of survey work, just like the estimation of sampling errors (see Recommendation 9, Chapter 10).

5.3 Coordinated Measurement and the Control of Error

As detailed in the preceding section, our review of tandem surveys indicates that survey measurements of subjective phenomena can and do produce reliable results. That is to say, given adequate care in carrying out surveys, it is possible to reproduce the measurement process and thereby obtain satisfactory agreement of results. However, there are many noteworthy instances of measurements that were widely assumed to be equivalent but that produced highly discrepant results. Such results suggest the need for more routine consideration of nonsampling sources of variation in survey measurements and also of the types of institutional mechanisms that might improve the comparability of survey measurements (see Recommendations 16 and 17, Chapter 10). As a venture in this direction, the panel undertook some coordinated measurement studies both to provide insight

into the problems of concern to us and to explore the value of such strategies in general. We briefly report this work in the hope that it will be helpful to others who may face similar problems with their survey measurements.

As noted in Chapter 1, our study was prompted, in part, by the discovery of a series of discrepancies in some commonly reported and widely used survey measurements. During the interim between the first ad hoc planning meeting for this project and the drafting of this volume, several experiments were conducted to explore the causes of the variability in those measurements.

INITIAL EXPERIMENTS

Four pilot experiments were carried out using two identically worded questions, one on general happiness and a more specific one on marital happiness:

> [*General Happiness*] Taken all together, how would you say things are these days—would you say that you are very happy, pretty happy, or not too happy?
>
> [*Marital Happiness*] Taking things all together, how would you describe your marriage? Would you say that your marriage is very happy, pretty happy, or not too happy?

Three initial experiments were conducted by the George Fine Organization for the *Washington Post* in May 1979; by the Survey Research Center of Michigan in August 1979; and by the National Opinion Research Center in February–March 1980.

In each of these experiments the order of presentation of the survey questions was manipulated. In the *controlled context* (approximately one-half of the survey interviews),[21] the general happiness item followed the question on marital happiness; in the *uncontrolled context* (the remaining half of the interviews), the item followed whatever else was in that section of the survey questionnaire. Since the three surveys were otherwise completely different, the question immediately preceding the general happiness item varied in the uncontrolled context. So, for example, in the *Washington Post* survey, the general happiness item followed a question on income in the uncontrolled context; in the SRC survey, it followed a series of items on the gas shortage and a question asking marital status; and in the NORC survey it followed a series of five questions asking "how much satisfaction" respondents received from their: city or place of residence, hobbies, family life, friendships, health and physical condition. A parallel variation in measurement context occured for the marital happiness item: In the controlled context, it followed the question on general happiness; and in

the remainder of the interviews, it preceded the general happiness item and, thus, followed whatever else was in that section of the survey questionnaire (the same items as noted before).

Figure 5.1 displays the proportion of respondents saying they were "very happy" in response to the general and marital happiness questions in each survey.[22] While the results of these experiments are perplexing in some regards, the data do consistently support the hypothesis that general questions are more vulnerable to context effects than specific ones. Thus, the more specific question about marital happiness yields equivalent estimates in all three surveys and in both measurement contexts. In contrast, the general happiness question yields consistent estimates only when the measurement context is controlled. When the measurement context is uncontrolled, the proportion of respondents who say they are "very happy" varies from .335 to .527. Thus, when we fit alternative log-linear models to the cross-tabulation of happiness response $\{H\}$ by survey $\{S\}$ by measurement context $\{C\}$, we find that a three way interaction term $\{HSC\}$ is required to fit the data for the general happiness question. In contrast, a model [No. 1: $\{H\}$, $\{CS\}$] which posits that the measurements are independent of both the measurement context and the survey organization, provides an adequate fit to the *marital* happiness data from these three experiments ($p = .37$). Further details of these analyses are presented in Table 5.10.

We do note that an interesting alternative parameterization is possible for these data. The general approach is that of Duncan (1979c). The linear parameterization and estimates of the log parameters under comparable models (4* and 5*) for the general happiness measurements are presented in Table 5.10 (b). In model 5*, we constrain all associations involving happiness, that is, {HS}, {HC}, and {HSC}, to be linear (in the log frequency) with respect to the three categories of the happiness response. In the parameterization shown in Table 5.10 (b), parameter a pertains to the (linearized) {CH} association, parameters b and c to the {SH} association, and parameter d to the 3-way {HSC} interaction. However, d is used in such a way that the NORC/*Post* contrast with respect to the {CH} association is just the inverse of the SRC/*Post* contrast with respect to that association.[23]

Comparing model 4* with model 4 shows very little deterioration of fit when the {HS} and {HC} associations are linearized, despite the gain of 3 d.f. Since neither model fits, however, neither should be given serious attention. Model 5* includes the linearized 3-way interaction, and the fit is excellent, even though 6 d.f. have not been used. This result implies that the effects of context and survey organization on measurements of happiness (as represented in the 3-way tabulation) are well modeled by assuming that the three categories of happiness form a linear scale.

While the results shown in Table 5.10 and Figure 5.1 fit the pattern hypothesized in the preceding section (that is, the general question pro-

TABLE 5.10

Test of Alternative Models for Behavior of Happiness Measurements

A. GENERAL HAPPINESS MEASUREMENTS

Model	Marginals Fit	d.f.	L^2	p
1. Stable measurements	{H} {CS}	10	28.0	.002
2. Context effect	{HC} {CS}	8	27.7	.001
3. Survey effect	{HS} {CS}	6	20.8	.002
4. Context and survey effects	{HS} {HC} {CS}	4	20.0	.001
5. Interaction effect	{HSC}	0	0.0	—

B. MARITAL HAPPINESS MEASUREMENTS

Model	Marginals Fit	d.f.	L^2	p
1. Stable measurements	{H} {CS}	10	10.8	.37
2. Context effect	{HC} {CS}	8	10.7	.21
3. Survey effect	{HS} {CS}	6	5.9	.44
4. Context and survey effects	{HS} {HC} {CS}	4	5.5	.24
5. Interaction effect	{HSC}	0	0.0	—

C. PARAMETERIZATION FOR GENERAL HAPPINESS EXPERIMENT

		Response		
Survey	Context	Very Happy	Pretty Happy	Not Too Happy
NORC-GSS, 1980	MG	1	bd	b^2d^2
	GM	1	ab	a^2b^2
Post, 1979	MG	1	1	1
	GM	1	a	a^2
SRC, 1979	MG	1	c	c^2
	GM	1	acd	$a^2c^2d^2$

Model	*Constraints*	*d.f.*	L^2
4*	$d = 1$	7	20.837
5*	——	6	3.369

Estimates of *log* parameters (with standard errors) under model 5*

a – .00902 (.06993)

b – .36639 (.09869)

c – .04628 (.10970)

d – .41935 (.10060)

NOTE: Models were fit using procedures developed by Goodman (1971) and Duncan (1979c). L^2 values are likelihood ratio chi-square statistics. Variables included in this analysis are:

H: Response to happiness question (three categories: very happy; pretty happy; not too happy). Respondents who did not answer this question (circa 1 percent or less) were excluded from sample.

S: Survey (three categories: NORC; SRC; *Washington Post*)

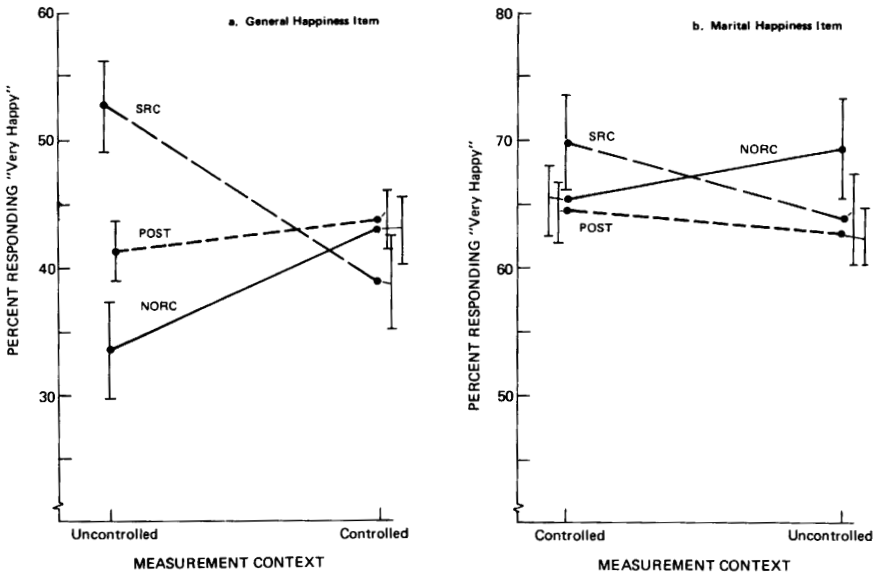
C: Measurement Context (two categories: (1) controlled context; (2) uncontrolled context).

duced more labile measurements than the specific question), it is important to note that the direction and magnitude of the artifacts found in this experiment would have been difficult to predict in advance. Thus, it is hard to intuit why SRC's measurements in the uncontrolled context are so much higher than those of NORC and the *Post*. Nonetheless, since the different organizations' measurements agree in the controlled context, we can effectively rule out organizational differences in sampling, processing, and so on, as an explanation for the disagreements we have observed.

These results indicate that the discrepancies in our survey measurements often have identifiable causes, and that there may be systematic differences between the types of questions that are more (and less) vulnerable to measurement artifacts. It thus appears that some improvements in survey practices are possible. In the present case we would suggest that special caution needs to be exercised with general questions on "happiness." Replications that do not control the questionnaire context in which such measurements

FIGURE 5.1

Percentage of respondents reporting they were "very happy" in reponse to questions on (a) general happiness, and (b) marital happiness. See text for discussion of measurement contexts. (Error bars demark approximately ± 1 standard error around estimates.)



SOURCE: From surveys conducted by the Survey Research Center (August, 1979), the National Opinion Research Center (February-March, 1980) and the *Washington Post* Poll (May, 1979). Samples restricted to married respondents with telephones in their residences.

are made appear to be vulnerable to contamination by substantial nonsampling artifacts.

RECONCILIATION OF MEASUREMENTS

While our initial experimental studies demonstrated the vulnerability (or lability) of measurements of general happiness, it remained to be demonstrated that the *specific* discrepancies (see Figure 1.1) that prompted our work could be explained. This demonstration was achieved in the fourth experimental study conducted for the panel in November 1980 by the Opinion Research Corporation. As part of the experimental design, general happiness measurements were made in contexts that replicated the low and high NORC-GSS measurements (see Figure 1.1).

In the negative (low) context, the general happiness question followed three questions inquiring about the respondent's financial situation, questions that apparently depressed respondents' estimates of their happiness in the 1973 NORC-GSS survey:

> We are interested in how people are getting along financially these days. So far as you and your family are concerned, would you say that you are pretty well satisfied with your present financial situation, more or less satisfied, or not satisfied at all?

> During the last few years, has your financial situation been getting better, getting worse, or has it stayed the same?

> Compared with American families in general, would you say your family income is—far below average, below average, average, above average, or far above average?

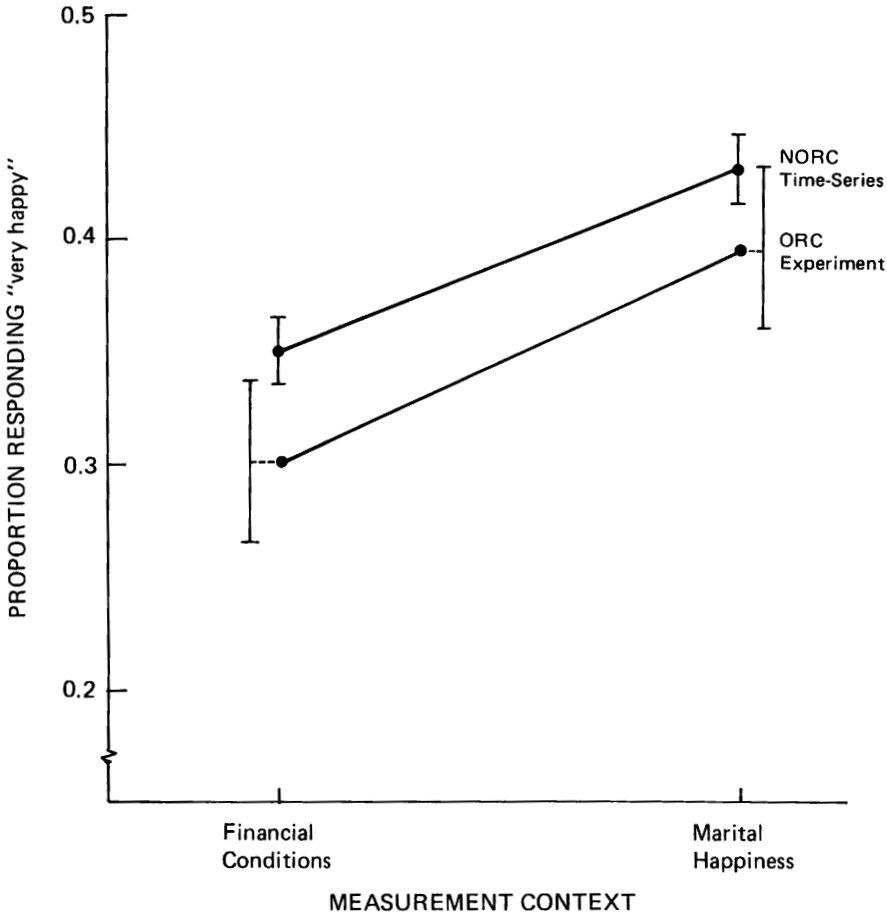
(We were led to incorporate the financial questions in the experiment by the discovery that the bivariate associations between responses to the financial and happiness questions varied significantly between 1972 and 1973 in the NORC series; see Turner, Chapter 7, Volume 2, for details.) In the positive (high) context, the general happiness question followed the marital happiness question (see previous discussion for wording). These two sequences replicate those used in the 1972 (low) and 1973 (high) NORC-GSS measurements. Our hypothesis was that this context manipulation in 1980 would produce a variation in the distribution of responses identical to that observed for the 1972 and 1973 NORC measurements, which would imply that the apparent change in happiness between 1972 and 1973 was artifactually produced. (Our hypothesis of no temporal change between 1972–73 followed from the facts that: (1) an independent series of NORC-CNS mea-

surements was stable during this period, and (2) the rise in the NORC-GSS series was largely attributable to a shift in the responses of married respondents.)

The relevant results are presented in Figure 5.2 and Table 5.11. It will be seen that (for married persons) 30 percent of respondents said they were

FIGURE 5.2

Proportion of respondents saying they were "very happy" in response to general happiness question in two measurement contexts. See text for discussion of measurement contexts. (Error bars demark approximately ± 1 standard error around estimates.)



SOURCE: From NORC General Social Surveys of 1972 and 1973 and ORC survey experiment of 1980. Samples restricted to married respondents.

"very happy" in the negative context and 39 percent in the positive context. Moreover, when we compared the experimental results in 1980 with NORC data for 1972 and 1973, we found that the only parameter strictly required to fit the data is that for measurement context. Our hypothesis that the measurement context, and not true change, explains the difference between the 1972 and 1973 NORC results is thus supported.

In this instance, a coordinated attack on measurement problems discovered in the review of past measurements has yielded both further insights into some of the characteristics of questions that seem to yield measurement anomalies and also an explanation of discrepancies found in the archival record of past measurements.

TABLE 5.11

Comparison of Results of ORC Experiment and 1972–73 NORC Measurements

		Response to General Happiness Question				
Survey	**Context**	**Very Happy**	**Pretty Happy**	**Not too Happy**	**Total[a]**	
ORC experiment, 1980	Marital	39%	51%	11%	100%	(358)
	Financial	30	52	18	100%	(393)
NORC time series, 1972–73	Marital	43	48	9	100%	(1,073)
	Financial	35	52	14	100%	(1,156)

NOTE: Includes only married respondents; "don't know" and "no" answers excluded.
[a]Numbers may not add up to 100 percent due to rounding.

Fit of Alternative Models

Model	**Marginals Constrained[a]**	**L^2**	**d.f.**	**p**
1. No effects	{CS} {H}	39.9	6	.0005
2. House-year effect only	{CS} {HS}	31.5	4	.0002
3. Context effect only	{CS} {HC}	8.6	4	.07
4. Context and house-year effects	{CS} {HC} {HS}	0.4	2	>.5
5. Interaction effects	{CSH}	0.0	0	—

NOTE: Models were fit using procedures developed by Goodman (1971). L^2 values are likelihood ratio chi-square statistics.
[a]Variables included in this analysis are:

H: Response to happiness question (three categories: very happy; pretty happy; not too happy). Respondents who did not answer this question were excluded from sample.

S: Survey (two categories: NORC–General Social Surveys: 1972, 1973; ORC survey experiment, 1980. Note that survey house is inextricably confounded with time, i.e., NORC = 1972–73 while ORC = 1980.

C: Measurement Context (two categories: (1) financial conditions context; (2) marital happiness context).

CONCLUSION

There is much that can be done to improve understanding of the reasons why survey data sometimes yield inconsistent estimates of subjective phenomena. The preceding examples are instructive of one fruitful line of attack. Coordinated measurement programs are common in other disciplines (see Chapter 1), and we recommend that similar programs be considered for survey measurement (see Chapter 10). In particular, it may be worthwhile for survey organizations themselves to consider cooperative ventures involving coordinated measurements, for example, by incorporating a number of baseline measurements in their surveys that replicate those in surveys conducted by other organizations. Such data might be useful in deciding when and how data from different sources could be validly integrated; it would aid in the detection of errant survey practices and problematic questions; and it could result in improvement in the quality of survey measurements. Coordination between survey houses would also facilitate systematic study of the various methodological problems that have been discussed in this and the preceding chapter.

The study and control of variation is a basic requirement of scientific measurement. And so, the calibration of measurement instruments—across time and across organizations—is a necessary part of the scientific use of survey measurements for the study of social phenomena and social change.

We conclude by noting that our suggestion of increased coordination is not novel, and we are not the first panel to make such a recommendation. Cantril in 1945 made a similar appeal, and the SSRC's report on the failure of the electoral polls in 1948 echoed this concern. While we have no desire to bemoan past inaction, we would hope that 30 years hence it will not require yet another panel such as ours to repeat this recommendation.

Notes

1. In this section (except where otherwise noted) we treat systematic effects, not random effects. And since most of the evidence comes from between-subjects experiments, we deal with distributions of responses rather than with individual responses. It is possible, of course, for a distribution to remain unchanged because systematic effects at the individual level balance out. We should also note that when a question is turned into an election referendum, the results, ipso facto, do take on absolute meaning (that is, passage or failure to pass), whatever the results that other wordings of the referendum question might have shown.

2. A split ballot is simply an experiment in which a random half of the questionnaires have a question or set of questions asked in one way and the other half in a different way. Divisions other than 50-50 can be used, as can more complex factorial designs.

We should note here that throughout this chapter we employ statistical tests which presume that the data were derived from simple random samples of the population. In the majority of cases this is not so, and the probability statements therefore must be considered approximate. We note, however, an explicit study of sample design effects which was conducted in the context of some of the experimental studies we review (Schuman and Presser 1981). This study found design effects that ranged from 0.70 to 1.40 (for the difference in proportions across questionnaire forms), with an about equal number falling on either side of 1.0.

3. There is some suggestive evidence that wording changes exert their greatest effect when they are fairly subtle and that more obvious attempts to "bias" answers are apt to be unsuccessful, presumably because they are recognized by respondents and discounted; see Schuman and Presser, 1981:Chapter 11.

4. See Cannell, Fisher, and Bakker (1965) for interesting examples involving retrospective reports of hospitalization in government surveys.

5. We use percentages here for simplicity, but these involve floor and ceiling effects that can be avoided by using logit transformations.

6. Given only this latter result, it would have to be concluded that no change had taken place ($L^2 = 0.27$, d.f. $= 1$, *ns*).

7. The term "random response error" continues to be appropriate, since unreliability in measurement has much the same meaning with subjective questions as with other measures.

8. Most of the original data used in the comparison were dichotomous or had been dichotomized by Stouffer. The rest were dichotomized by combining adjoining categories to produce as near to a 50-50 split as possible.

9. We have looked at a sample of precoded responses (intentionally excluded from analysis by Stember) and found statistically significant differences in five of ten items. For both open and closed questions, the differences, while statistically significant, were typically small, averaging less than 3 percent.

10. Comparisons from the appendix were used if data were from total sample and not based on a demographic question. All response categories were used in making comparison; if tabulation presented data on "don't know" responses, they were included in the comparison.

11. In theory, with marginals for a dichotomy near 50-50 and two simple random samples of 1,200, the root mean square difference expected would be approximately 2 percentage points.

12. The panel speculated that this resulted from the lower case-response rate: that is, by losing more uncooperative people at the door, SRC was able to get more cooperation on individual items. Unfortunately, reported item nonresponse was not divided into respondent-oriented (refusals) and interviewer-oriented (no answers) nonresponse, so support for this hypothesis is not available.

13. Comparison is made in manner described in note 10.

14. The authors note that with the unspecified refusals excluded from analysis, there is high agreement in the ranking of reasons for refusals (although there still is a significant difference in the distributions).

15. Wilks had been appointed in 1945 as vice-chairman of the joint SSRC-NRC Committee on the Measurement of Opinion, Attitudes and Consumer Wants (see Chapter 1).

16. This is a matter of judgment for which no direct evidence is available. We note, however, that the widespread adoption of telephone interviewing in survey research has made the adoption of random household selection considerably simpler since interviewers do not have to travel to (randomly selected) respondents' homes.

17. And, thus, to be more likely to produce significant intersurvey differences in estimates when independent measurements are made and sampling variation alone is used as a guide to statistical inference.

18. See Smith (1978, 1980b) for the exact criteria of comparability used. Items usually had exactly the same wordings, but in a few instances putatively equivalent wordings were compared. Time differences between measurements averaged less than three months. The instances consist almost exclusively of measurements of inherently subjective phenomena; however, a few of the comparisons shown in Table 5.8 included measurements of objective phenomena, such as reports of voting. (For 20 measurements, exact sample sizes were not known and estimates were used.)

19. A somewhat similar pattern was found for reports of thefts. Reports of larcenies of under $50, for example, evidenced significant context-induced variations (Z = 3.07), but those involving larger amounts of money and reports of auto thefts did not vary significantly between measurement contexts (Z = 1.20, larceny of $50 or more; 0.15, theft of car, from Cowan, Murphy, and Wiener, 1978:282).

20. In these comparisons and apparently in the analyses of others (except Cohen, 1980), simple random sampling was assumed—even though the sampling design, in fact, employed some clustering. Since this assumption results in an underestimate of sampling variation, we will accept some differences as statistically significant when they are not.

21. In the SRC and *Post* experiments, the sample division was approximately 50-50; in the NORC experiment, it was approximately 66-33.

22. All samples include only married respondents since the marital happiness question could not be asked of unmarried respondents. The samples were drawn from the adult (18+) population of the continental United States. The SRC and *Post* surveys used telephone interviews and random digit dialing samples; the NORC surveys were face-to-face interviews using a multistage area probability sample (see National Opinion Research Center [1980] for description of sample design). There was random assignment of respondents to experimental conditions. Selection of individual respondents in households in the NORC and SRC surveys was done by random selection from among all eligible members of the household; in the *Post* survey, selection was from among those eligible members of the household who were at home at the time of the initial contact.

All data are unweighted. Due to an error in the preliminary processing of the SRC data, six eligible respondents were eliminated from the tabulation; this error changes none of the SRC results by more than 1 percent.

Data from the NORC survey include only respondents who reported that they had a telephone in their residence in order to make the sample comparable to those of SRC and the *Washington Post*. (Respondents who refused to give their telephone number to the interviewer [about 3 percent of sample] and who, thus, were not asked the location of their phone, were included in this tabulation.)

23. That particular constraint is perhaps a little artificial, although it fits the data very well, and is not essential to the idea.

6

Measurement and Structure: Strategies for the Design and Analysis of Subjective Survey Data

In this chapter we do not treat the topic of survey analysis comprehensively, but rather call attention to some approaches to the scientific analysis of survey data that are either novel or underused. We hope that the following material will stimulate readers to consider some of the less obvious possibilities for detecting and analyzing structure in survey data. Here, as in Part IV (where we review selected approaches that may lead to improvement in survey design), we identify some lines of work that seem promising, without prejudice to many other approaches that may also merit consideration.[1]

We believe that, in principle, survey analysis requires even-handed and balanced attention both to the respondents and the ways in which they differ among themselves and to the questions or items investigated. If this

principle seems platitudinous, it may be because some of its implications have not been carefully considered. In Section 6.1 we develop the distinction between objects and subjects, that is, between items and respondents, and we review various examples of item-centered analysis. We suggest that some of the more impressive substantive achievements of survey research have emerged from object-centered analysis and that continuing emphasis on this approach will be salutary.

In Section 6.2 we review some studies that attempt to describe the relationships between measurements of subjective phenomena and more objective (factual) measurements. Some of this work is related to the item-centered research described in the first section, and some of it is respondent-centered. In both kinds of research we are interested in the subjective-objective connection for two major reasons. On one hand, there are instances in which a systematic relationship between subjective and objective variables serves, in a sense, to validate our measurements of the former. On the other hand, if such a relationship were always found, one could argue that subjective variables are redundant and superfluous. The examples we review show that no simplistic formulation of this kind is adequate and that measurements of subjective phenomena can indeed be of great theoretical interest.

In Section 6.3 we consider some of the recent developments concerning statistical methods and models for analyzing response structures. A common feature of the problems discussed in this section is that they involve responses to two or more items rather than the response to a single question that is the explanandum in the conventional poll report. Critics of polling often deplore the reliance on the single question, but they are not always explicit about just what the alternative involves: this section suggests a few of the possibilities.

Finally, in Section 6.4 we provide a brief and elementary introduction to a new approach to the use of multiple responses in surveys, the Rasch measurement model (see also Duncan, Chapter 12, Volume 2).[2] The thematic connection of the Rasch model with the first three sections of this chapter is its symmetrical treatment of item parameters and person parameters. The gist of the model, for a dichotomous response, is that the odds that a respondent will give one answer, rather than the other, is a product of two parameters, one unique to the item or question and the other a parameter unique to the person responding.[3] The Rasch model is of great conceptual interest because it urgently invites us to analyze survey responses as a joint function of both the particular questions that are asked and the people who answer. It thus enforces an appropriate and firm distinction between statements about the items in a questionnaire and statements about the respondents.

6.1 Objects and Subjects

The main argument of this section is that some of the potential contributions of surveys are missed if there is less than adequate attention to item-centered analysis. The point can be made without in any way questioning the value of the classic approach of survey analysis (as expounded, for example, by Rosenberg 1968), which is concerned with variation among respondents.

For concreteness, consider the generalized form of a data matrix that arises in both survey research and psychometrics:

	Items (i.e., objects)						
Respondents (i.e., subjects)	**1**	**2**	**3**	.	.	.	**k**
1							
2							
3							
.							
.							
.							
n							

Each row (which might stand for one or more punch cards or for a single record on a computer tape) contains the coded responses to all items obtained from one respondent. Each column contains the coded responses to a single item from all respondents. The numbers coded in the data matrix may refer to magnitudes or to ordered categories, or they may be the mere labels of the categories of a nominal variable. In the usual kind of survey analysis, one is concerned with a selected column because it contains information on the "dependent variable," the response for which an explanation is sought. The search for explanation is pursued by sorting on one or more of the other columns in anticipation of finding variables associated with the dependent variable. Or the values in one column will be correlated with (regressed on, or whatever, according to the statistical model being used) values in one or more other columns. In the examples of survey analysis most obviously relevant to this report, the dependent variable will be a measure of some subjective phenomenon, and the independent variables may be either other such subjective variables or factual characteristics of respondents, such as their age, sex, ethnic group, education, and socioeconomic status. In specialized applications, survey research may take account of a network of social relationships among respondents. In a household survey, for example, in which both spouses in a husband-wife household

are interviewed, the analysis will somehow take into account the marital connection between respondents in the two different rows. But such elaborations of the typical survey-analysis strategy—which have been strongly urged by Coleman (1964:89) and which seem to be called for by arguments linking social clusters or networks to attitude change (Abelson, 1979)—do not compromise the distinction featured in the present discussion.

An alternative way to examine the data matrix is to consider the reasons for variation along a row (or the aggregate of all rows). Of course, if items are disparate in form or topic, the main explanation for variation among items is simply that different questions are being answered. But often a survey includes a number of questions of the same form, with identical precoded response categories, addressed to different topics. A striking feature of the annual General Social Survey (GSS) is that it includes nearly a score of such item sets (apart from items concerning factual matters). For example, Q. 62 (as listed in the 1980 GSS codebook) mentions various "problems in this country" and for each problem the respondent is asked "whether you think we're spending too much money on it, too little money, or about the right amount." There are 11 problems altogether, beginning with "space exploration program" and concluding with "welfare." Among other things an analyst might learn from the resulting data is that in 1980 about 67 percent of respondents said the United States was spending "too little" on "halting the rising crime rate" but only 5 percent accepted the "too little" category for "foreign aid." The data thus invite a comparison among items that will presumably have a rather different focus than an inquiry into what kinds of respondents favor and oppose government spending.[4]

A rather different example is provided by Q. 120, which presents the respondent with a list of 13 "qualities" that one might desire a child to have and asks that each quality be rated as extremely, very, fairly, not too important, or not at all important. The leading quality, by a considerable margin, appears to be "that he is honest," which was rated "extremely important" in 1980 by 66 percent of respondents; by coincidence, just 21 percent rated as "extremely important" both the traits "is a good student" and "gets along well with other children."

The social processes giving rise to such decisive contrasts in aggregate judgments are not well understood. One reason for that may be that researchers eager to get on with a research problem defined in respondent-centered terms may regard the items as interesting only as indicators rather than as important research issues in their own right. In the research of Melvin Kohn, from which the question on child qualities originated, the investigator was most interested in the discovery that respondents in higher social class positions were more likely to evaluate highly the qualities suggestive of "self-direction" and less likely so to evaluate qualities reflect-

ing "conformity" than were respondents in lower class positions (Kohn, 1969:Chapter 4). Kohn combined information on all 13 child qualities to produce a single measure of self-direction versus conformity and showed a subtantial positive correlation between class and self-direction.[5] An investigator with different interests might have questioned why some items are more highly evaluated than others rather than considering what kind of respondents have high and low evaluations of one or another item. While Kohn's research is perhaps representative of survey research using multiple items, there are, nevertheless, some interesting examples of item-centered survey research. The rest of this section reviews some of those examples.

OCCUPATIONAL PRESTIGE

The social standing of occupations was systematically investigated at least as early as 1925. As a topic for a modern kind of survey it has been prominent since the development of the classic North-Hatt scale on the basis of data collected in a 1947 survey by NORC. The scale pertained to some 90 occupations that were rated by each of some 2,900 respondents as having excellent, good, average, somewhat below average, or poor standing. (The term "prestige" was not part of the instructions to the respondent, although it has come to be the shorthand term used by sociologists to refer to "general standing" or social status of occupations.) The study was sophisticated in some ways. It was preceded by a fairly large pretest. Occupations were listed in four blocks that were systematically rotated between respondents to avoid fatigue effects. Although the work was interrupted by the premature death of one of the investigators, the legacy of the study permitted quite intensive methodological studies by secondary analysts.

Reiss carefully examined sources of variation in rating and reached two important conclusions that countered much contemporary sociological wisdom (in Reiss et al., 1961:195):

> (1) despite the fact that the social position of a person has an effect on his ratings of some occupations, the prestige status of occupations is viewed in substantially the same way by major social groupings in American society; (2) despite the relative lack of consensus on the criteria for rating a job as having excellent standing, there is almost complete agreement on the nature of the occupational prestige structure of American society, at least where the rated occupations are highly salient. It seems apparent that even if persons do not have a uniform conceptualization of the "general standing" of occupations, or a common frame of reference in terms of which they view the occupational structure, there is a convergence of their perspectives such that they perceive the occupational structure in much the same way.

Without waiting for this good news, sociologists had already begun to make extensive use of the North-Hatt scale. Its popularity as a research instru-

ment was great enough to justify an updating and revision of scale values using survey data collected by NORC in 1963–65 (see Appendix G, Cumulative Code Book, General Social Survey, 1980). The 1963 results demonstrated another most important property of occupational prestige, its stability over time—at least for the period 1925–63 (Hodge, Siegel, and Rossi, 1964). In the meantime, Inkeles and Rossi (1956) had shown that the prestige structure is much the same in countries thought to be quite dissimilar in many ways, the United States and the Union of Soviet Socialist Republics. Reiss's analysis and a variety of other studies further documented that at least the ordinal properties of the measured prestige structure are quite robust against variations in the form of the rating task performed by respondents, in methods of summarizing or scaling the responses, and in composition of samples of respondents. Closely similar results are obtained from men and women, from students and the general adult population, from white-collar and manual workers, and from well-educated and less well-educated respondents.

The generality of the phenomenon of occupational prestige and its almost invariant character under variations of time, place, method of ascertainment, and characteristics of raters have been documented most extensively by Treiman (1977). In a review of some 60 studies in as many different countries, he found an average correlation of about .8 between the prestige scores of pairs of countries taken at random from the set of 60. This finding leaves room, of course, for influences of cultural idiosyncracies as well as of measurement error, but it points to the likelihood "that a very strong common dimension underlies the manifest prestige structure of all societies" (Treiman, 1977:80). The ultimate product of Treiman's compilation of information on this topic is a Standard International Occupational Prestige Scale. This instrument promises to be as valuable in comparative studies of social stratification and mobility (among other topics) as its predecessor U.S. scales have been in American sociology.

It should be noted that the robustness of measures of occupational standing is not something that appeals to common sense or that was anticipated by social scientists working on the topic. In many instances, research workers expected greater differences among prestige structures over time, between countries, or among raters than they actually found. Only in retrospect might we think of the results of this sequence of research as other than remarkable.

One of the interesting consequences of the research on occupational prestige is that quite good estimates of the prestige structure can be obtained using samples of raters much smaller than the usual national samples; the normal precautions of a rigorous sample design are not really critical to the quality of the results. This follows from the fact that what is actually being measured is what the psychologist S. S. Stevens (1966) called

"the social consensus" and what Emile Durkheim, had he known about it, would have called a "collective representation." If there really were a perfect consensus, apart from idiosyncratic ratings and random variation of individual raters, then one might recruit respondents in almost any fashion (except according to known idiosyncracies of their responses), average out their errors, and obtain an estimate of each occupation's prestige with a calculable measure of precision. The actual situation is not quite perfect, but Hope (1982) has demonstrated that it is close to being so.

Recent methodological work and a beautifully designed and executed study of occupational standing in Britain (Goldthorpe and Hope, 1974) demonstrate that there is still room for much work to strengthen the foundations of our understanding of the phenomenon of prestige (a point to which we return in Section 6.2), even if it seems unlikely that major revisions of current estimates of relative prestige will occur. In a nicely designed experimental test of the "axiom of choice" of R. D. Luce (1959, 1979), Skvoretz, Windell, and Farraro (1974) used a model for paired-comparison data to generate prestige-scale values. Their suggestion that the method be extended to a national survey has not been taken up, although the British study clearly demonstrated the feasibility of using sophisticated scaling procedures in a national survey. The work suggested by Skvoretz and his colleagues would be in the spirit of the Rasch model (see Section 6.4), and we suggest it is time for survey research to consider the use of rigorous models of measurement.

SERIOUSNESS OF CRIMES

Among the examples of "a metric for the social consensus" given by Stevens (1966) is one that has been followed up on a large scale in recent survey research. As part of the ongoing National Crime Survey (administered by the Bureau of the Census for the Bureau of Justice Statistics), Wolfgang (1978) and his coworkers prepared a list of 204 offenses, with such items as

—A person steals property worth $10 from outside a building.

—A person robs a victim at gunpoint. The victim struggles and is shot to death.

—A man forcibly rapes a women. Her physical injuries require hospitalization.

Each respondent assigned scores to 25 of the offenses to represent "how serious YOU think certain crimes are." Zero was designated as the score to be used if the act was not regarded by the respondent as a crime; and the respondent was instructed to assume that "A person steals a bicycle parked on the street" has been given a score of 10 to show its seriousness. Then she or he was told to assign scores to the other crimes by comparing each of them with the bicycle theft, taking account of whether a fraction or a multi-

ple of 10 would be a suitable score. There was no upper limit. The final estimate for each crime was the geometric mean of the scores assigned by all respondents who rated that crime. Different respondents were given different sets of 25 crimes so that scores were obtained for all 204 offenses. The study was replicated in six successive months of 1977 and similar results were obtained from each month, except for some indication that as interviewers became more familiar with the task after the first month or two the interviewer contribution to variation in ratings diminished.

For five of the offenses, all thefts, the only difference among the crimes is the dollar value of what is taken. The values $10, $50, $100, $1,000, and $10,000 were used. The geometric mean of seriousness score is related to the dollar value by the function

$$Y = 21.88\ X^{0.27}$$

where Y is seriousness and X is dollar value, the fit of the function (which is linear when logarithms are taken) being quite close. This result is taken by the investigators to demonstrate the scalability of seriousness of crimes by the magnitude-estimation method. Note that since the exponent is much smaller than 1, doubling the value of the item taken does not come close to doubling the seriousness of the offense; rather, it increases the seriousness by a ratio of about 1.2.[6]

Wolfgang has not reported (as of 1981) on the possible variations in perceived seriousness of crimes by groups differing in regard to demographic, regional, community size, and victimization experience variables, although such analysis is contemplated. Earlier research on a less ambitious scale suggests that the ratings may be found to be robust although the degree of consensus may not be the same for all offenses.

Other examples of stimulus scaling with implications for survey methodology include work by Lodge and others (1975) on scaling of adjective phrases that could be used to describe a respondent's "strength of support" for a governmental entity such as the U.S. Senate or the Suffolk County police, or for a specific action such as the Supreme Court's decision on abortion.

As a general comment on the body of work inspired by S. S. Stevens's methods of psychophysical research (not just the studies by Wolfgang and by Lodge and others), we observe that the statistical underpinning of the results is no stronger than that of the work on occupational prestige before the 1970s. The followers of Stevens, like Stevens himself, have been greatly concerned with the form of the function that relates objective magnitudes to perceived magnitudes, and they have shown that systematically different results are obtained using categorical judgments and magnitude estimation. But they have provided no model for an individual response other than an

implicit model of random errors that may be averaged out either within or between subjects. The lack of sophisticated statistical evaluation stands in marked contrast to the paired-comparison method, which provides quite rigorous tests of hypotheses about scalability.

THURSTONE ATTITUDE SCALE

We take brief note of Thurstone's method, not because of any substantial use of it in surveys, but because the method explicitly separates the grading of objects (statements) as to their favorableness or unfavorableness and the assessments of positions taken by subjects or respondents. McNemar (1946:301) provided a succinct summary of the method as applied, for example, in a study of attitudes toward "the church:"

> . . . the scheme is one for arranging items on an 11-point scale according to the degree of favorableness or unfavorableness as determined by having a large number of judges sort quite a number of statements into 11 piles. The median judged location for an item is its assigned scale value. Objective criteria were developed for eliminating items which are ambiguous and which are irrelevant in the sense of not belonging to the continuum from the viewpoint of the judges as sorters. The process . . . begins with the assembling of a much larger number of statements than needed in order to be able to select . . . a desired number . . . with scale values nearly evenly spaced along the assumed continuum. . . . In taking the test, the respondent is instructed to check those statements with which he agrees, and his score is the median of the scale values of the checked statements.

Much of the criticism of the Thurstone method pertained to the last step, the scoring of respondents' attitudes. But the method fell into disuse not because of the criticisms but because other methods that bypassed the work of the "judges" seemed to yield equally satisfactory results, at least in terms of scoring respondents. In the Likert method, for example, each statement is accompanied by answer categories, such as strongly agree, agree, disagree, strongly disagree, which are assigned the values of successive integers. The score is the sum of the values for all questions answered. No scaling of items results from this approach, and the questions themselves are not presumed to have any particular interest as entities in their own right. The Likert procedure, therefore, goes about as far as possible toward the practice followed in classic ability-test procedures. Of course, item analysis can be carried out to determine whether items are correlated with the total score in a manner deemed satisfactory, and item statistics similar to the "difficulties" of items in an ability test can be calculated. But such procedures are seen as strictly subordinate to the aim of scoring respondents. (We take note in Section 6.3 of a different alternative to Thurstone's technique, the Guttman scale, which offers the possibility of simultaneously obtaining a scale of items [objects, or statements] and a scoring of subjects.)

ETHNIC DISTANCE

The work of E. S. Bogardus on ethnic distance is interesting in that it combines features of both the Thurstone and Guttman approaches, although it anticipated them by some years. Moreover, two distinct structures are investigated by means of this scale. Bogardus, whose first paper on this topic was published in 1925, developed 60 items, including

I would admit _____ to my street as a neighbor.
I would marry _____.
I would exclude _____ from my country.

A group of supposed experts rated the items by the amount of intimacy or "sympathetic understanding" expressed, using a 7-point scale. The average rating for each item was computed and the scale was constructed from 7 items whose averages were closest to integer values (Laumann, 1973: 215–216). The instrument as administered to respondents has the following instruction: "For each race or nationality listed . . . circle each of the classifications to which you would be willing to admit the average member of that race or nationality (not the best members you have known, or the worst). Answer in terms of your first feeling reactions." The "classifications" are

1. to close kinship by marriage;
2. to my club as personal chums;
3. to my street as neighbors;
4. to employment in my occupation;
5. to citizenship in my country;
6. as visitors only to my country;
7. would exclude from my country.

Selltiz, Wrightsman, and Cook (1976:422) note that respondents generally treat the 7 statements as conforming to the principle of a cumulative scale, although reversals sometimes occur, perhaps because of "the intrusion of some factor other than the individual's own attitude toward the group in question—for example, the respondent's image of how other people would interpret his living in a certain neighborhood, or his expectation concerning the impact on real estate values of admitting minority-group members to residence on his street." Rose (1951:52) notes some advantages of the Bogardus instrument:

> First, the Bogardus scale, while it elicits expression of verbal opinions, tries to make these opinions as realistic, as simple, and as close to actual behavior as possible. . . . Second, the Bogardus Scale assumes realistically that a person's behavior is not a monolithic "for" or "against" a social object, but a function of the social situation in which the favorable or unfavorable behavior is to be

> expressed. . . . A third advantage of the Bogardus scale is still not fully appreciated today, and consequently there may be some controversy over my evaluation. It is apparent that Bogardus had a rather thorough "intimate" acquaintance with the area of ethnic behavior and attitudes before he constructed his scale. From this, he could set the reasonable outer limits of favorable and unfavorable attitudes. . . . He did not, as Thurstone did, choose his items from the literature at random, but he had knowledge and a point of view in selecting them.

Whatever the merits of Rose's specific observations, the point is well taken that item-centered analysis is likely to be productive to the extent that the selection of items (objects) is congruent with a well-grounded understanding of the empirical domain under study.

In applying the social distance scale, a second form of item-centered analysis comes into the picture, for interest is now focused on the distances assigned to various ethnic groups. (Incidentally, Bogardus used the term "race" in a way that would not be acceptable today.) In 1926, Bogardus secured responses from some 1,725 "selected persons throughout the U.S." (evidently, college students for the most part) concerning 28 ethnic groups, and he computed the mean distance for each group. The means ranged from 1.06 for English to 3.91 for Indians (from India), when the numbers assigned to the statements in the list already given serve as scores. The work was repeated in 1946, 1956, and 1966, with the addition of two ethnic groups to the list (Bogardus, 1968). There are a number of changes, mostly small, in the rank order of ethnic groups by mean distance score, but the impressive finding from these data spanning some four decades is the stability of the structure in terms of rank order. In terms of actual score values, it is noteworthy that the groups with the highest social distances show a rather pronounced decline of social distance over time.

Bogardus's work does not meet modern standards for quantitative research in regard to sampling of respondents and treatment of data. However, as with the work on occupational prestige, it may well be that for purposes of finding the rank order of ethnic groups its deviations from standard research design are not terribly important. As far as we could determine, the Bogardus scale has not been used in modern surveys of the general population. However, a somewhat similar approach has been taken, namely, to measure the prestige of ethnic groups in the same way that the prestige of occupations is measured. Pineo (1977) presents prestige scores for 36 ethnic groups obtained from about 300 English Canadian and 93 French Canadian respondents in a national survey. He notes a high degree of between-group consensus on the prestige ratings within the English and the French populations. For example, the correlation of prestige scores awarded by blue-collar workers and by white-collar workers was .99 within the English sample and .96 within the French sample. But the between-

population agreement is less: the aggregate of English ratings correlated .84 with the aggregate of French ratings. Pineo also cites a correlation of .95 between the English ratings of 19 ethnic groups and ratings obtained from 445 U.S. respondents in a national survey. He further finds that for 19 ethnic groups in the 1956 Bogardus study of college students, the correlation is .95 with the Canadian prestige scores. (The similarity of the two sets of results is a happenstance. The 19 groups are not the same in the two analyses. Moreover, if Bogardus's 1966 data are used, the rank correlation is .90. Pineo apparently assumed that the 1956 results were more reliable.)

VIGNETTE DESIGN

In the item-centered techniques so far discussed, the respondent or "judge" is just given the name of the object (item) or, in the case of the Thurstone technique, is given the object itself, a verbal statement. But it is possible to elaborate the procedure by presenting standardized descriptions of objects incorporating two or more dimensions on which the objects may vary. For example, Triandis and Triandis (1965:208, 210) asked

> . . . white American subjects to react to hypothetical persons who were described in greater detail than is customary in studies of social distance. Thus subjects were asked to indicate their acceptance of a Negro who is a physician, Portuguese, and a Roman Catholic. . . . The first study that used this technique showed that American white subjects paid some attention to all four of the characteristics used in the study and attached the following weights to them: race 77, occupation 17, religion 5, and nationality 1.

Somewhat similar procedures were suggested by Westie (1952) and Hamblin (1971).

Further development of this strategy is summarized by Rossi (1979), who attributes the technique of "vignette analysis" to unpublished communications of Paul F. Lazarsfeld (circa 1950). Rossi illustrates its use in eliciting criteria of social status, but he also indicates other uses. The technique involves presenting to respondents standardized descriptions of households that vary in regard to six characteristics: husband's occupation, wife's occupation, each spouses's educational attainment, ethnicity, and type of housing. Some 40 different combinations of values of the 6 variables were constructed, and 496 respondents rated the 40 combinations (vignettes). The ratings were then regressed on the characteristics of the vignette households and also on characteristics of the respondent. (Note the interesting variation on the techniques described earlier, for which only the average scores of the objects were computed.) However, it turns out that respondent characteristics do not have much to do with the ratings, and their effects can be reduced to insignificance by standardizing each respondent's

score distribution over the 40 vignettes rated. The regression on vignette characteristics shows that all six characteristics contribute significantly to the rating, although husbands' occupations and educations have somewhat larger coefficients than the corresponding characteristics of wives.

Rossi mentions some other examples of vignette analysis. In one study, incomes were attached to the characteristics of vignette households, and respondents were asked whether the persons portrayed were fairly paid, overpaid, or underpaid. This information was used to analyze the public's concept of distributive justice. In another study, the vignettes described characteristics of criminals—their offense, age, and previous record—and the vignettes were presented to public officials to elicit the principles used in sentencing. A study in the area of child abuse obtained ratings of "seriousness" of abuses. In all these examples, there is clearly an effort to probe the social definitions of conduct held by the public or by various groups. The research design is well suited to the determination of a "social consensus," if one exists, or to detect differences among groups in regard to their definitions.

6.2 Relating Measurements of Subjective and Objective Phenomena

In the examples presented to illustrate the idea of item-centered analysis, respondents' (subjective) judgments were elicited in association with specific stimuli. The stimuli in the examples are names of occupations, descriptions of offenses, adjectives, statements about some object or issue, names of ethnic groups, or descriptions of households in terms of status characteristics. In all these studies, the interest is in differential response to different stimuli. If the stimuli were themselves measured and the responses to them were quantified, the possibility would exist of proposing a mathematical relationship or "law" to describe the subjective-objective nexus. And some two centuries of work in psychophysics (Marks, 1974:Chapter 1) could be drawn upon for methodological prototypes. To a limited extent, such an approach was followed in some of the research summarized above. In the analyses of seriousness of crimes, a power function was found to describe the relationship of seriousness score to the dollar value of the goods involved in a theft. In the study of determinants of household status by means of vignette analysis, the subjective status scores were regressed on the quantified (objective) status characteristics. But this kind of subjective-objective correlation was not part of the design of the studies of occupational prestige, attitude scale construction, and ethnic distance. In some in-

stances, working out that correlation has been a task undertaken in subsequent research.

The data of the original North-Hatt occupational prestige study were analyzed by Duncan (1961) from the standpoint of objective correlates of the ratings of general standing. Occupational titles that could be considered reasonably well matched in the Census occupational classification were identified. For each of the 45 occupations, three variables were constructed: X_1 = percentage of respondents rating occupation "excellent" or "good"; X_2 = percentage of men in the 1950 experienced civilian labor force with incomes of $3,500 or more in 1949; X_3 = percentage of men in the 1950 experienced civilian labor force having graduated from high school. The latter two variables were constructed from 1950 census data giving detailed characteristics of the incumbents of each occupation. The multiple regression equation

$$\hat{X}_1 = 0.59X_2 + 0.55X_3 - 6.0 \qquad (R^2 = .83)$$

was obtained. The relationship was deemed close enough to justify use of the census data on income and education to construct socioeconomic scores for occupations for which prestige ratings had not been obtained. A much more extensive study using 426 rather than 45 occupations with more recent census data and prestige scores obtained a similar equation, with $R^2 = .78$ and with a somewhat higher relative weight for education. The variation in results was attributed by the authors (Stevens and Featherman, 1981) to Duncan's use of a biased sample of occupations.

Empirical results of this sort provided the point of departure for Treiman's attempt at a theoretical rationalization of the worldwide similarity of occupational prestige structures. In summary, his argument (Treiman 1977:2) is "that the division of labor creates inherent differences in the power associated with various occupational roles wherever they are found, that these differences, in turn, create differences in privilege, and that power and privilege beget prestige. Since a similar complement of occupational roles tends to arise in all complex societies, the resulting prestige hierarchies will also be similar." Comparative data supporting the argument are fragmentary. Formal educational attainment is not a completely adequate measure of the skills and training required by occupations, and income is only one form of power. But it is challenging that the available data (Treiman, 1977:115) do indicate a "basic intersocietal similarity in the relative education required to perform various occupations and in the relative income returned to them." Moreover, Treiman (1977:115) notes that "for the most part the connections between educational requirements, income gained, and prestige are similar throughout the world." Later research no

doubt will modify or qualify Treiman's conclusions (see Haller and Bills, 1979).

An interesting use of historical data that partly confirms Treiman's theory is reported by Sewell (1982), who compiled status indicators from marriage licenses in nineteenth century France for a list of 53 occupational groups. His "summary scale" correlates .74 to .78 in three different periods with Treiman's international scale, suggesting a notable degree of stability in occupational rankings for a century or more.

In regard to the measures of social distance for ethnic groups, Duncan and Lieberson (1959) studied the spatial segregation and assimilation of ethnic groups in Chicago. They ranked 16 European nationalities on the basis of the segregation of foreign-born whites of those nationalities from all native whites of native parents in the city, as of 1930, and found a correlation of .73 with the ranking of these groups in terms of social distance as indicated by the Bogardus "neighbors" question. For the same 16 groups, the rank correlation with social distance was .80 for segregation of native whites of foreign or mixed parents versus native whites of native parents. Moreover, a crude indicator of the extent to which the foreign groups had intermarried with the native population correlated .83 with the ethnic group ranking based on the Bogardus "marriage" question. Duncan and Lieberson (1959:373) note: "Bogardus found that the social distance rankings of ethnic groups have been relatively stable over time, as were our segregation indexes. Consequently, it is difficult to detect any significant change in the relationship between the two." Apparently, the subjective measure of ethnic distance captures something real about the pattern of intergroup relationships. The form of the evidence is not, however, such that one would want to base strong theoretical conclusions on it.

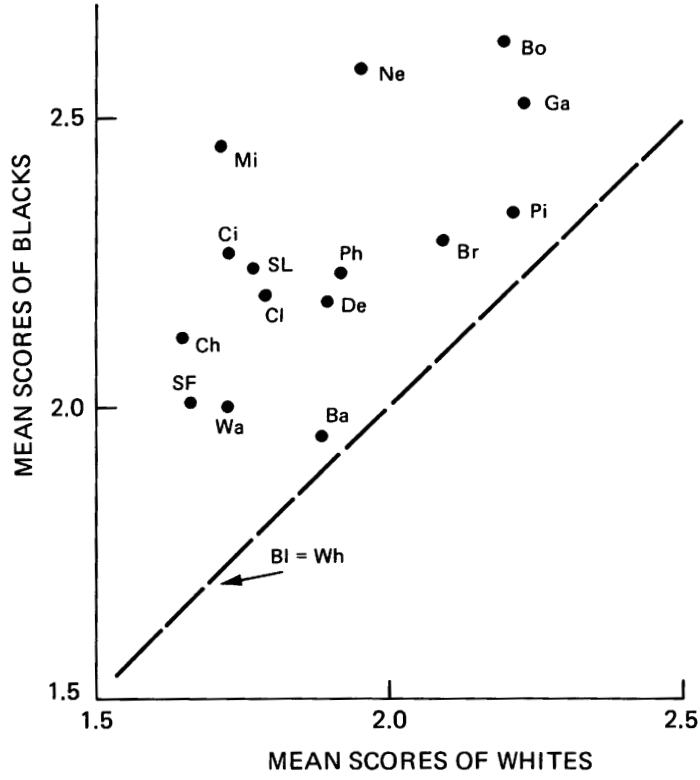
An interesting feature of these two examples is that the data describing objective properties of the patterns of occupational inequality and ethnic segregation come from a different source than the subjective measures. Thus, the utility of the survey data may be greatly enhanced when the objects are reasonably well-defined and socially recognized, because in that event auxiliary data sources may well be available (see Recommendation 16, Chapter 10).

In all the examples considered so far, each respondent was invited to make an explicit comparison of items. A study by Schuman and Gruenberg (1970) illustrates a design in which each respondent evaluates just one object, but a comparison of objects is made possible by the aggregate data. Approximately 175 white and 175 black respondents in each of 15 cities evaluated services provided by the city government. The mean scores of blacks were plotted against the mean scores of whites, as shown in Figure 6.1. The correlation coefficient for the 15 pairs of means is .61. The authors point out (Schuman and Gruenberg, 1970:234, note 16):

We must be careful not to infer that blacks and whites *within* cities agree in absolute scores; on the contrary, . . . racial scores usually differ within cities. Despite the similarity in city *orderings* . . . within each of the fifteen cities, black dissatisfaction is greater than white. Nor does a high positive correlation indicate that blacks and whites consciously tend to place their city in the same rank position vis-à-vis the other fourteen cities. Respondents neither were nor could have been asked such a question. Our interpretation of a high positive correlation is that it indicates that each city possesses a reality which affects all its inhabitants in much the same way, even though the levels of effects for blacks and whites differ across the sample as a whole.

FIGURE 6.1

Black versus white mean scores on "Dissatisfaction with neighborhood services," 15 cities, 1968. (Cities are Baltimore, Boston, Brooklyn, Chicago, Cincinnati, Cleveland, Detroit, Gary, Milwaukee, Newark, Philadelphia, Pittsburgh, Saint Louis, San Francisco, Washington D.C.)



SOURCE: Schuman and Gruenberg (1970: Figure 3 and Table B1).

Research designs involving respondents' reactions to objects in their external milieu account for only a small though important part of work using subjective survey measurements. Other designs ask a respondent to take herself or himself as an object, as it were, and to report something about that object. Here, too, there are interesting questions concerning the correlation between objective and subjective variables. We consider two examples, the contrasting outcomes of which are instructive. In 1955 and again in 1971, married women in the Detroit area were asked to report feelings about certain aspects of marriage, including "how you feel about your standard of living—the kind of house, clothes, and so forth." There was no statistically significant change in satisfaction from 1955 to 1971, but in both years a positive regression of satisfaction on family income was noted. Analyzing the two years together showed that the data were consistent with a common within-year regression, but required a lower intercept for 1971 than for 1955 when the regressor was current dollar income (or its log). When the 1971 incomes were deflated by approximately 1.5 to take account of the increase in consumer prices, the year effect was lessened, but it was still in the direction of suggesting lower satisfaction in 1971 than in 1955 at comparable incomes. However, the year effect was fully removed (or even reversed) when the regressor was changed from income per se to position in the income distribution (that is, the proportion of Detroit families with incomes below the respondent's). The investigator (Duncan, 1975:267) interpreted these results as supporting the thesis of Easterlin (1973) that "Individuals assess their material well-being, not in terms of the absolute amount of goods they have, but relative to a social norm of what goods they ought to have." Evidence that the norm itself evolves was found in Rainwater's analysis of results of 18 Gallup Poll surveys taken between 1946 and 1969 that asked the question, "What is the smallest amount of money a family of four needs to get along in this community?" According to Rainwater's estimate, this amount, expressed in constant dollars, increased by a factor of 1.36 between 1954 and 1969 (about the same period covered by the Detroit data). Between 1955 and 1971 median family income in constant dollars was estimated to have increased by a factor of 1.42. Evidently, standards escalated along with real incomes; hence there was no net gain in satisfaction.

The implications of results like these are not trivial, for they call into question the whole thesis that material progress is the route to a more satisfactory life. Easterlin (1973, 1974) analyzed data for 19 countries and reported: "In all societies, more money for the individual typically means more individual happiness. However, raising the incomes of all does not increase the happiness of all." *Within* a country, satisfaction or happiness is, to be sure, positively related to affluence. But between countries with widely differing material levels of living, there are only slight differences in

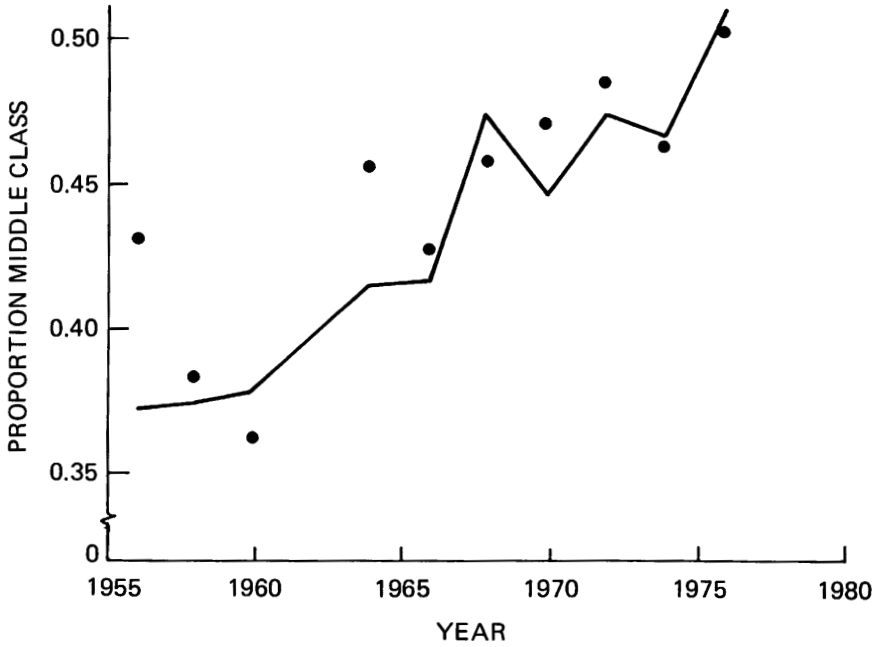
happiness, just as the time series of happiness readings in one country, followed over a considerable period of time characterized by generally rising real income, does not show an upward trend.

A different result is obtained when the subjective indicator is modified. Cannon (1980) studied a 20-year series of surveys that asked respondents to categorize themselves as middle class or working class. There was a general upward trend in the proportion of respondents saying they were middle class, in contrast to the level or erratically fluctuating series of happiness readings examined by Easterlin. Moreover, the trend in proportion of those saying middle class could be fully explained by the regression of response on respondents' occupational class, occupational prestige, educational attainment, and family income (see Figure 6.2).

In this case, one could argue, the subjective datum simply amounts to a perception by a respondent of class position conceived of as some kind of

FIGURE 6.2

Proportion of American adults describing themselves as middle class, 1955–76.



SOURCE: Points are observed sample proportions, white nonfarm males (Cannon, 1980: Table I). Line shows calculated proportion, based on regression on year (not significant), occupational prestige, years of education, logarithm of family income (1967 dollars), and manual vs. nonmanual occupation (Cannon, 1980: Table III). Means of log income kindly supplied by L. W. Cannon; means of other variables from Cannon (1980: Table I).

aggregate of location on several socioeconomic status scales. The time series of subjective measurements then merely registers more or less veridically the general upward movement of socioeconomic status in the United States in recent decades.

By contrast, in the satisfaction and happiness research the subjective data are not mere perceptions, but perceptions filtered by explicit evaluations, made (presumably) according to standards set by the subjects themselves. The variation in standards across time is the uncontrolled variable that evidently accounts for the failure of the subjective variable to increase despite an increase in the level of an important objective factor related to it.

In all of these examples there is obviously room for much theoretical argument concerning the meaning of the results. Our intention has not been to resolve any theoretical issues but rather to suggest the fruitfulness of encounters between objective and subjective variables in posing issues for theory and further research. We have not tried to review any broad selection of studies related to these issues. Perhaps the most active sector at the moment is the field of so-called subjective social indicators. Articles and reviews pertaining to them are featured in the journal *Social Indicators Research*. Some rather elaborate models have been proposed to account for the generation of evaluations of, for example, satisfaction and happiness. And the role and importance of subjective indicators in a program of social indicator development continue to be debated.

6.3 Response Structures

In classical survey analysis, the primary concern was with a situation in which one variable (which we term the "response") is conceived as being dependent on one or more other variables (sometimes termed "factors").[7] Our concern in this section, however, is with complex patterns of association among responses. (A particular model for multiple responses is considered in the last section.) Analysis of multiple responses of any kind may be a preliminary step toward some parameterization or data reduction that leads into something like the classical kind of survey analysis. The latter, moreover, can be combined with aspects of an item-centered approach, and we will present an example illustrating this possibility.

A great variety of possibilities for penetrating analysis of survey data exist today that were not even imagined a decade ago. This should surely have implications for the economics and logistics of future survey research (see Recommendation 8, Chapter 10). We should also point out that the following exposition does not refer to multivariate models that, like factor analysis

and its elaborations (such as Jöreskog's LISREL), depend on the assumption of multivariate normality. This condition infrequently holds for the type of survey data that concerns us. In contrast, the newer methods for analysis of categorical data that we discuss feature distributional specifications that are appropriate to such data.

GUTTMAN SCALE

We described earlier the approach of Thurstone to the scaling of items, a procedure regarded as a preliminary to the scoring of the attitudes of subjects. The approach taken by Guttman (1950) to the same problem sought to derive simultaneously an ordering of items (for example, attitude statements) and a weak ordering of subjects (survey respondents). A set of statements is said to comprise a perfect scale if each statement is always endorsed by subjects endorsing any other statement endorsed by a larger fraction of respondents. In the cross-classification of any two items from a Guttman scale (we take items to be dichotomous for simplicity) there will be one cell, corresponding to "yes" on one of the items and "no" on the other, with zero frequency. (The zero frequency in the sample arises because, according to the model, the probability of that response combination is zero, not because of some accident of sampling.) In practice, when a survey analyst works with attitude statements (whatever may be true of other kinds of items), perfect scales or close approximations to them are not ordinarily obtained. This result occurs because the model of a perfect scale as formulated by Guttman is a deterministic model. Suppose there are four statements; for each statement let the letter *Y* signify that the response is "yes" (rather than "no"). The response pattern *YYYY* means that all four statements are answered "yes"; and *YYYN* means that the first three are "yes" while the last is "no." If the 4 items comprise a Guttman scale, only 5 response patterns can occur: *YYYY*, *YYYN*, *YYNN*, *YNNN*, and *NNNN* (or *NNNN*, *NNNY*, *NNYY*, *NYYY*, *YYYY*). But there are a total of 2^4 (that is, 16) possible response patterns. Hence, there are many possibilities for "error" in response. Most of the work done with the scale idea uses some rule of thumb for deciding whether there is too much error for the items to be deemed scalable and for combining the error-infected responses with scale-type responses to achieve a scale scoring for all respondents.

A more rigorous approach is to adopt some model for responses that incorporates the principle of a Guttman scale but that also includes an explicit mechanism for generating errors. Clogg and Sawyer (1981) review several such models. (The Rasch model, discussed in Section 6.4, may be viewed as still another one.) We consider briefly the model of Goodman (1978), who postulates that there are two kinds of respondents: the intrinsically scalable respondents who give one of the scale-type responses with a

probability of 1.0 and the intrinsically unscalable respondents who choose one of the 16 responses in a random manner (random being explicitly specified by parameters of the model). Under this model, the hypothesis of quasi-independence of items, once the five scale-type responses are deleted, must hold. Hence a test of it provides a test of whether the modified scale model is appropriate in the population under study. For example, Clogg and Sawyer's (1981) analysis of data for the six abortion items in the General Social Survey shows that the modified scale model does not fit the data acceptably well (at the .05 level).[8] However, if one supposes that there are actually two different orderings of the items in use in the population, a satisfactory fit can be obtained. The tentative conclusion is that the abortion items comprise a biform rather than a unidimensional scale. Even so, it is estimated that 19 percent of the population does not use either scale but falls into the intrinsically unscalable class of respondents. (That is, the pattern of response they give to these six items is not one of the patterns permitted by either ordering.) Perhaps these are people who have what Converse (1964) called "nonattitudes"; at any rate, their responses are not structured in either of the two preponderant ways.

LATENT STRUCTURE

Although Guttman's proposal concerning scaling and Lazarsfeld's latent structure approach both were developed in survey research in the military setting of World War II, their subsequent careers have been rather different. Supposed Guttman scales have been constructed quite frequently by research workers, although most of that work would undoubtedly not withstand close scrutiny under the rigorous testing of a scaling model of the kind proposed by Goodman (1978). Latent structures, however, have not been used at all widely and remain a topic largely of interest to methodologists. To some extent, this lack of interest may merely reflect the fact that methods of estimating and testing latent-structure models readily adapted to realistic research situations became available only a few years ago.

Given a multi-way distribution of several categorical responses, the latent class model proposes that all observed association or interaction (2-way or higher order) of the responses is due to an unobserved classification of respondents by a set of latent categories.[9] Within a latent class, responses to items are mutually independent. Each class has a unique set of parameters describing conditional probabilities of response to all items, given membership in the class. (The latent classes, therefore, function in somewhat the same way as the "ulstrata" we discuss in Section 6.4.) All the "structure" in the model has to do with the distribution of the respondents between the latent classes and the probabilities of response that are conditional on class membership.

It should be emphasized that the latent-structure model is a purely statistical model that makes a mathematically convenient assumption about the mixing of distributions that supposedly takes place when the observed data are aggregated over the latent classes. This point is brought out in the discussion by Madansky (1968:34), who was careful to note that even when the data can be fitted to a latent class model, "it may happen that the . . . latent classes will have no reasonable interpretation, let alone the hoped-for interpretation."

Recent work on latent structures (Goodman, 1978; Haberman, 1979; Clogg, 1979) features models in which constraints are placed on the conditional probabilities in the interest of representing substantive assumptions suggested by the conceptual framework of the analysis. Among other things, as Clogg illustrates, appropriately defined constraints may correspond to assumptions made about the ordinal properties of the manifest variables and latent classes.

There is no basis for forecasting general utility of latent-structure models with most of the multi-item data sets used in current and past surveys. The model makes strong assumptions, which are likely to be justified only if the manifest indicators are carefully constructed or selected with an eye to the plausibility of the assumptions that will have to be made when the latent structure model is estimated. Unfortunately, this is easy to say but hard to implement in the absence of the large body of experimental work that will surely be needed before these models can prove their worth in serious substantive inquiries.

RESPONSE CONSISTENCY

In the latent-class model, observed responses to different items are associated for one and only one reason: the manifest data are a mixture of responses of members of different latent classes. In many situations, this may seem to be an overly strong assumption. There could be a variety of unknown reasons for any pair of items to be associated, some of substantive interest and some, perhaps, of artifactual origin. Abelson (1979:250), for example, notes that attitudes may be correlated for cognitive reasons or because "social clustering begets opinion clustering." In models featuring response-consistency parameters (Duncan, 1979b; Duncan, Sloane, and Brody, 1982) the pairwise (net) associations of items are taken for granted and the model purports to provide only part of the explanation of them. The focus is on 3-way or higher-order interactions of responses of the kind generated when three or more items are answered consistently (all in the favorable direction, say, or all in the unfavorable direction) more frequently (or, perchance, less frequently) than would be anticipated on the basis of pairwise associations alone.

Consider the 4-item situation used previously as an illustration for a

Guttman scale. The 4-way table in Goodman's notation is {ABCD}. In a hierarchical approach to the analysis of the table (Goodman, 1978:Chapter 4), an analyst would ordinarily consider models that fit each of the 1-way marginals {A}, {B}, {C}, and {D}; some or all of the 2-way marginals {AB}, {AC}, {AD}, {BC}, {BD}, and {CD}; and some or all of the 3-way marginals {ABC}, {ABD}, {ACD}, {BCD}. According to the hierarchical principle, fitting any set of 3-way marginals entails fitting the three sets of 2-way marginals contained therein and the three sets of 1-way marginals contained therein. But suppose we consider a model that fits all the 2-way marginals and also includes a parameter δ_1 that pertains to just the one cell corresponding to response pattern *YYYY* and another parameter δ_2 that pertains just to pattern *NNNN*. The model is no longer hierarchical; it includes a constrained form of 4-way interaction. Consider the estimation of the net association of a pair of items, for example, *B* and *D*. If the model is the hierarchical one that fits only 2-way marginals, the association of *B* and *D*, denoted (*BD*), will be the same for any combination of responses to the other two variables, *A* and *C*. Thus $(BD)_{YY} = (BD)_{YN} = (BD)_{NY} = (BD)_{NN}$, where the subscripts denote, respectively, the responses to A and C. But the nonhierarchical model with the two additional response-consistency parameters implies three different values for the net association of a pair of responses; hence, for example, $(BD)_{YY}$, $(BD)_{NN}$, and $(BD)_{YN} = (BD)_{NY}$ will be of three different magnitudes. The response-consistency effect, as measured by the parameters δ_1 and δ_2, will then be reflected in the ratios

$$\delta_1 = \frac{(BD)_{YY}}{(BD)_{YN}}$$

and

$$\delta_2 = \frac{(BD)_{NN}}{(BD)_{YN}}.$$

These parameters work the same way for any pair of responses even though the net associations for different pairs of responses are not equal. For example, $(BD)_{YY}/(BD)_{YN} = (AB)_{YY}/(AB)_{YN}$ even though $(BD)_{YY} \neq (AB)_{YY}$, and so on. Hence, the positive response-consistency parameter has the same effect on all pairwise associations, although those associations are allowed to differ for other reasons; and the negative response-consistency parameter has the same effect on all pairwise associations. In working with the model, attention is naturally given to hypotheses such as $\delta_1 = \delta_2 = 1.0$ (no response-consistency effects), $\delta_1 = \delta_2$ (positive and negative response-consistency effects the same), and $\delta_1 = 1.0$ or $\delta_2 = 1.0$.

Response-consistency effects have been observed in several data sets. One data set included four items reflecting preference for traditional (sex-typed) assignment of household chores to children versus egalitarian assignment. In the Detroit Area Study data, both the consistently traditional and the consistently egalitarian responses were subject to response-consistency effects significantly greater than 1.0 in 1976 and 1971, but in 1953 only the egalitarian response pattern showed that effect. The analyst conjectured (Duncan, 1979b) that the women's movement of the 1960s generated not only an ideological commitment to egalitarianism on the part of some respondents but also a traditional backlash on the part of a smaller number of respondents.

The second example involves four questions on belief in the principle of freedom of speech as applied to speeches criticizing the president, against religion, in favor of communism, or in favor of fascism (Duncan, Sloane, and Brody, 1982). In this instance the only significant response-consistency effect pertains to the YYYY response (all answers affirming the principle of free speech). The conjecture is that ideological adherence to the Bill of Rights is characteristic of some substantial proportion of respondents, but that ideological opposition to free speech as a principle (as distinct from its exercise in connection with a particular issue) is not frequent enough to register in a sample of conventional size. The third example involves six questions on abortion in the General Social Survey, which show response-consistency effects pertaining to both pro-abortion and the anti-abortion positions. Duncan, Sloane and Brody, 1982, conjecture that such effects will usually be found when opinions have been crystallized by debate that features appeal to political, ethical, or religious ideas providing a general rationale for a position on the issue, whatever its specific formulation.

In other sets of items, response-consistency effects have not been detected. This lack of effect was found for a set of four questions regarded as indicators of anomia and for another four questions hitherto treated as measures of political efficacy. These sets of items do not seem to tap some currently debated issue, such as the social role of women or abortion, nor do they relate in a direct way to some general proposition with normative force, such as the free speech provision of the Bill of Rights. Absent some such force that induces respondents to take an ideologically consistent position, the associations among items (though significant and in the direction expected on the basis of thematic or semantic relationships) will not be enhanced by a consistency effect.

The attraction of this kind of model, therefore, is that it may register something about the dynamics of public opinion formation. The task for theory is to specify more fully the circumstances under which opinions tend to become ideologically consistent and thereby lead to responses in which a consistency effect is detected. This discussion is intended only to suggest

the potential value of pursuing such investigations. Much of the literature on "nonattitudes" (reviewed in Smith, Chapter 8, Volume 2) is relevant here, as is the recent work on issue consistency (Nie, Verba, and Petrocik, 1979).

ASSOCIATION OF POLYTOMOUS VARIABLES

In survey analysis, the advantages of dichotomous items are well known, and so response distributions for many questions worded to have three or more answer categories are collapsed for the convenience of the analyst. This practice may not be as innocuous as it seems (Schuman and Duncan, 1974), and, at the very least, it involves discarding expensively acquired information. Although log-linear models have often been presented with illustrative data pertaining mainly to dichotomous variables (Goodman, 1978), the generalization to polytomies is explicitly available. Even more important, specialized models precisely suited to the analysis of contingency tables involving polytomies have been proposed (Goodman, 1972, 1979a, 1979b; Simon, 1974; Haberman, 1974; Duncan, 1979a; Clogg, 1982a, 1982b). Inasmuch as this literature is not yet widely used, we want to give just a little of the flavor of these models without getting into technical details. For illustrative purposes we present a very special case, but one with ample complexity beyond that encountered with dichotomies. The reader interested in general formulations is referred to the cited literature.

We consider a 3×3 contingency table and, to motivate the models that are illustrated, we assume that the same classification pertains to the rows and columns and that its categories are ordered. (Other models can be suggested when any of these conditions is modified.) Such a table might arise, for example, in a cross-classification of the General Social Survey questions on happiness:

[109] Taken all together, how would you say things are these days—would you say that you are very happy, pretty happy, or not too happy?

[110] Taking things all together, how would you describe your marriage? Would you say that your marriage is very happy, pretty happy, or not too happy?

We take the marginal distributions as given and concern ourselves with the *association* of the responses.[10] We are not seeking (necessarily) an overall measure of the strength of the association in the fashion of the earlier work of Goodman and Kruskal (reprinted 1979). Our goal is to describe the pattern or structure of the association, using no more parameters than may be necessary. The values of those parameters will serve in lieu of a measure of association in the pre-1970 sense.

Table 6.1(i) provides notation for the expected frequencies F_{ij} in a 3×3

TABLE 6.1
Notation and Illustrative Patterns of Association in a 3 × 3 Contingency Table

(i)	F_{11}	F_{12}	F_{13}	$\theta_{11} = F_{11}F_{22}/F_{12}F_{21}$	$\theta_{12} = F_{12}F_{23}/F_{13}F_{22}$
	F_{21}	F_{22}	F_{23}	$\theta_{21} = F_{21}F_{32}/F_{22}F_{31}$	$\theta_{22} = F_{22}F_{33}/F_{23}F_{32}$
	F_{31}	F_{32}	F_{33}		
(ii)	s	1	$1/t$		
	1	1	1	$\theta_{11} = s$ $\theta_{12} = t$	
	$1/u$	1	v	$\theta_{21} = u$ $\theta_{22} = v$	
(iii)	1	1	1		
	1	a	a^2	$\theta_{ij} = a$	
	1	a^2	a^4		
(iv)	1	1	1		
	1	a	a^2	$\theta_{11} = \theta_{12} = a$	
	1	a^2b	a^4b^2	$\theta_{21} = \theta_{22} = ab$	
(v)	1	1	1		
	1	a	a^2c	$\theta_{11} = a$ $\theta_{12} = ac$	
	1	a^2	a^4c^2	$\theta_{21} = a$ $\theta_{22} = ac$	
(vi)	1	1	1		
	1	a	a^2c	$\theta_{11} = a$ $\theta_{12} = ac$	
	1	a^2b	$a^4b^2c^2$	$\theta_{21} = ab$ $\theta_{22} = abc$	
(vii)	d	1	1		
	1	d	1	$\theta_{11} = d^2$	$\theta_{12} = 1/d$
	1	1	d	$\theta_{21} = 1/d$	$\theta_{22} = d^2$
(viii)	d	1	1		
	1	da	a^2	$\theta_{11} = d^2a$	$\theta_{12} = a/d$
	1	a^2	da^4	$\theta_{21} = a/d$	$\theta_{22} = d^2a$
(ix)	d_1	1	1		
	1	d_2	1	$\theta_{11} = d_1d_2$	$\theta_{12} = 1/d_2$
	1	1	d_3	$\theta_{21} = 1/d_2$	$\theta_{22} = d_2d_3$
(x)	d	1	1		
	1	da	a^2b	$\theta_{11} = d^2a$	$\theta_{12} = ab/d$
	1	a^2b	da^4b^4	$\theta_{21} = ab/d$	$\theta_{22} = d^2ab^2$

table under whatever model is specified. Inasmuch as the table has four degrees of freedom for association, its pattern of association can be exhaustively described by the four cross-product ratios obtained from the four 2 × 2 subtables formed by a pair of adjacent rows and a pair of adjacent columns of the 3 × 3 table. The right side of Table 6.1(i) illustrates the computation of these four odds ratios (cross-product ratios). The θ_{ij} may be regarded as parameters describing the association. Alternatively, we may propose some

other way of parameterizing it and see how the alternative parameters relate to—or, actually, "generate"—the θ's. In part (ii) of the table we display parameters s, t, u, v in certain cells: these quantities can be used in forming odds ratios in just the same way that the expected frequencies above were used. We are abstracting from the fact that the frequencies have multiplicative components due to the marginal distributions of the row and column variables, because these components cancel out when the odds ratios are computed. In (ii) we see that there is a one-to-one relationship of s, t, u, v to the θ's. Hence (ii) represents a saturated model of the association. While there are many other ways to express a saturated model, considering them does not advance our discussion, and we next consider a selection (by no means exhaustive) of unsaturated models.

In (iii) there is a single parameter for association that appears in the cell with exponent 0, 1, 2, or 4. The pattern is such that every one of the four basic odds ratios has the same value and is, therefore, termed the pattern of uniform association. In this case, the model justifies the use of a single scalar as a measure of association because the association is, in a clearly defined sense, the same everywhere in the table.

In (iv) we find, superimposed on the uniform-association pattern, an association due as well to a second parameter. It has the effect of altering the association in regard to rows 2 and 3 from that obtaining for rows 1 and 2. Hence, this is the row-effects association model. In (v) a transposition of this modification yields the column-effects association model. Note that the two odds ratios pertaining to columns 1 and 2 are the same, but a different value obtains for columns 2 and 3. In (vi) the pattern involves both row and column effects. Three parameters are used: hence this model has $1 = (4 - 3)$ d.f., following the rule of subtracting the number of parameters used to describe association from the total of four available to describe association exhaustively. The 1 d.f. in the row-and-column effects model corresponds to the constraint $\theta_{11}\theta_{22}/\theta_{12}\theta_{21} = 1.0$. This same constraint obtains in (iii), (iv), and (v) as well, but there are additional constraints for those models.

In (vii) we return to a model with a single parameter, but one quite different in form from uniform association. In this model, all the association in the table is attributed to a diagonal effect. Again we could describe association with a single number, the value of d, but it could be misleading, in a sense, to think of this as a measure of "strength" of association for the whole table. If d is greater than 1.0, we see that $\theta_{11} = \theta_{22} > 1.0$ implies positive association, while $\theta_{12} = \theta_{21} < 1.0$ implies negative association.

In (viii) we have a model that includes the diagonal parameter along with the uniform association parameter. Duncan and Schuman (1980) found that this model could be used to describe the cross-classification of two religious indicators used in the Detroit Area Study in 1958, 1959, and 1971:[11]

> Would you say you attend religious services more often, about the same, or less often than you did 10 or 15 years ago?
>
> All things considered, do you think you are more *interested*, about as interested, or less interested in religion than you were 10 or 15 years ago?

In model (viii) the pattern of association may either be isotropic in the terminology of G. U. Yule (all four theta's deviate from 1.0 in the same direction) or anisotropic, according to the relative magnitudes of the parameters a and d. Again, it could be misleading to attempt a summary of the association in the table with a single index or coefficient of association.

Model (ix) illustrates the possibility that not all diagonal effects are the same, although the entirety of the association in the table may still be attributed to the diagonals (with a qualification noted presently). This pattern of association is the one observed by Goodman (1965, 1968, 1969) in a series of papers on occupational mobility (in which the row classification pertains to father's occupational status and the column classification to son's occupational status).

Model (x) returns to the idea of a single diagonal parameter juxtaposed with the uniform association parameter, as in (viii). But in this model there is an additional parameter for row-and-column effects entering the model symmetrically; this is brought about by setting b equal to c in (vi). At first sight, this seems like quite a different conception of the pattern of association from that represented by (ix). But in fact the two models are equivalent. This is demonstrated by the elementary but tedious algebra of solving for the parameters of (viii) in terms of those in (ix), or vice versa, after equating $\theta_{ij}^{(\text{viii})} = \theta_{ij}^{(\text{ix})}$. Inasmuch as the models are equivalent, there is no statistical meaning to the question of which of them is to be preferred. In fact, the analyst can gain a somewhat deeper understanding of the data by looking at them from the perspective of both models.

We do not deal with the apparatus of statistical inference needed for estimating parameters, assaying goodness of fit of models, and choosing between models. The computations are facilitated greatly by Haberman's program FREQ (Haberman, 1979: Appendix). In the earlier literature on models of this kind, various expedient methods of computation were outlined.

The illustrative case discussed here can be generalized in various directions: by using variables with more categories and by considering association of more variables. Clogg (1982b) outlines, with interesting empirical examples, ways of analyzing multi-way cross-classifications from the point of view of models for conditional association, models relating to partial association, and models used to analyze symmetric association. In another paper, Clogg (1982a) takes as the vehicle for his exposition not the log-linear model of association (of the kind illustrated previously), but the log-multiplicative

(or log-log-linear) model suggested by Goodman (1979b) as an alternative in some situations. Of special interest is Clogg's demonstration that the log-multiplicative model is useful in inferring the distances separating the categories of a variable whose categories are known to be ordered but not known to be equally spaced. The metric obtained in fitting the log-multiplicative model sometimes is rather different from that implied by the assumption of equal intervals. (This characteristic turns out to be true for the GSS happiness question, for example.)

RANKING OF OBJECTS

Occasionally a survey question will request the respondent to place in rank order of importance or value several different objects, such as qualities of a child or policies that might be followed by the government. In item-centered analysis, it may suffice to compute the mean rank for each object. Statistical inference for that approach and for the more complicated one in which mean rank is a function of respondent characteristics has been treated by Koch, Freeman, and Lehnen (1975; 1976). However, summarizing the distribution of ranks by the mean loses some of the fine structure of the responses. With five or fewer objects to be ranked, it is feasible to consider specialized models for the whole distribution, given the usual size of survey samples. A surprisingly large number of distinct models are available for even as few as three objects in the set being ranked. When the rankings are related to respondent characteristics, log-linear models specified to bring out particular contrasts may be of interest. In the past, rankings have probably not been used as much as they profitably could, since methods for analyzing the somewhat cumbersome data were not widely known. A survey of some of the serviceable models (Duncan and Brody, 1982) may prove useful in revealing some analytical possibilities that have previously been overlooked.

DIFFERENTIAL RESPONSE

The example of two questions on happiness—one on happiness in general, the other on happiness of the respondent's marriage—again will serve as an example for discussing differential response. We noted earlier that the distributions of the two responses are rather different (rather more people are "very happy" with their marriages than are "very happy" in general). Suppose an analyst wishes to consider the hypothesis that the magnitude of the discrepancy between the two distributions is related to a third variable, such as education. Inasmuch as both the happiness variables and education are polytomous, this formulation of a problem invites attention to the form of association between education and each happiness response. But the additional complication here, if data from the General Social Survey are to

be used, is that the responses are obtained from the same respondents. Education by marriage happiness and education by general happiness are not two halves of a 3-way contingency table: they are two contingency tables based on the same respondents. Despite this dependency, it is possible to specify various models representing the two associations as having the same form and same parameter values and then to see if the data are consistent with such a specification. The necessary modifications of the standard models for multi-way tables have been discussed (Duncan, 1981) in the context of an essay on panel analysis for which a formally parallel problem arises.

BETWEEN-ITEM COMPARISONS

In the first section of this chapter, we noted that analyses combining features of item-centered analysis and the standard respondent-centered kind of survey analysis are entirely possible. We note further, at this point, that when multiple responses are involved, one or another of the approaches to response-structure analysis that we mentioned (or still others) might also be helpful. There are two general points to be made. First, while two or more decades ago the problem of analyzing data was generally seen as requiring one to squeeze the analysis into some form of correlation exercise (or, alternatively, a comparison of percentage-point differences), today we work in a model-rich environment that invites analysts to tailor the analytical approach to the requirements of the problem and to the design of the available data. Second, many of the parts of a reasonably complicated problem require techniques or models that have not yet found their way into the standard elementary and intermediate textbooks of survey analysis and statistical methods. Under these circumstances, a versatile analyst will have to be in some measure what sociologists, somewhat ambivalently, term a "methodologist." That is, she or he will need to participate in the adaptation or implementation of available methods to the problems at hand and not be satisfied to rely on a cookbook approach, even if sanctified by its presentation in the manual of a computer package. Only by becoming a competent methodologist can an investigator prevent his or her problem from being dominated by methodological prejudices—obsolete ones at that (see Recommendation 14, Chapter 10).

To support these hortatory injunctions with instances from actual research is not easy, given limitations of space and the reader's patience. Any single research study will, of course, have the defects of its qualities and may be unconvincing for that reason. Nevertheless, with whatever disclaimers may be needed, we submit a summary of one piece of research in hopes that it may be suggestive of what can be done by an investigator willing to take advantage of the recent developments in methods of analyzing survey data.

In a study of the nuclear controversy Duncan (1978a) was interested in making use of a special feature of the 1975 Harris survey that served as the source of data. Among other questions in the survey, there was a set of nine purported advantages of nuclear power. Respondents were asked whether they did or did not "think so" when presented with a statement of each of these advantages. In addition to the main sample of U.S. adults, Harris obtained responses from 47 officials of nuclear regulatory agencies and 52 leaders of environmental groups. In his analysis of the data, Duncan reported that the mean number of advantages recognized by the general public was 5.23, fairly close to the 5.96 claimed by regulators but not as close to the 3.47 listed by environmental leaders. When each question was scored by "percent think so," the rank ordering of scores given by the public correlated .72 with the rank ordering by regulators and .66 with the rank ordering by environmentalists. (The correlation was .66 between the rankings of the two leadership groups.) Information of this kind gives a certain texture to the question of how opinion is divided on a controversial issue. The public in general tends to be closer to the position of the officials who work with the nuclear power industry, and this is reflected not only in the foregoing data but also in responses to general questions on the advisability of continuing nuclear development in the future.

But we can go beyond this kind of summary statement. Each supposed advantage of nuclear power was evaluated as to how controversial it is by computing the cross-product ratio, $p_R(1-p_E)/(1-p_R)p_E$, where p_R is the proportion of the regulators who "think so" and p_E is the proportion of environmentalists who do. Thus, the statement "is a reliable form of energy for the U.S. to depend on in the long run" is quite controversial, with a cross-product ratio of 16.0, while "will be a cheaper form of energy in the future than it is now" is not very controversial, with a cross-product ratio of only 1.4. The other seven statements fall between these two ratios, but all the statements have ratios in excess of 1, since the environmentalists were consistently more skeptical than the regulators.

These cross-product ratios were used to characterize the items (that is, the nine statements about nuclear power), and further analysis was carried out to investigate whether the controversiality of the statements affected responses to them. For each statement, a 3-way table was prepared using the statement of a supposed advantage of nuclear power cross-classified by the response to the question, "How much confidence do you have in the people running environmentalist groups?" and by the response to a similarly worded question on confidence in the Atomic Energy Commission. Nine distinct 3-way analyses were carried out, one for each advantage of nuclear power.

The entire set of analyses could be summarized in the following manner: for the two most controversial statements, each confidence question was strongly related to response (in the direction one would expect if confidence

in leaders is positively associated with agreement with their positions), and the effects of the two confidence questions taken together were additive. However, for the three least controversial statements, the effects of neither of the confidence questions were significant. The remaining five statements, with intermediate levels of controversiality, showed mixed results, including significant 3-way interactions, that preclude clear-cut statements of the effects of confidence in leaders on endorsement of the statements.

The analysis just summarized is a complex one. The ultimate finding takes the form of a significant 4-way interaction among the variables: response to a statement concerning an advantage of nuclear power, controversiality of that statement (as independently measured in data from elite samples), confidence in leaders of the environmental movement, and confidence in the major government agency leading the development of nuclear power. The gist of the result is that public opinion tends to reflect the positions taken by leadership groups *when those positions lead to controversy that clarifies issues.* The research design is a somewhat intricate combination of item-centered and respondent-centered analyses. The finding is one that has a somewhat more dynamic flavor than such simple, although undeniably important, findings as that women are less favorable to the development of nuclear power than men.

6.4 The Latent Trait Approach in Survey Research: The Rasch Measurement Model

There are two ingrained impressions of survey researchers that argue for a latent trait approach to measurement. First, they are aware of considerable instability in individual response: the same question asked twice of each respondent will receive different answers from many respondents even when there is no reason to suspect that any appreciable fraction of the population is changing its attitudes. Second, it often seems possible to ask what appears to be the same question in several different ways or to ask two or more questions that seem to be getting at the "same" attitude, even though they do not have identical univariate response distributions. In such a case, as is noted in Section 5.1, the investigator may hope to observe "form-resistant correlations" of responses with other variables.

The first impression is consistent with the postulate that response is probabilistic rather than deterministic. The second impression can be formalized by a postulate stating that questions (items) are not themselves the variables of interest; they are clues to or indicators of some variable or trait that cannot be directly observed. A third motivation for some kind of latent

trait approach is not so firmly grounded in experience or data analysis but has an almost irresistible conceptual appeal. It is the supposition that an attitude (say), whatever its qualitative texture, can be thought of, in quantitative terms, as varying along a single dimension from pro to con, unfavorable to favorable, or low to high approval of some object or issue position.

Whatever the appeal of these ideas considered singly, whether they make sense in combination can best be judged by considering a model that somehow brings out their joint implications. Of such a model the question asked would not be, "Is it true?" but, "Is it instructive?" and (if so), "Is it applicable or can it usefully be made applicable under interesting circumstances?" What is "interesting" is a matter of theory (or proto-theoretical discussion) or of application. If applicability of a model is demonstrated, the demonstration becomes a challenge to theory. But the model itself is not a theory. It is a recipe or set of criteria—proposed conditions that must be met for an investigator to support the claim that something has actually been measured. A theory stipulates the conditions under which the model is expected to apply.

While we do not have space to review the work of Rasch and his followers in psychometrics, something of the import of the work is conveyed in the following quotation from Loevinger (1965:151, cited by Wright, "Foreword" to Rasch, 1980):

> Rasch (1960) has devised a truly new approach to psychometric problems. . . . He makes use of none of the classical psychometrics, but rather applies algebra anew to a probabilistic model. The probability that a person will answer an item correctly is assumed to be the product of an ability parameter pertaining only to the person and a difficulty parameter pertaining only to the item. Beyond specifying one person as the standard of ability or one item as the standard of difficulty, the ability assigned to an individual is independent of that of other members of the group and of the particular items with which he is tested; similarly for the item difficulty. . . . Indeed, these two properties were once suggested as criteria for absolute scaling (Loevinger, 1947); at that time proposed schemes for absolute scaling had not been shown to satisfy the criteria, nor does Guttman scaling do so. Thus, Rasch must be credited with an outstanding contribution to one of the two central psychometric problems, the achievement of nonarbitrary measures. . . . When his model fits, the results are independent of the sample of persons and of the particular items within some broad limits. Within these limits, generality is, one might say, complete.

The reader interested in measurement theory is referred to the demonstration by Perline, Wright, and Wainer (1979) that "Rasch's psychometric model [is] a special case of additive conjoint measurement."[12] Tjur (1980) has shown that the psychometric work is in some respects convergent

with recent developments in statistics concerning log-linear models for categorical data (see texts by Fienberg, 1980; Goodman, 1978; Haberman, 1978–79). We cannot review those statistical methods here; readers may wish to consult the cited texts. In this illustration (and in Duncan, Chapter 12, Volume 2), we undertake to express the Rasch model in a form that leads to easy testing and estimation using the procedures discussed in the cited literature on log-linear models.[13]

In our discussion we do not attempt a general exposition of latent trait models nor do we speculate on the kinds of theoretical development that might ensue if any of them found significant application in survey work. The discussion is illustrative, and we consider only the very simplest of the models proposed by Rasch for item analysis of ability tests.

In the 1977 General Social Survey there were two questions on women's working:

[Q. 25] Do you approve or disapprove of a married women earning money in business or industry if she has a husband capable of supporting her?

[Q. 30] If there is a limited number of jobs, do you approve or disapprove of a married woman holding a job in business or industry when her husband is able to support her?

For Q. 25, the results were approve, 65.4 percent; disapprove, 33.2 percent; don't know, 1.4 percent (N = 1,528 with 2 "no answers" omitted). For Q. 30, the results were approve, 35.6; disapprove, 62.4 percent; don't know, 2.0 percent (N = 1,526 with 4 "no answers" omitted).

Between the two questions there were a question on willingness to vote for a woman for president, a four-part question concerning effects on a family of the mother's working, a three-part question on the Equal Rights Amendment, and a question on whether it is all right for the wife to refuse to have children when the husband wants children.[14]

When results of this kind are quoted, they are often discussed in terms of the lability or modifiability of response or in terms of bias in the wording of questions. As to the former, it is clear that if approval drops from 65 percent to 36 percent, quite a few respondents must have "changed their opinion" and changed, moreover, within the short interval of four questions, perhaps 10 minutes or less of interviewing time. Equally disturbing is the fact that an unscrupulous user of survey data could choose whether to quote the results for Q. 25 or those for Q. 30, depending on whether the aim was to portray majority approval or majority disapproval for women working. But even an ethical analyst would have difficulty in deciding which of the questions is unbiased. To be sure, Q. 30 introduces what may look like an extraneous issue—whether the number of jobs is limited. But, surely, if

there were not *some* conditions under which people believe women should not work, the question of disapproval is hardly interesting anyway. Some critics might feel that whether a women has a husband is extraneous. But this kind of inconclusive discussion is just what we would like to avoid.

Let us look at the situation in a different manner. If one were trying to assess the "jumping ability" of athletes, one might set the bar of a high jump very low. In that event, most athletes would clear it, but occasionally even a good athlete might not, owing to "bad luck" or inattention to the task. If one set the bar rather high, many fewer would clear it, but even a fairly poor jumper might do so occasionally by extraordinary effort. The "test" of jumping ability, with the different bar settings as "items," is "answered" probabilistically. But the low settings, the intermediate ones, and the high one are all measuring "jumping ability" (Campbell, 1963).

Now consider again the measurement of "attitudes toward women working." Those attitudes are not directly observable, any more than is "jumping ability." But we can put "tests" to the respondent. If we attach negative qualifications to an item (such as "if jobs are scarce"), we make it "harder" to approve the activity described in the item (women working). If we drop the negative qualifications or stipulate some positive condition or consequence, we make it "easier" to approve the activity. A respondent who approves when there is a negative qualification must have a more positive attitude toward women working than a respondent who disapproves when there is no such qualification. Both items are measuring attitude toward women working; the statement of qualifications or conditons is a device (perhaps not the best one, but not an absurd one) for raising or lowering the bar. In the following paragraph we formalize this idea and state additional postulates that permit one to deduce consequences from it.

Consider dichotomous items A, B, C, . . . intended to serve as indirect measures of an unobserved or latent variable x. Let λ_{Ai} be the odds on a positive response to item A for the ith respondent, and let the odds for the other items be similarly denoted. The odds is defined as $P_{Ai}/(1 - P_{Ai})$, where P_{Ai} is the probability of a favorable response; hence

$$P_{Ai} = \lambda_{Ai}/(1 + \lambda_{Ai}).$$

The basic postulate of the model is

$$\lambda_{Ai} = ax_i$$

where a is a parameter for item A, which may be termed the "easiness" of A (following Rasch, 1966:91), and x_i is the value of the latent variable x for the ith respondent. It represents what we want to mean by the degree of "favorability" of her or his attitude. Item parameters a, b, c, . . . are assumed to be invariant across respondents, and x_i is a property or parameter

of the individual respondent that does not depend on which question is asked. Finally, we adopt the postulate of local independence: *for each individual respondent*, responses to A, B, C, . . . are independent. (Of course, we also assume, as in the commonly used statistical models for survey data, which neglect departures from simple random sampling, that responses of different respondents are independent.) Thus, when we consider the joint distribution of responses to two or more items, we can use the multiplication rule for independent events.

In the present illustration, we have just two items, A (Q. 25) and B (Q. 30). The multiplicative model for the odds $\lambda_{Ai} = ax_i$, and $\lambda_{Bi} = bx_i$, translates into the logistic model for the probabilities (substituting in the expression stating the probability as a function of the odds):

$$P_{Ai} = \frac{ax_i}{1 + ax_i}, \quad 1 - P_{Ai} = \frac{1}{1 + ax_i}$$

and

$$P_{Bi} = \frac{bx_i}{1 + bx_i}, \quad 1 - P_{Bi} = \frac{1}{1 + bx_i}.$$

Hence, applying the multiplication rule to obtain the joint probabilities, the probability of two positive responses is

$$P_{11i} = P_{11i}^{AB} = \frac{abx_i^2}{(1 + ax_i)(1 + bx_i)}.$$

For the probability of a favorable response to A and unfavorable to B, we obtain

$$P_{10i} = \frac{ax_i}{(1 + ax_i)(1 + bx_i)},$$

and for the opposite case,

$$P_{01i} = \frac{bx_i}{(1 + ax_i)(1 + bx_i)}.$$

And, finally, for the probability of two unfavorable responses

$$P_{00i} = \frac{1}{(1 + ax_i)(1 + bx_i)}.$$

Note that the sum of the four numerators is:

$$1 + ax_i + bx_i + abx_i^2 = (1 + ax_i)(1 + bx_i),$$

which is the denominator of each fraction. Hence the sum of the P's is 1.0, as it must be.

Now, P_{11i} is the probability of a favorable response to both items for the ith respondent. It can also be regarded as the expected frequency of such a joint response in a sample consisting of one respondent. To find the expected frequency of this particular response pattern in a sample of n respondents we can add the P_{11i} over the index i, $1 \leq i \leq n$. The same operation for each of the four possible response patterns yields the following set of expected counts:

$$F_{11} = ab \sum_{i=1}^{n} \frac{x_i^2}{\Delta_i} \qquad F_{10} = a \sum_{i=1}^{n} \frac{x_i}{\Delta_i}$$

$$F_{01} = b \sum_{i=1}^{n} \frac{x_i}{\Delta_i} \qquad F_{00} = \sum_{i=1}^{n} \frac{1}{\Delta_i}$$

where, for brevity, $\Delta_i = (1 + ax_i)(1 + bx_i)$.

Suppose we had a sample cross-classification of items A and B for a population in which the Rasch model was known to apply. We could regard the sample counts as estimates of the foregoing expected counts and, in particular, take $\hat{F}_{10} = f_{10}$ and $\hat{F}_{01} = f_{01}$, where f is an observed sample count and $\hat{F}$ is a maximum-likelihood estimate of the expected count.[15] Consider, for illustration, the data in Table 6.2. Before proceeding, we combine "disapprove" and "don't know," without formal justification. (The decision is, of course, debatable, though perhaps not of great moment in the present example. We do note, as a matter of interest, that only two respondents were "don't know" on both questions; this may be a warning that univariate "don't know" proportions are overestimates of "nonattitude" holders.) We then have $\hat{F}_{10} = 519 = 502 + 17$ and $\hat{F}_{01} = 65 = 58 + 7$. From the formulas we see that $F_{10}/F_{01} = a/b$, inasmuch as the terms involving the summations are identical. Therefore, an estimate of the ratio a/b is 519/65 or 7.98. In other words, item A is about 8 times as easy as item B. Such ratios are meaningful for the Rasch model.

We have several comments about this elementary exercise, which is patterned after the discussion by Rasch (1980:171–172). First, we note that the

estimate of relative easiness of the two items does not depend upon the numbers answering both questions "approve" and both "disapprove." This is reasonable, inasmuch as people answering the questions in the same way do not give us any information about which of the items is the more attractive. Second, we see that *any* collection of respondents from a population in which the model holds—provided only that it includes respondents giving both of the relevant patterns and is not selected on the basis of the responses as such—constitutes a sample for the purpose of estimating a/b. The model is "sample-free" in this specific sense. (Of course, we do not offer this comment as a justification for relaxing the usual standards of survey sampling. But it could be a very useful property of the model when comparisons of specially defined subpopulations are of interest.) Third, the calculation we have made depends on the assumption that the Rasch model holds in the population sampled by the General Social Survey. The data in Table 6.2 do not provide a test of that assumption, but a test can be carried out with the survey data (see below).

In Table 6.3 we consider the cross-classification of the two responses for each of several strata of the GSS sample. In conformity with the second comment in the preceding paragraph, we could use (for example) any one of the four sex-color strata as the basis for estimating a/b. Of course, even if the model holds in all four strata and the item parameters are constant across strata, we will not get the same estimates from all strata or estimates of equal precision. But we are, nevertheless, estimating the same ratio. There is an important consequence of this circumstance.[16] The hypothesis of independence of row and column classification must hold for the 4×2 table whose four rows are the four sex-color strata and whose two columns are the two middle columns of Table 6.3, that is, the two response patterns approve/disapprove and disapprove/approve. Carrying out the usual test of this hypothesis, we find for these data L^2 (likelihood-ratio chi-square statistic) = 9.82, d.f. = 3. The hypothesis of independence is rejected at the .05

TABLE 6.2
Cross-classification of Responses to Two Questions in 1977 General Social Survey

	B [Q. 30]			
A [Q. 25]	**Approve**	**Disapprove**	**Don't Know**	**Total**
Approve	478	502	17	997
Disapprove	58	437	11	506
Don't know	7	13	2	22
Total	543	952	30	1,525

SOURCE: 1980 Codebook, General Social Survey (No. 147 and No. 193).

TABLE 6.3

Cross-classification of Responses to Two Questions in 1977 General Social Survey, for Strata Defined by Sex and Color and by Sex and Marital Status

Stratum		Response Pattern			
	A [Q. 25] B [Q. 30]	Approve Approve	Approve Disapprove[a]	Disapprove[a] Approve	Disapprove[a] Disapprove[a]
Sex	*Color*				
Female	Black	36	32	12	27
	Other	206	262	30	230
Male	Black	24	18	2	25
	Other	212	207	21	181
Color and Sex	*Marital Status*				
Other female	Married	140	167	17	139
	Widowed	17	30	3	59
	Divorced	16	23	3	16
	Separated	7	8	2	5
	Never married	26	34	5	11
Other male	Married	137	155	10	127
	Widowed	5	10	1	15
	Divorced	10	10	1	10
	Separated	5	4	2	7
	Never married	55	28	7	22

SOURCE: See Table 6.2.
[a]"Disapprove" includes "Don't know."

level; and, since the Rasch model implies that hypothesis, the model is also rejected. Further study of the 4 × 2 table, however, turns up an interesting result: with black females excluded, the hypothesis of independence for the remaining three rows yields $L^2 = 0.17$; and when black women are compared with aggregate of the remaining groups, $L^2 = 9.65$, d.f. = 1. These results are consistent with the supposition that the model holds in the population exclusive of black females or that it holds both for that population and for black females but with different item parameters in the two groups. (Of course, we must bear in mind the fact that the test of the hypothesis with black women excluded was suggested by the data.)

This revised conclusion need not be taken too seriously. The data are consistent with it, but in no sense require it. The main general point to be made is that a test of the model of this type will have no great statistical

power (that is, the ability to reject the model when it is false) unless the strata are somewhat strongly related to the latent variable. (This point does not seem to be emphasized in the literature on the Rasch model.) For the data at hand, it turns out that, within the white population, sex is not at all strongly related to either item. Hence, there is little realistic probability of rejecting the model once black females are excluded. The exercise is not, however, in vain, for we have learned something about the limited applicability of the model, although the data do not really disclose the reason for the deviant pattern for black women. The result itself, however, would hardly be surprising to an analyst acquainted with actual labor force participation patterns in view of the high participation rate of black women.

In the lower portion of Table 6.3, we consider a different stratification of the sample, the sex-by-marital status classification of white persons (including a few respondents classified as "other" than black or white). Proceeding as before to consider only the two middle columns, we entertain the hypothesis of independence for the 10×2 table and obtain $L^2 = 10.21$, d.f. $= 9$. The model is not called into question by these results. For the hypothesis of independence of the two response patterns and sex within marital status categories, we have $L^2 = 2.35$, d.f. 5; and for the hypothesis of independence of the two response patterns and marital status within sex, we have $L^2 = 10.05$, d.f. $= 8$. Again, the results are consistent with acceptance of the model within the "white" population. As far as the test involving sex is concerned, the result is of little interest for the reason already mentioned. But marital status does have nonnegligible associations with the two responses. Hence the test involving marital status does supply some support for the model.

More stringent tests are available when the stratifications in Table 6.4 are used. From the upper part of the table we extract the 14×2 table pertaining to 14 age classes and the two response patterns, approve/disapprove versus disapprove/approve. For that part of the table, $L^2 = 12.58$, d.f. $= 13$, for the usual hypothesis of independence. Moreover, for the uniform association model (Goodman, 1979b), we find $L^2 = 10.91$, d.f. $= 12$. The difference, $L^2 = 12.58 - 10.91 = 1.67$, d.f. $= 1$, tests the hypothesis of no linear trend in the log of the ratio a/b across age classes. That hypothesis cannot be rejected. This result is quite useful because of the low power of the conventional test of the hypothesis of independence in a table with 13 degrees of freedom.[17] But in comparing the two hypotheses, independence versus uniform association, we consider quite a sharp alternative to independence. It is also relevant to observe that age is not a negligible factor in response. If we use all columns in the upper part of Table 6.4, we have a 3-way ($14 \times 2 \times 2$) contingency table. The hypothesis that the two responses taken jointly are independent of age is decisively rejected; $L^2 = 185.56$, d.f. $= 39$. Further tests make it clear that age is

TABLE 6.4

Cross-classification of Responses to Two Questions in 1977 General Social Survey, for Strata Defined by Age and by Years of Schooling

Stratum	A [Q. 25] B [Q. 30]	Response Pattern: Approve Approve	Approve Disapprove[a]	Disapprove[a] Approve	Disapprove[a] Disapprove[a]
Age (years)					
18–22		50	46	11	21
23–27		84	59	7	30
28–32		67	47	10	19
33–37		67	60	6	35
38–42		38	52	3	47
43–47		31	37	3	39
48–52		37	40	5	41
53–57		31	57	4	44
58–62		27	44	8	38
63–67		19	33	4	40
68–72		8	22	2	43
73–77		12	14	1	38
78–82		3	3	0	17
83 and over		1	3	1	9
Years of schooling					
0–2		1	2	0	1
3		2	2	1	6
4		2	8	0	10
5		1	5	0	18
6		5	18	0	13
7		5	13	1	29
8		24	27	6	62
9		12	24	6	30
10		22	33	5	46
11		30	46	8	30
12		153	188	20	145
13		44	27	6	13
14		36	38	4	24
15		23	12	2	7
16		65	45	1	17
17		18	10	2	2
18		19	11	1	1
19		6	4	0	1
20		9	2	2	2

SOURCE: See Table 6.2.

[a]"Disapprove" includes "Don't know."

significantly related to each of the items considered one at a time. Hence, the success of the Rasch model in the previously mentioned tests is genuine support for the model and not a trivial consequence of using an irrelevant stratification of the sample.

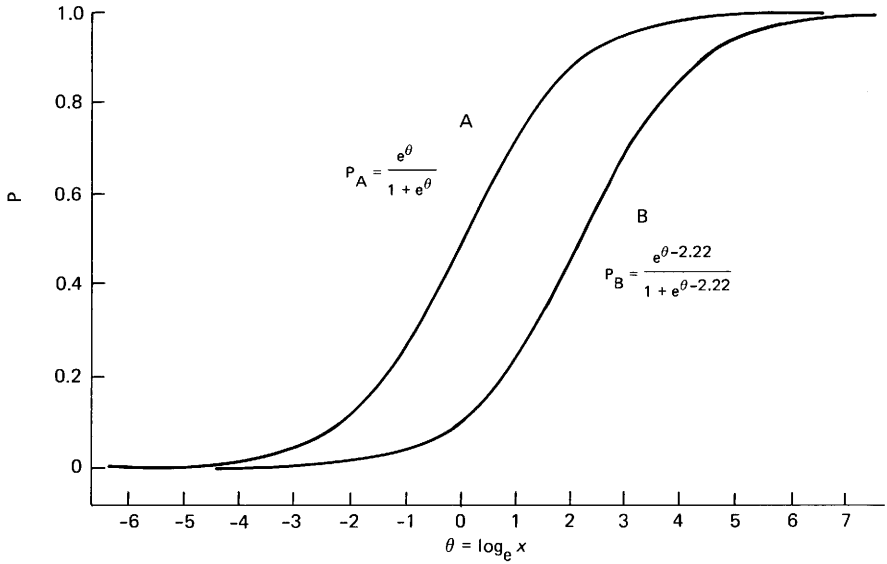
Results to the same effect are obtained from the lower part of Table 6.4. Here, for the 19 × 2 table, education by the two middle columns, we have $L^2 = 25.96$, d.f. = 18, p = .1, for the hypothesis of independence, and $L^2 = 25.95$, d.f. = 17, for the uniform association model. Again, analysis of the entire table, education by the two responses, makes it clear that education can be regarded as an important factor affecting response so that the support for the Rasch model is not negligible.

In summary, our evidence indicates that the Rasch model holds within the population exclusive of black women and in particular for each stratum defined by marital status, by age, or by education. (Strictly speaking, we should have eliminated black females from the analyses for age and education, but this would not have changed the nature of the outcome.) But if the model holds for black females (on which issue the data examined do not bear), the item parameters are not the same for them as for the other groups. Or, we might say, the items have a different meaning for black females. To return to the problem of estimating a/b in the light of these results, we delete black females and find that there remain 487 disapprove/approve and 53 approve/disapprove responses, so that a/b is estimated at 487/53, or 9.19. In the subpopulation for which the model is presumed to hold, item A is about nine times as "easy" as item B. The natural logarithm of 9.19 is 2.22. Hence, in Figure 6.3, the so-called item characteristic curves (which graphically depict the gist of the model) for A and B are separated by a horizontal distance of 2.22 on the scale of log x. The location of the zero on this scale is arbitrary; here the scale is chosen so that at zero the probability of a favorable response on item A is 0.5.

This is a convenient point to insert a comment concerning the relationship of the Rasch model to Guttman scaling. In the case of the Guttman scale, each item characteristic curve will take the form of a stair step. There is a critical or threshold value of x for each item such that $P_{Ai} = 0$ if $x_i < x_A$, $P_{Ai} = 1.0$ if $x_i \geq x_A$, and so on for items A, B, C. . . . The scale model thus provides an ordering of items and a weak ordering of respondents. But it does not assign parameter values to items or to respondents. It is curious that the stochastic model of Rasch, which might be said to involve "weaker" assumptions than Guttman uses, actually leads to a stronger measurement model. Suppose the results of our empirical exercise are taken at face value, and we conclude that A and B are both measures of attitude toward women working. What have we gained? One thing is that we now have two items instead of one to work with. Item A would be especially suitable for studying subpopulations in which attitude is relatively unfavorable and item B for

FIGURE 6.3

Item characteristic curves for items A and B: Probability of favorable response as a function of latent variable, x.



subpopulations in which attitude is relatively favorable. For a subpopulation with a wide dispersion of attitude, either item could be used in the ordinary fashion of survey analysis to investigate determinants of attitude. But there is a second, somewhat surprising implication of the model that comes to light when there are two (or more) items.

Consider expressions such as those given earlier for expected frequencies, but with a classification by stratum made explicit. Now x_{ij} is the value of the latent attitude variable for the ith individual in the jth stratum. For a compact notation we define

$$\Delta_{ij} = (1 + ax_{ij})(1 + bx_{ij})$$

and the stratum parameters,

$$S_{tj} = \sum_{i=1}^{n_j} (x_{ij}^t / \Delta_{ij})$$

where t = number of favorable responses to the two items, $0 \leqslant t \leqslant 2$, and n_j is the number of respondents in the jth stratum. With these definitions,

$F_{h_A h_B j}$ is the expected count in the jth stratum of respondents giving response pattern $h_A h_B$, where $h = 1$ for a favorable response (approve) and $h = 0$ for an unfavorable response (disapprove). Hence we have

$$F_{11j} = ab\ S_{2j} \qquad F_{01j} = b\ S_{1j}$$
$$F_{10j} = a\ S_{1j} \qquad F_{00j} = S_{0j}$$

where $1 \leqslant j \leqslant J$ if there are J strata altogether. From the definitions we see that the S-parameters are complex mixtures of item parameters (a and b) and respondent (subject) parameters x_{ij}. They have no substantively useful interpretation, but they are complex functions of the "composition" of the strata and, therefore, will not be the same from stratum to stratum. (The term "composition" is used in its demographic sense to refer to the stratum distribution of values of x.) Let us now inquire, what is the association of items A and B within a stratum? As in any log-linear model, we define association in terms of the cross-product ratio, $F_{11j}F_{00j}/F_{10j}F_{01j}$. Substituting for the Fs their definitions in terms of parameters, we have $abS_{2j}S_{0j}/aS_{1j}bS_{1j} = S_{2j}S_{0j}/S_{1j}^2$. Thus, association of items within strata is entirely due to heterogeneity of the respondents within each stratum with respect to the respondent parameter x_{ij} in the following sense. Suppose there was no such heterogeneity, or that $x_{ij} = k_j$ for all members of the jth stratum. In that event, we would have $S_{2j}S_{0j}/S_{1j}^2 = 1.0$: that is, there would be no association of A and B within the stratum. If such a situation obtained for all J strata, we might say that our sample was classified by "ulstrata" (ultimate strata).[18] By definition, when the sample is stratified into ulstrata, we have completely "explained" the association of A and B in the aggregate of all strata or of the whole sample. Of course, if x_i is continuous, so that there are presumably no ties, we will not achieve such a stratification short of treating each individual as a distinct stratum. But it is not absurd to suppose that we might, on occasion, achieve an approximation to such a stratification short of this limiting case. Let us explore the implications of this possibility.

The usual goal of survey analysis is to explain attitudes by means of personal characteristics, membership in groups and social relationships, exposure to persuasive arguments, and so on. Unlike the situation in ability testing, the precise estimation of each individual's latent trait parameter is not a primary objective of the survey interview. Nor is such estimation necessary for the investigation of the determinants of attitude if the Rasch model applies and if we have just a few items that effectively span the range of variation in attitude in our sample. As we have just seen, when the strata defined in an analysis approximate ulstrata, the association between any two items will tend to vanish.

In most survey analysis one is not inclined to carry stratification ("partialling") very far for fear of small cell counts. But the hazard here is greatly

mitigated by a theorem stated by Haberman and reemphasized by Fienberg. In comparing two models for a set of sample counts that differ only with regard to one or more parameters included in the more complex model but excluded from the less complex one, "differences in likelihood-ratio chi-square statistics have approximate chi-square distributions under quite general conditions" (Haberman, 1979:421), and the "conditions" include those of "large sparse multinomial structures" (Fienberg, 1979:63).

An investigator might be faced with a substantial association (AB) between two items, computed from the sample as a whole. In seeking to move toward a set of ulstrata the investigator could consider models incorporating the interactions $(AZ_1Z_2Z_3 \ldots)$, $(BZ_1Z_2Z_3 \ldots)$, or sets of lower-order interactions approximating them, such as (AZ_1Z_2), (AZ_1Z_3), (AZ_2Z_3), $(AZ_1 \ldots)$, $(AZ_2 \ldots)$, $(AZ_3 \ldots)$, $(A \ldots)$, and similar interactions involving B rather than A. (These are the interactions that are defined in terms of 2-way and multi-way marginal distributions of variables in log-linear models.) Call this Model M_1. Model M_2 is the same as M_1 except that it also includes a parameter for the (net) (AB) association. The difference between likelihood-ratio chi-square statistics, $L^2(M_1) - L^2(M_2)$, with d.f. = 1, tests whether the (AB) association has been accounted for by the Zs. Note that because we assume a Rasch model (after appropriate tests of the assumption), it is not necessary to investigate interactions involving Zs along with both A and B, such as (ABZ_1) or (ABZ_2Z_3). Such apparent associations are due to respondent heterogeneity within strata formed by combinations of values of the Zs. And we need no redundant indication of such residual heterogeneity since the chi-square test just mentioned is precisely directed to this issue. From the standpoint of computation, a program such as ECTA (Goodman, 1978:468) would seem to be well suited to the handling of the "large sparse multinomial" table to be analyzed. The properties of the Rasch model, therefore, appear to offer something that has hitherto been lacking in survey analysis, an indication of when the analysis is "complete." There is no guarantee, of course, that a "complete" analysis is feasible in terms of the variables that may be available in a particular survey interview. If it is not, we will be in the inconclusive situation of almost all survey analysis today. Having carried stratification of the sample as far as we can or care to, we will nevertheless rest uneasy that all our supposed "causal" coefficients could be drastically changed if the sample were stratified by just one more (unknown) variable. That would not happen, of course, if we had already achieved a classification by ulstrata.

The preceding ideas concerning completeness and ulstrata are offered quite tentatively. Successful empirical survey analysis employing these ideas remains to be demonstrated. But no other part of this section depends on these ideas.

Additional illustrative data are presented in Duncan, Chapter 12,

Volume 2, in which further algebraic properties of the model are developed in the situation in which there are more than two items. We conclude this section with nine statements concerning implications. Those statements that take the form of theorems are presented without proof, although proofs are available in the literature or can be constructed using the kind of algebraic technique already illustrated.

First, the separability of item and subject parameters that is achieved with the Rasch model is a most attractive property of the model. Indeed, it is a little difficult to say in what sense *measurement* is achieved if that property is violated in an arbitrary way. If the yardstick is stretched or compressed in an unknown way as we go from one respondent to another, it will not be of much use in measuring heights. But it is not only true that the model has this property, it is the only model of its general type that does so (see Rasch, 1966:105). For example, if the logistic item characteristic curve of Figure 6.3 is replaced by the graphically very similar normal ogive, separability no longer obtains (Rasch, 1980:122). (Of course, a contrary conclusion can be defended by rejecting Rasch's criterion of separability.)

Second, as we have seen, separability of item and respondent parameters at the micro level does not produce separability of item and stratum parameters at the macro level. On the contrary, as long as the strata are heterogeneous with respect to values of the latent attitude variable, there will be statistical associations and interactions the coefficients of which mix up in a very messy way the item parameters and the composition of the stratum with respect to x. Discussions (for example, Cronbach, 1975) stressing the ubiquity of interactions might well be reconsidered in this light, for the empirical basis of them may well be research in which measurement in the rigorous sense demanded by Rasch was not achieved or the stratification of respondents did not go far enough to reduce to a negligible quantity the intrastratum variation in x_{ij}.

Third, if we have T items of the kind here discussed, it seems natural to compute a total score, t, $0 \leqslant t \leqslant T$, as the number of items answered in the favorable direction. Indeed, t_i is known to be a sufficient statistic for the estimation of x_i under the model. However, the regression of t_i on x_i is not linear. Hence, the application of linear models to the explanation of t_i does not lead in quite the direction one might think. In particular, moreover, the R^2 statistic for the proportion of variation in t_i explained by regressors Z_1, Z_2, Z_3, . . . (whatever the functional form of the regression) is very misleading as a measure of success in explaining x_i. Even when explanation is rendered virtually "complete" by the production of a set of ulstrata, this R^2 can be far below unity.

Fourth, in investigations with multiple indicators, a common practice has been to inspect relationships of the several indicators, taken one at a time,

to explanatory variables. It would seem that if the indicators were measures of the "same" attitude, they should have the same relationships to other variables. But it is an elementary consequence of the model that this will *not* be true, except in the limiting case in which the explanatory variable actually is tantamount to a set of ulstrata. Thus, the naive search for "form-resistant correlations" is not well advised.

Fifth, the Rasch model resembles the latent-class and latent-distance models of Lazarsfeld and Henry (1968) in that the postulate of local independence plays a key role. However, the models are quite different in that Lazarsfeld places no constraints on the item parameters; these may vary in an arbitrary way from one latent class to another or from one step to the next on the latent-distance scale. Work by Goodman (1978) and Clogg (1979), however, has shown how various constraints may be imposed on the basic latent-structure models. Further work, therefore, may disclose a closer affinity between the Rasch and Lazarsfeld approaches than is apparent at first glance.

Sixth, in the Rasch model there are no "biased" or "unbiased" questions. An item is either one of a set of items that measure the latent attitude or it is not. (With three or more items, indeed, one can test the hypothesis that if one dimension is measured by some pair of items, no more than one dimension is measured by some subset of three or more items, including that pair.) Moreover, the zero point of the scale for item parameters and the zero point of the scale for respondents (both scales being in the same units) are arbitrary, and no claim for the substantive importance of the zero point can be sustained. Still, it is possible to calibrate any item that measures the latent attitude against any other item chosen as a standard. If Rasch measurement is convincingly demonstrated, such calibration will be a useful byproduct of the demonstration. If that occurs for attitudes hitherto investigated by single-question methods characteristic of some polling organizations, we may hope to see somewhat less emphasis on the idea of the poll as mini-referendum.

Seventh, the model is subjected to statistical tests. The ethical investigator, of course, will make the tests as powerful as possible within the limits of resources for research. Model failures need not always be regarded as disastrous. One form of failure can result if the items have different meanings in different subpopulations (as, perhaps, the questions on women's work may have different meanings for black women than for others). Hence, it will always be an empirical problem to delineate the population within which the model applies. Similar remarks apply to the problem of items whose meanings change with time, an aspect of the problem described by Cronbach (1975:122) as "the times as a source of interactions." In principle, it should actually be possible to separate items whose meanings have altered from those with stable meanings, provided there are at least

two of the latter. With volatile material, a program of nearly continuous calibration and recalibration will be needed, as is the case with such changeable natural phenomena as geologic faults or volcanic activity.

Eighth, the model considered here is only the simplest one of a class that has the property of separability of parameters. Models involving more than two response categories have been studied. And models relaxing in a particular way the postulate of independence between item responses within respondents have been proposed. For example, Kempf, Hampapa, and Mach (1975) described and illustrated with survey data a "dynamic test model" that features "*transfer parameters* that describe how an individual's responses are influenced by his answers to the preceding items." (For further discussion see the symposium edited by Kempf and Repp [1977] and the review of it by Miller and Wasserman [1979].) In many cases, however, in place of trying to model a complex situation the investigator will be well advised to redesign the instrument. Rasch (1966:107) warns:

> *Observations may easily be made in such a way that specific objectivity* [i.e., separability of item and subject parameters], *otherwise available, gets lost.* For instance, this can easily happen when qualitative observations with, say, five categories of responses for convenience are condensed into three categories. If the basic model holds for the five categories, it is mathematically almost impossible for the three-category model also to hold. . . . I must point out that the problem of the relation of data to models is not only one of trying to fit data to an adequately chosen model from our inventory to see whether it works; it is also *how to make observations in such a way that specific objectivity obtains.*

Ninth, if one is convinced that demonstrated applicability of Rasch measurement in survey research would be of substantive and theoretical interest, what should be done? As a matter of economy, one might first try relevant models for some of the numerous multi-item data sets in extant survey data archives. It may well be that satisfactory fits will not often be obtained since these batteries of items were constructed without benefit of any indoctrination in a rigorous theory of measurement. In that event, effort should go into diagnosing reasons for model failure. For example, using statistical methods for log-linear models one could try including in a model the interactions in addition to those called for on the assumption that the items conform to the Rasch model. Detection of such interactions might afford clues to the redesign of survey instruments. On the pessimistic premise that Rasch models will not readily be turned up, investigators should begin long-term projects for developing appropriate instruments. No doubt in their early stages such projects will involve much trial and error. Despite the heretical character of the advice, however, one could suggest that time and other resources could be conserved by doing developmental studies on readily accessible populations, reserving the expensive

machinery of the national probability sample for final demonstration and calibration. If, on the other hand, the developmental work is successful, attention might turn in part to an effort to construct attitude theories that stipulate the conditions of measurement. At intermediate stages of research, all that we know concerning social action and human response is potentially useful for the construction of instruments and the explanation of why instruments do or do not work. Finally, *pari passu* with the foregoing, studies of the sources of variation in item parameters should be vigorously pursued (see Section 6.1), and classic survey analysis oriented to explanation of variation in attitudes of individuals should also be carried on, with appropriate modifications of traditional practice in the light of the observations offered earlier in this section.

Notes

1. Among the topics *not* covered in this chapter are classical survey analysis (see, for example, Selvin, 1968; Rosenberg, 1968); factor analysis (see, for example, J. Kruskal, 1978; Harman, 1976); recent extensions of Guttman scaling (Shye, 1978); multidimensional scaling, clustering, and similarity analysis (see, for example, Carroll and Kruskal, 1978; Romney, Shepard, and Nerlove, 1972); structural equation models (see, for example, Goldberger and Duncan, 1973; Jöreskog and Sörbom, 1979); and hierarchical log-linear models (see, for example, Goodman, 1978:Chapters 1, 3, 4). The very scope of these topics precludes a detailed review within the confines of this report.

Readers should also note that in this chapter probability statements are included that assume that the data were derived from simple random samples; this is not the case in most of the examples in this volume. While the probability statements must, therefore, be considered approximate, we are following the practice of statisticians who have worked with these same (and similar) data. Haberman (1979, Vol. 2: 487), for example, prefaces his analysis of data from the NORC General Social Survey (whose data are used frequently throughout this volume) by noting:

> Assume that N [observations] are independently distributed . . . . This assumption is approximate given the complex sampling method used, but is probably adequate for most analysis.

Hopefully current and future work on the (multivariate) analysis of data from complex sampling designs (for example, Fay, 1979, 1982; Fellegi, 1980; Rao and Scott, 1981) will provide more precise treatments of such data in the future.

2. The Rasch model has been developed in the field of mental testing. In view of the reservations of some panel members about the applicability and utility of the model, we are making no affirmative recommendation on its behalf other than that it warrants further investigation, perhaps in the directions suggested at the end of the section.

3. The latter parameter takes the form of an unobserved score on a latent trait that the item is supposedly measuring.

4. We hasten to disclaim any endorsement of such analysis that would involve taking the answers at face value. There may be artifactual reasons for the apparent outcome for "crime" vis-à-vis "foreign aid". Item-centered analysis must be at least as circumspect as respondent-centered analysis.

5. Kohn's data actually show that all five of his social classes evaluated "considerate of others" more highly than "neat and clean"; however, that kind of comparison was not of immediate interest. For his purposes, it was more relevant to observe that evaluation of "neat and clean" was negatively related to social class while evaluation of "considerate of others" was positively related to social class.

6. The reader may note that a seriousness score of 10 for stealing a bicycle is far off this regression on any reasonable assumption as to the value of a bicycle. This anomaly is not explained in the initial progress report on this study, which is the source quoted here.

7. Rosenberg (1968:20), in his felicitous exposition of survey analysis, explicitly narrows his focus to this case: "While symmetrical, reciprocal, and asymmetrical relationships all have theoretical significance, the process of elaboration, which is the focus of this book, is undertaken only with *asymmetrical relationships.*" Quite a bit more could be said about the statistical side of this problem than Rosenberg did, inasmuch as methods explicitly designed for categorical data (which most survey data are) have been developed and diffused largely during the period since Rosenberg wrote. (For reviews of this literature, see, for example, Duncan, 1978b, 1982, Swafford, 1980.)

8. The questions read (National Opinion Research Center, 1978:131–32): "Please tell me whether or not *you* think it should be possible for a pregnant woman to obtain a *legal* abortion if . . .

A. If there is a strong chance of a serious defect in the baby?
B. If she is married and does not want any more children?
C. If the woman's health is seriously endangered by the pregnancy?
D. If the family has a very low income and cannot afford any more children?
E. If she became pregnant as a result of rape?
F. If she is not married and does not want to marry the man?"

9. In this discussion we consider only models with a finite number of discrete latent classes, whether ordered or not. Latent structure models with "continuous latent space," as presented by Lazarsfeld and Henry (1968), have been of interest primarily in the psychometric setting. (The Rasch model is a model of this class.)

10. The two questions have rather different marginal distributions; this is a fact of considerable potential interest, but not the matter on which the present discussion focuses.

11. Actually, the second question was worded somewhat differently in the 1959 survey and in half of the 1971 interviews; the investigators were interested among other things, in detecting an effect of the wording change, a goal facilitated by the explicit parameterization of the pattern of association.

12. A variety of applications of Rasch measurement in ability testing are summarized in Benjamin Wright's "Afterword" to Rasch (1980), and the literature cited there makes it clear that Rasch's approach is proving to be a seminal one, even though there was a lag in its development owing, perhaps, to the inaccessibility of Rasch's original publications (see also

Weiss and Davison, 1981). There is also an active group of European social psychologists who have exploited the approach productively (see Kempf and Repp, 1977).

13. The estimates produced by this approach are conditional maximum likelihood estimates of the Rasch model item parameters.

14. The data and question wordings are taken from the 1980 cumulative codebook of the General Social Survey, in which the two questions are listed as No. 147 and No. 193, respectively.

15. A formal justification for the use of these estimates is based on a conditioning argument that focuses on the off-diagonal cells, in a fashion similar to the approach used in McNemar's test for symmetry in a 2 × 2 table. Indeed, McNemar's test is the appropriate test for the hypothesis H_0: $a/b = 1$ versus the alternative H_A: $a/b \neq 1$.

16. Using the conditioning argument referred to in note 15.

17. The uniform association model is not being proposed as a generalization of the Rasch model but is considered for heuristic reasons because it is a sharply defined alternative to the hypothesis of independence.

18. The neologism was suggested by "ulstrith," a term coined for a somewhat different situation by Toops (1948).

PART IV

THE SURVEY INTERVIEW PROCESS

. . . one has only learnt to get the better of words
For the thing one no longer has to say, or the way in which
One is no longer disposed to say it. And so each venture
Is a new beginning, a raid on the inarticulate
With shabby equipment always deteriorating
In the general mess of imprecision of feeling. . . .

T.S. ELIOT, East Coker

Primary responsibility for original drafting of this part of the panel's report was assumed by Abelson (9.3), Fischhoff (7.2), MacKuen (7.1), Martin (8.1–8.5, 9.1, 9.2, 9.4), and Nerlove (7.3). See Note on Authorship, page vi.

In this part, we deal with the interview as a psychosocial process, using ideas and techniques from several social science disciplines in order to understand sources of ambiguity in survey measurements of subjective phenomena. Most of these ideas and techniques have not been systematically tested in the survey context, and this part of the book is therefore more speculative than the first three parts. Our aim is to lay some of the necessary groundwork for a more integrated theoretical approach to the survey measurement process.

Survey measurement is an attempt to quantify a theoretical or social construct (Y) by means of a set of procedures (X) to produce a measurement ($\hat{Y}$). The measurement, $\hat{Y}$, resulting from the application of X may be erroneous or inconsistent for several different reasons. Each of the following three chapters concerns a different type of uncertainty in survey measurements of subjective phenomena.

Chapter 7 is concerned with problems of conceptual ambiguity. The construct, Y, may itself be poorly defined or poorly understood so that the resulting measurement is of uncertain meaning and validity. Conceptual ambiguity can exist on (at least) two levels: the theoretical constructs can be poorly defined so that there is no way of knowing whether $\hat{Y}$ corresponds to Y; or real-world concepts can be subject to uncertain and varying interpretations. Two examples of conceptual ambiguity are presented in Chapter 7 and some techniques for exploring and resolving disparities between an investigator's and respondents' conceptual frames of reference are offered.

A second type of measurement ambiguity occurs when the measurement procedure measures not only Y, the construct of interest, but also extraneous variables, Z_i. That is, $\hat{Y}$ varies not only with Y, but also with Z_i. This problem is particularly severe when the extraneous variables include factors that vary systematically among respondents, among interviewers, or among interview settings. For example, we believe that survey questions about racial attitudes measure not only true racial attitudes, but also the respondent's willingness to speak frankly about racial matters to the interviewer, and the latter varies systematically with the interviewer's race. One important source of extraneous influences on survey measurements is the social setting in which the interview occurs, and in Chapter 8 we discuss various factors that may influence the way respondents define their role vis-à-vis an interviewer and a survey.

A third source of ambiguity in survey measurement is that the process is psychologically intrusive. A respondent is required to interpret the interviewer's question and then construct and articulate some reaction to it (even if only to say "I don't know"). Thus, every response has a synthetic quality; it is not exactly the same as the subjective state that is the real object of measurement. In order to assess the validity of a respondent's

report of her or his subjective state (attitude, belief, and so forth), we require an understanding of the mental processes involved in producing verbal responses to survey questions. In Chapter 9 we consider some of the psychological dynamics involved in interpreting survey questions and formulating answers to them.

7

Conceptual Ambiguity in Surveys

Concepts used in social science theories have meaning not only in theory, but also in the real world. The theoretically appropriate definition of a social construct—such as public opinion, or risk, or unemployment—thus depends not only on scientific and measurement criteria, but also on how the concept is understood and used in the world.

Lay and scientific (or technical) concepts do not necessarily match. For example, lay definitions of unemployment probably do not correspond exactly to the definition that is used in the Current Population Survey. When there is disjuncture between concepts as they are used by survey researchers and as they are used by the public, the published results and public debates about them may be misleading. In addition, if the concepts used in survey questions are not understood in the same way by the survey researcher and the respondent, then responses to the questions are likely to be misinterpreted by the researcher. A survey researcher, then, must be attentive to lay concepts and classifications and the relation of those to scientific concepts. (By contrast, physicists need not be concerned with how physical concepts—such as the quark—are understood by nonscientists, and, since they do not rely upon self-reports as a source of scientific data,

they certainly need not worry that quarks may not define themselves as physicists do.)

The fact that many constructs in social science have social as well as scientific meaning creates several pitfalls for researchers. One such pitfall is a temptation to rely on implicit understandings of the construct, without invoking independent normative or scientific criteria to define it. An example of this is the fundamental concept of public opinion, which is discussed in Section 7.1.

Even when a researcher's measurements are grounded in carefully defined theoretical constructs, however, another pitfall may exist if respondents do not share the researcher's conceptual framework. This has proved to be the case in attempts to understand people's attitudes concerning risk (see Section 7.2). Most surveys of public attitudes regarding policy issues assume that the issue, as presented by policy makers, is well formulated. But when an issue, and the language in which it is expressed, are rapidly evolving, policy makers may be framing issues in what is actually an incoherent or illogical fashion or using concepts without clear meaning.

One way of exploring differences between a researcher's and respondents' conceptual frames of reference is an approach known as ethnographic semantics (discussed in Section 7.3). Anthropologists and psycholinguists have developed procedures to enable an investigator to determine the verbal coding schemes in which people understand and interpret their reality. The procedures can be used to conduct more rigorous and systematic pretesting for surveys than is presently done; they can also be incorporated as an integral phase of the survey itself.

7.1 The Concept of Public Opinion

The single area of scientific inquiry that has been aided most by the use of survey-based measures is the study of public opinion. Although by the 1930s a great deal of creative and primarily speculative work on public opinion had already been undertaken, classical researchers were stymied because the data needed for further research seemed unobtainable: public opinion resided inside the heads of far too many people. But with the development and popularization of survey sampling and interviewing methods during the 1930s (see Converse, Chapter 1, Volume 2, and Converse, 1981), students of public opinion were able for the first time to test their theories empirically. The use of the new techniques enabled and encouraged a creative outburst of exploration. With the tremendous progress, however, has come a certain insensitivity to the limits of the popular tech-

nique. In particular, the research community, in exploiting new avenues of data collection, bypassed some of the concerns that occupied classical public opinion theorists and instead became preoccupied with questions that apparently could be answered by the use of surveys solely.

There are those who maintain that in studying survey responses they are, by definition, studying public opinion. But in the classical view, public opinion does not consist of an aggregate of individual opinions. Ideas that have not been formed and tested through discourse and public debate, for example, would not be deemed worthy of being called public opinion. With the advent of the survey method, the classical view was often lost because survey researchers implicitly equated public opinion with whatever public opinion polls measure. So, for example, Gollin (1980:448) notes that " 'Public opinion,' whatever that phrase once meant, now is taken by most people most of the time to mean poll findings." Accepting such a position quickly settles the matter of the utility of surveys in public opinion research. However, as Blumer (1948) emphatically pointed out more than a quarter century ago, such a stance provides only a semantic escape from what is meant to be a serious question: What is public opinion? In order to use an instrument to examine an object, he pointed out, investigators must have an idea of what they are trying to get at in order to be able to understand the importance of what they see. Without an external framework for analysis, survey researchers cannot be sure they study public opinion, only that they study public opinion polls.

All this would be relatively unimportant, a squabble among academicians, were it not for the fact that the concept of public opinion plays a very large and legitimated role in the social and political life of (at least) industrial democracies. Having the unambiguous backing of public opinion gives an unmistakable boost to the political fortunes of any program. In fact, one major definition of democracy identifies it with a government that effectively reflects public opinion. The ability to define public opinion to give substance to this claim is thus of real consequence. Hence, it is useful to understand the origin of the concept of public opinion.

The idea of public opinion originated during the seventeenth and eighteenth centuries as rulers found themselves relying on the support of the rising bourgeoisie. Without a political organization to give it form, that support became centered around the ideas generated for discussion and evaluation in the coffeehouses and salons of European cities. For the first time in centuries, social critics found an audience independent of the ruling hierarchy, a clientele for whom the rule of ideas, of rational debate, became the idealized (though not always the effective) grounds of legitimacy. In that age of enlightenment, it was the seemingly rational character of the process of idea testing through social intercourse that gave promise to the ideal of democracy, to the rule of public opinion. Yet it is not at all clear that the

proponents of the eighteenth and nineteenth century liberal revolution had the views of ordinary citizens in mind when they argued for the central role for public opinion. Indeed, the traditional understanding of popular opinions, that is, of the views found in the streets, was that they owed more to imagination and passion and less to rational thought and knowledge (Speier, 1949) and hence were not what was meant by public opinion.

Today, the policy question is not what portion of the citizenry has a legitimate claim in the formation of public opinion. Rather, the question is what qualities distinguish citizen views that ought to be listened to from those that ought to be ignored by the political system. For those working in the classical normative tradition, research centers around the extent to which a particular expression of public opinion reflects thought and knowledge rather than imagination and emotion. The task then is not merely to organize one's observations in order to develop an adequate description of society, but also to interpret those observations in the light of a normative meaning. It is because the users of survey data, social scientists or poll readers, carry certain normative views of public opinion that they must be careful about judging the quality of the information about the citizenry which surveys provide.

Much of contemporary research into the nature of public opinion has been concerned with ideas more or less derived from alternative interpretations of the classical concept. The research agenda of scientific work on public opinion includes (but is not limited to) four general areas: the motivational basis for attitudes; the interpersonal communication structure within which attitudes are set; the centrality of any particular attitude in a personal and cultural context; and the implications and locus for action that can be associated with any view. Each of these may be considered a separate research domain in its own right, but each also has used survey research technology and, importantly, other measurement tools to enrich investigators' understandings of how the attitudes explicitly measured by surveys might reflect various ideas of what constitutes public opinion.

MOTIVATIONAL BASE

It seems clear that individuals form and hold attitudes (including cognitive and affective elements) for various reasons. A number of scholars (for example, Smith, Bruner, and White, 1956; Katz, 1960; Kelman, 1961) have emphasized the importance of motivations for understanding in what ways people hold and change their attitudes. Broadly speaking, people adopt attitudes in order to adjust to the social and political world by understanding and evaluating its elements or in order to work out internal psychological dynamics such as ego-defensive or self-expressive mechanisms (in Lasswell's [1930:173] famous phrase, "displacing private affects upon political objects"). Democratic theorists clearly had the first set of motives in

mind when they argued for governmental sensitivity to public opinion; yet we know that a part of what is measured in surveys derives from the second. The question of the relative role of these dynamics is central to studying public opinion, but it is clear that most surveys are not particularly well suited to address the question.

Intensive work exploring the relationship between opinions and personality (Smith, Bruner, and White, 1956; Lane, 1962) has suggested that the dynamic psychological processes involved are complex and difficult to generalize. A great mass of subtle information is needed for case-by-case analysis, information that is not easily unearthed by common survey practice. This is not to say that the interview format cannot be designed to obtain useful data; only that an investigator would have to go to extraordinary lengths to find out about each respondent's constellation of psychological forces. Such intensive encounters are rarely made because of the expense in time and effort (as well as the seemingly limited generalizability), yet the few results available clearly indicate that apparently public opinions often include a generous portion of private feelings. Thus, merely assuming that the responses to an ordinary survey are directly related to the public domain is empirically unjustified, and political inferences about how a policy might be changed to meet public opinion might be misleading. If the root of an individual's dissatisfaction with government lies within his or her inner personal world, then no change in public policy can alter the dissatisfaction—it can only deflect it.

INTERPERSONAL COMMUNICATION STRUCTURE

By far the favorite prescription of the classical public opinion theorists is that opinion ought to rule only to the extent to which it has been sufficiently considered in the public domain. While the specifics of the discussion process might vary, the general idea lies in recreating for citizens the experience of the eighteenth century coffeehouses and salons. Only after an idea has circulated, only after its adherents have confronted their opponents in rational debate, is it to be considered sovereign. In this sense, people's views that are derived from isolated and passive reception, as in a mass society whose citizens receive information from only one central source, are not worthy of the term public opinion. Also less worthy in the classical view are ideas generated by communication with others that allows only positive feedback, say the product of individuals in a crowd milling about strengthening each other's ideas.

To the extent that this quality of public opinion—its root in social communications—is the focus of public opinion research, then most contemporary sample surveys, taken by themselves, will be inadequate tools. Most survey analysts treat respondents as atomistic individuals, the elements of a mass, and neglect the social groupings which would provide the information

needed to pursue the matter of communication networks.[1] However, some current research emphasizes the importance of further examining the character of social communication (for one discussion, see Noelle-Neumann, 1979), a process that will clearly call for a reintroduction of expensive social network sampling. If surveys do not actually trace out these patterns of interaction, then the researcher will understand neither the dynamics of opinion formation and change processes nor the quality of interaction which takes place.

The importance of the communication process in the formation of opinions was clearly suggested some time ago by Lord Bryce (1899:253):

> In examining the process by which opinion is formed, we cannot fail to note how small a part of the view which the average man entertains when he goes to vote is really of his own making. His original impression was faint and perhaps shapeless: its present definiteness and strength are mainly due to what he has heard and read. He has been told what to think and why to think it. Arguments have been supplied to him from without, and controversy has imbedded them in his mind. Although he supposes his view to be his own, he holds it rather because his acquaintances, his newspapers, his party leaders all hold it. His acquaintances do the like. Each man believes and repeats certain phrases, because he thinks that everybody else on his own side believes them, and of what each believes only a small part is his own original impression, the far larger part being the result of the commingling and mutual action and reaction of the impressions of a multitude of individuals, in which the element of pure personal conviction, based on individual thinking, is but small.

Bryce worked from personal observation. In recent years, surveys have been used in conjunction with measures of media and environmental context in order to assess the applicability of his observation in a modern context. The linkage between the social environment and individual attitudes proved more complex than earlier thought (for a review of early work, see Klapper, 1960), but it does exist (see review by Kraus and Davis, 1976).

A growing field of research examines the ways that individuals form opinions in the context of a broader communication system (see Beniger, Chapter 15; MacKuen, Chapter 14; Martin, Chapter 17, Volume 2). To the extent that the quality of an individual's views are derived from outside sources such as the news media or conversations with others, then the public opinion analyst can turn to the objective records of communication to get at something that would otherwise be hidden—the quality of thought supporting those views. For example, if a respondent indicates a desire for a tough foreign policy, the researcher does not know if the respondent is motivated by a belief that such a policy would better serve the national interest or by personal feelings projected on a public policy question. One

clue would be the nature of the public debate that has colored the respondent's feelings: whether it is couched in terms of global strategy, anticipated reactions of other nations or national resources, or in personalized terms such as shame, weakness, or vengeance. In this sense, it is the context in which an opinion is set that defines its meaning, and thus the context must be understood to grasp the meaning. (Caution, however, must be exercised in assuming a direct link between the tenor of public debate and the quality of public attitudes—initial empirical results on the linkage have been mixed [see Neuman, 1979].) It is also true that, to the extent that an individual's views are embedded in a larger interplay of opinions, the usual atomistic survey will only measure a segment of the social situation. Yet it is the entire process of opinion formation that is of interest, and it would make no more sense to measure public opinion by analyzing the press's output or political leaders' actions by themselves.

A sense of the interplay among all the elements (given that they are not homogeneous forces), an understanding of the range and strength of action and reaction, will increase our understanding of public opinion. Doing so certainly requires the ability to measure attitudes held by citizens at any given moment—which surveys do best—but it also demands much more extensive measurement of the content of other aspects of public opinion. In this sense, the object of attention is not an individual's opinion or set of opinions but the system that generates them.

PERSONAL AND CULTURAL CONTEXT

Less often remarked on than the social context of a person's opinions but also of great interest in studying public opinion is the value and belief context within which people hold their opinions. The importance of any opinion can be said to depend on its centrality within a framework of other elements, either within an individual's inner world or, more significantly, within the society's cultural background. Ideas that appeal to citizens' fancies, due to either their rational or their emotional characteristics, will eventually be retained only if they can coexist with the weight of opinions already in stock. Without a wealth of consistent and associated beliefs and values, any attitude is likely (though not certain) to weaken.

The interview format of ordinary surveys can certainly be used to analyze the psychological structure of attitudes and beliefs. In fact, consideration of the coherence of attitude structures has been an area of extensive research (see, for example, Converse, 1964; Nie, Verba, and Petrocik, 1979; and Smith, Chapter 8 in Volume 2). While a good deal has been learned by evaluating citizen responses to externally defined frames of reference, this area of inquiry might benefit from an application of techniques that allow respondents to reveal their world views in their own terms (see Section

7.3). Respondent-centered procedures, adapted for use in surveys, might bridge the gap between intensive but discursive interview sessions and the fixed format schedules commonly used.

Of greater importance for society than individual attitude structures is the relation of people's beliefs with one another and with a shared cultural heritage. In terms of the classical view of public opinion, the fact that an attitude has been adopted by some numerical majority does not imply that it represents public opinion. As Lowell (1913:4–5) pointed out:

> If two highwaymen meet a belated traveller on a dark road and propose to relieve him of his watch and wallet, it would clearly be an abuse of terms to say that in the assemblage on that lonely spot there was a public opinion in favor of a redistribution of property. Nor would it make any difference, for this purpose, whether there were two highwaymen and one traveller, or one robber and two victims.

Public opinion cannot be defined simply as the sentiment shared by 51 percent of the population (or any particular majority). Clearly what is missing on the highway is a sense of community with shared goals. Lowell argued that only to the extent that a common basis for mutual interest exists, so as to frame disagreement in a context of greater cooperation, can the opinion of some sizable element begin to command the allegiance of the minority with the force of public opinion. Mutuality of interest and the compatibility of alternative viewpoints could certainly be assessed with survey interviews (Hovland, Harvey, and Sherif, 1957), although in contemporary work respondents are infrequently asked to evaluate their tolerance of other positions (see Stouffer, 1955; Davis, 1975a).

As with individual orientations, the contextual field for opinions in societies can also be applied to a cross-temporal perspective. Here the matter of cultural beliefs and values comes into play, with the clear understanding that any opinion handed down over the years represents the fundamental baseline against which current fashions of thought must be measured. A cultural heritage may be complex and subtle, but any set of beliefs fundamentally at odds with it will find itself poorly based. For example, an economic program designed for aggressive and optimistic entrepreneurs will have difficulty in parts of the world where traditions of communal sharing and of resistance to innovation prevail. Of course new ideas can and do change cultural views, but only by studying such ideas within their cultural context can their importance be discerned. The problem for a researcher is to identify what is culturally meaningful and what is not. Over time, sample surveys may provide useful information about cultural norms and their stability, and when contrasted with expectations, may reveal otherwise hidden cultural values (see Martin, Chapter 17, Volume 2).

CONSEQUENCES FOR SUBSEQUENT ACTION

During the Enlightenment, the idea of public opinion commanded notice because it was meant to represent or imply latent public action. In particular, rulers were concerned about the political views on which the emerging bourgeoisie would base its lending decisions. While public opinion has long been treated as a norm guiding elite behavior, it may also represent a threat of action to enforce the public's intentions. Thus, from a classical viewpoint the extent to which given views are related to future political action provides crucial depth to the outlines of public desires.

The action context of any attitude involves two distinct elements: whether the attitude implies a subsequent action to be taken by the individual or is only a passive judgment about the world; and whether such action is to be taken as a private individual or as a member of society. These two elements combined with the element of motivation (previously discussed) leads to four states: private feelings, private actions, public feelings, and public actions. It is the last state—attitudes that suggest activity in the public sphere—that most clearly matches the classical idea of public opinion.

Certainly many opinions held by the people in a society are not related in any way to anticipated action; they are merely reflections of the way in which their understandings of ongoing events match their preferences. People may have strong opinions about the value of a balanced budget or of a prohibition on abortion without giving any thought to the possibility that those views ought to shape life decisions. In fact, given the relative distance of much of what is termed public affairs from the daily life of most people, many of the public's views about community life have a fantasy air completely removed from any responsibility for personal action.

On the other hand, people who are sufficiently moved by their attitudes to take action may do it as private individuals or they may try to affect a decision by the society. It is clear that many people consider matters primarily in terms of individual consequence and act so as to advance their own personal interests, while others perceive them as parts of the collective arena and work along more generalized political avenues (Verba and Nie, 1972; Schuman and Presser, 1979). Concern about pollution may lead individuals to leave urban environments, to write letters complaining about a particular nuisance, or to form an organization to lobby for new laws. Clearly these actions are different in meaning, and only by understanding respondents' dispositions to various future actions can public sentiments be interpreted as reflecting a form of active public opinion.

Surveys can begin to assess a respondent's inclination to take subsequent action dependent on his views, and the arena of that action. Reports of past activity may help order the likelihood that a change in attitudes will lead to

a change in society's behavior, either as an aggregate of individual actions or as a collective action. A change in the assessment of the seriousness of an energy crisis by individuals oriented to private action may result in a change in gasoline sales, while that change among citizens oriented to public action may result in the formation of a new interest group.

Finally, it is worth noting that when surveys have been used to measure current respondent activity, they have shown that only a small segment of the population takes a meaningful part in any policy debate. The channels through which public views are transmitted to policy makers attract so small a portion of the general population that those participating in the forming of effective public opinion cannot readily be represented in an ordinary sampling frame; trying to measure those activists requires specialized samples. Furthermore, special care must be taken in interviewing techniques with elite samples (Dexter, 1970) as the respondents are likely to know more about the subject matter than the investigator and hence be put off by overly structured, simplistic, or tedious probes (see, for example, Lang, 1978, 1981). Nevertheless, the basic technology of the sample survey seems applicable here, though such specialized samplings of public opinion are infrequently taken and even less frequently interpreted in a broader context.

IMPLICATIONS

Much of what is reported in public opinion surveys (1) represents a projection of respondents' inner psychic needs and has little to do with public affairs; (2) is derived directly from external sources and has not been subjected to serious personal discussion or critical evaluation; (3) represents views that are out of character with the individual's or society's basic world view and thus ephemeral in nature; and (4) is entirely removed from consequential action so that individuals are unlikely to translate their views into any action, particularly public action. In part, what is measured in public opinion polls is nothing more than the public's fantasy life, and so serious political response based on it would be a caricature of democracy and ought to be avoided. But a part of what is portrayed does reflect public-minded, well-considered, thoroughly consistent, and consequential attitudes. Clearly these are the kinds of attitudes that carry the widely accepted force of public opinion, and these are the views that ought to be emphasized in polls bearing that name.

Thus, it is incumbent on the analysts and reporters of surveys to make clear to readers the extent to which the views reported correspond to a careful definition of public opinion. In order to do so, they must go beyond simply asking respondents their preferences in a set of multiple-choice

questions and summing up the answers. To do nothing more is irresponsible. An analyst must begin to examine the meaning context within which individual views are held. For some analyses this will require conducting much more intensive interviews (perhaps with small numbers or groups) in order to discover the deeper meanings not easily elicited by structured questionnaires. (See Section 7.3 for some techniques for exploring systems of meaning.) In other cases, it will be necessary to analyze the content of formal communication (including but not limited to the elite press) in order to understand the character of the public debate and thus the environment surrounding individual expressions of opinion (for illustrative analyses, see Beniger, Chapter 15; MacKuen, Chapter 14, Volume 2). An obligation to provide contextual meaning requires nothing technically new from an opinion analyst—all of the necessary techniques are currently part of some practice in different projects. If the public opinion researchers, analysts, and reporters take their work seriously and are willing to go beyond providing entertainment for readers or propaganda for political interests, they must integrate explicitly, additional sources of information into a fuller, textured, portrait of public opinion. Unless they are willing to provide readers with the contextual feel necessary for interpretation they ought not report their results as public opinion at all.

7.2 Using Surveys to Clarify Concepts: The Example of Risk

Most surveys of public attitudes regarding policy issues assume that those issues, as presented by policy makers, are well formulated. The survey researcher's task is merely to put those questions into good English, avoiding overly complex or ambiguous terms, and try not to prejudice or pressure responses in any direction. Although this assumption may be fairly reasonable when one is dealing with familiar, articulated issues, it may not be at all reasonable when policies, and the vernacular language used to express them, are rapidly evolving. In such cases, policy makers may be framing issues in an incoherent, inconsistent, or even illogical fashion, or they may be using terms that do not have clear meanings.

In such situations the role of a survey in relation to policy may take on a somewhat different cast. Instead of just telling policy makers how to resolve issues or even how people feel about them, surveys may suggest how to formulate them, indicating what options to offer, what terms to use, and what interpretations to make. An indirect consequence of reliance on sur-

veys has often been to restrict communication between policy makers and the public to those opinions that can be expressed within the constraints of the particular questions posed by an impassive interviewer, able to say or hear little about how questions might be reformulated or reinterpreted. Several iterations of a survey might be needed to develop an issue and survey instrument to the point at which respondents can really express their attitudes. Both policy makers and the public would then have had a hand in shaping what an issue is all about and what possible resolutions should be considered.

Allowing for two-way communication may suggest alternative interpretations of some seemingly aberrant response patterns. Rather than representing apathy or ignorance, refusals to respond may reflect antipathy to the way a question is worded, frustration at the narrowness of the set of options proferred, or inability to express a complex opinion. Rather than representing confusion, apparently inconsistent responses to a set of questions on an issue may mean that the questions themselves did not adequately explicate respondents' points of view or that the questions were liable to conflicting interpretations. Rather than representing recalcitrance or contentiousness, mistrust of surveys may be due to a feeling that they are cutting off debate on an issue or unduly restricting it.

Several examples of the interaction between social theory and survey data can be seen in recent research into how people perceive the risks of hazardous technologies and how they evaluate the acceptability of those risks. This research initially adopted the formulations of risk problems that are dominant in the technical and policy-making communities, and in the conventional survey tradition, it explored the clearest way to pose questions and the easiest way to elicit responses. Then, following the newer or interactive tradition, it allowed respondents to reveal subtleties or ambiguities of the questions asked. In this way, the term "risk" was discovered to have different interpretations in the technical and lay communities studied (Slovic, Fischhoff, and Lichtenstein, 1979). Although each group can understand the other's definitions, they tend to use the terms differently. Clarification of such semantic differences is obviously necessary for communication between the groups whenever the meanings differ greatly (for example, when a technology has high risk from one perspective and low risk from another).

However, more than just semantics is involved when the parties disagree on what definition of risk should be used in policy making. That question is one for the political process to decide, not survey researchers. The role of the social scientist is to clarify the sources, extent, and implications of such disagreements. The hoped-for result of such research is improving communication between the parties in risk debates, developing better measure-

ment procedures for telling policy makers what the public wants, and deepening the debate by clarifying its formulation. Such an endeavor begins by assuming that the people involved do not know exactly what they are talking about, in the sense of understanding the full meaning of the frames of reference that they are or could be using. The following three examples elaborate on the problems and the difficulties of achieving conceptual clarity in debates about risk. (Fischhoff, Slovic, and Lichtenstein [1980, 1982], Slovic, Fischhoff, and Lichtenstein [1979, 1980] serve as general sources for the following discussion.)

ACCEPTABLE RISK

Many debates turn on whether the risk associated with a particular configuration of a technology is acceptable. Research has found substantial disagreements not only between people belonging to different population groups, but also within groups when the question is posed in different ways. Although these disagreements may be interpreted as reflecting conflicting social values or confused individual values, closer examination suggests that the acceptable-risk question itself may be poorly formulated (Fischhoff et al., 1981).

To be precise, one does not accept risks. One accepts options that entail some level of risk among their consequences. When a decision-making process is considering the benefits or the other (nonrisk) costs of several options, the most acceptable one need not be the one with least risk: one might choose (or accept) the option with the highest risk if it had enough compensating benefits. The attractiveness of an option depends on its full set of relevant positive and negative consequences.

In this light, the term "acceptable risk" is ill-defined: acceptable compared with what? Unless the options and consequences to be considered are specified, the term has little meaning. Once they are specified, acceptable risk might be used to denote the risk associated with the most acceptable alternative. When using the term, it may be difficult to remember how context dependent it is. That is, people may disagree about the acceptability of risks not only because they disagree on how to evaluate the consequences (that is, they have different values), but also because they disagree about what consequences and options are to be considered.

It is possible that such disagreement about what the problem is may underlie conflicts about the acceptability of different energy systems. The source of those conflicts is often said to be differing perceptions of risks and benefits. However, even two people whose perceptions agree may disagree about the acceptability of a system if one person is considering its attractiveness in relation to other energy production options while the other person

includes conservation programs among the alternatives. Or they may agree about the set of alternative production systems to be considered and their effects, but disagree about whether some effects (for example, shifts in the distribution of knowledge or economic wealth) are valid considerations. A combination of surveys and theoretical investigations of the acceptable-risk concept may help reveal what people mean by acceptable risk and the extent to which disagreements are due to alternative perceptions or alternative frames of reference.

VOLUNTARINESS

One of the most frequently invoked notions in discussions about risk is that more stringent levels of safety are required for involuntarily imposed risks than for voluntarily assumed ones (Lowrance, 1976; Starr, 1969). Hence, for example, skiing could (or should) legitimately be a more hazardous enterprise than living below a major dam.

Although, as intuition would suggest and as initial studies have found, most people agree about the voluntariness of dams and skiing, other technologies are more problematic. Considerable disagreement has been found in people's ratings of the voluntariness of such technologies as prescription antibiotics, automobiles, commercial aviation, hand guns, and home appliances. The disagreements may reflect differences in the reference groups considered: for example, the use of commercial aviation may be voluntary for vacationers, but involuntary for many business people. Or they may reflect disagreements about the nature of society or the meaning of the term: for example, each decision to ride in a car may be voluntarily undertaken and may in principle be foregone, but in a modern industrial society those choices may be somewhat fictitious. And even if one makes a clearly volitional decision, some of the risks that one assumes voluntarily may be indirectly and involuntarily imposed on one's family or the society that must pick up the pieces (for example, paying for hospitalization caused by skiing accidents).

Whenever alternative interpretations of voluntariness are possible, some preferred definition must be chosen in order for voluntariness to be judged. In the work done to date, the definition has been no more extensive than "Do people get into these risky situations voluntarily?" The fact that people often agree upon the ratings of technologies in this light may suggest that there is some consensual interpretation of this question or that different interpretations produce similar judgments for those technologies. In addition to empirical work clarifying what voluntariness means when posed in a general sense, theoretical work is needed to clarify what definition one should use: that is, how sensible is it to incorporate into public policy a double standard for voluntary and involuntary risks, using each possible interpretation?

CATASTROPHE

A somewhat more controversial notion in discussions about risk is putting a premium on avoiding catastrophic losses of life, in the sense of claiming that it is worse to lose N lives at one time than one life at each of N times. The subjective importance of catastrophic potential may be seen in the fact that the difference between lay and technical people's definitions of "risk" mentioned earlier seems to center on how to incorporate catastrophic potential.

Unlike voluntariness, catastrophic potential has been discussed in the risk literature in terms of its appropriateness as a consideration in social policy (Ferreira and Slesin, 1976; Wilson, 1975). One proposal has been offered for treating N lives lost at one time as equivalent to N^2 lives lost in separate events, and another for treating multifatality events no differently until the death toll passes some initial threshold (for example, twenty lives lost) and then giving them some extra weight. Arguments have also been advanced to the effect that society should be indifferent to how lives are lost or that it should only consider catastrophes as special to the extent that the fates of those who died are interrelated. For example, if the vitality or resilience of a community is threatened, more importance may be appropriate than if individuals leave no (grieving) survivors.

As with the study of voluntariness, the study of catastrophic potential would begin with a theoretical examination of the interpretations that are possible and of their reasonableness as bases for social policy. It would continue to the study of what people mean by catastrophic potential and what definition they would adopt were the alternatives laid out for them (along with the policy implications). It would look not only at the term itself, but also at how it is made operational. For example, what constitutes a catastrophic event? Does it include, say, deaths from mesothelioma that occur over a period of years as the result of exposure to asbestos at one time in the past?

Once a precise definition of catastrophic potential has been chosen, one can begin to explore seriously the disagreements about what that potential is. While voluntariness is a qualitative consideration, catastrophic potential should be amenable to quantification and even scientific estimation, although there are limits on such estimation because of the rarity and hypotheticality of such events. For example, the catastrophic potential of any novel technology must be in part a matter of conjecture, disciplined perhaps by theoretical analyses, but still awaiting evidence that can only come with its use. Understanding and eliminating definitional differences is a precondition for focusing debate on this important remaining uncertainty.

Finally, we need to explore whether the immediate loss of life is actually the driving force in people's concerns about catastrophic potential. Some studies suggest that people are alarmed by catastrophic accidents largely because of their informational value. When an accident or even a near-

accident occurs, it may signal to people that the technology involved is more poorly understood and more out of control than had been thought. As such, it serves as a harbinger of future problems, both catastrophic and distributed over time (and space). The direct social consequences of such signal events may be overshadowed by secondary effects, such as the (temporary or permanent) shut down of an industry or by higher-order consequences, such as dependence on more costly and more dangerous alternative technologies. Similarly, voluntariness may also serve as a surrogate for other risk features with which it is associated. For example, involuntary risks may be opposed because they tend to involve an inequitable distribution of risks and benefits, as when involuntary risks are imposed on people other than those who create the risky technologies and enjoy their benefits. Before posing survey questions or interpreting responses in terms of public policy, it is important to know just what it is that people are really concerned about when they express aversion to a particular kind of risk.

IMPLICATIONS

These examples suggest the potential usefulness of intertwined theoretical and empirical studies of risk (or any other subject), clarifying what concepts could, should, and commonly do mean. The results might help both to inform policy makers about what people think and to shape the ways in which policy debates are formulated (and, hence, the questions about which people are asked for their opinions). In this light, a survey becomes an intermediary for two-way communication between policy makers and the public, helping to structure, shape, and illuminate each other's questions and concerns.

In each of the preceding examples, problematic or inconsistent survey results posed a quandary that might be viewed as a choice between rejecting measures and rejecting concepts. They may also afford a somewhat different perspective on the validity of constructs. One of the higher forms of validating empirical measures is often considered to be construct validity (discussed above in Section 4.3), establishing whether the measures within a set are related to one another in a manner prescribed by an integrating theoretical construct. Failure to pass a test of construct validity is usually interpreted as a failure of the measure and its creators. In principle, however, an alternative interpretation is possible: that the concept itself leaves something to be desired. When concepts are abandoned, they usually take their measures with them. In the foregoing examples, however, the measures were allowed to stand while the concepts were called into question. The observed responses were taken to represent perfectly good, even important behavior, but the concepts to which they were bound were considered to be confused. New or more sophisticated concepts were needed, in the context of which the measures could demonstrate construct validity.

7.3 Techniques for Exploring Frames of Reference in Surveys

The presumption that a single frame of reference is shared by investigator and respondent—when it may neither exist nor be attainable by means of the kinds of pilot testing usually done by survey researchers (Belson, 1968) —results in a great deal of ambiguity in survey results. Thus, a respondent may misinterpret the investigator's meaning of a survey question. For example, the 1970 census question on Hispanic origins included the following response categories:

Mexican
Puerto Rican
Cuban
Central or South American
Other Spanish
No, none of these

The category "Central or South American" showed a surprisingly high response frequency; analysis revealed that a number of people from the Mississippi Valley and the Midwest had chosen the category. The question had confused those for whom it was not intended: some non-Hispanic respondents apparently did not think of themselves as belonging to the category "No, none of these" (Fernandez and McKenney, 1980).

A respondent's interpretation of a question is dependent on the set of response alternatives that is presented in the survey. Experimental results (Schuman and Duncan, 1974) demonstrate the importance of knowing the array of elements from which a particular response was chosen. In 1956, respondents in the Detroit Area Study were asked the following question:

> People feel differently about making changes in the way our country is run. In order to keep America great, which of these statements do you think is best? We should rarely, if ever, make changes in the way our country is run; We should be very cautious of making changes . . .; We should feel free to make changes . . .; We must constantly make changes. . . .

The question was replicated in a comparable 1971 survey, and the investigators reasoned that the meaning of the four original response alternatives would be captured by collapsing them into two ("cautious" versus "feel free"). Their hypothesis was tested in a split-ballot experiment, and it was dramatically refuted. Results presented in Table 7.1 show that if the investigators had assumed the equivalence of the two forms of the question, the direction as well as the magnitude of changes in responses from 1956 to

TABLE 7.1
Percentage Distributions of Responses to Question on Willingness to Change How Country Is Run: 1956 and 1971 Detroit Area Study

Response[a]	1956	1971A	1971B
1. Rarely	2.1	1.6	—
2. Cautious	48.9	42.5	63.2
3. Free	36.6	31.7	36.8
4. Constantly	12.4	24.1	—
(N)	(757)	(924)	(927)
(Unclassified)	(40)	(12)	(18)

SOURCE: Schuman and Duncan (1974:241).

[a]Question wording: "People feel differently about making changes in the way our country is run. (Show Card) [In order to keep America great], which of these (four/two) statements do you think is best?"

1. We should rarely, if ever, make changes in the way our country is run.
2. We should be very cautious of making changes in the way our country is run.
3. We should feel free to make changes in the way our country is run.
4. We must constantly make changes in the way our country is run.

(Wording in brackets was used only in 1956. Only statements 2 and 3, renumbered 1 and 2, were supplied in the 1971 form B interview.)

1971 would have been erroneously estimated. In addition, the 1971 two-alternative form suggests that the majority was conservative about making changes, while the four-alternative form suggests that the majority was liberal.

A variety of procedures and techniques for increasing agreement in frames of reference between respondent and investigator have been used or advocated by survey researchers over the years, such as followup interviews (Cantril, et al., 1944); the development of response alternatives based on analysis of verbatim responses to open questions in pilot or pretest studies (Lazarsfeld, 1944; McKennell, 1974; Moser and Kalton, 1972); rigorous pilot test debriefing (Belson, 1968); and the use of a battery of specific questions concluding with a general synthesizing question (Bradburn, 1969). Other researchers have called for the integration of more in-depth field work with survey research (see, for example, Glaser and Strauss, 1967; Myers, 1977; Sieber, 1973; and Stinchcombe, 1964).[2]

Less well-known among survey researchers is an approach to understand-

ing systems of meaning known as ethnographic semantics, which comes primarily from the fields of anthropology and psycholinguistics. With its emphasis on systematic discovery of both appropriate questions and the possible responses of specific populations or groups, it has potential for contributing to the development of questionnaires. Its procedures enable an investigator to elicit the verbal coding schemes in terms of which informants understand and interpret their reality.

Starting with a lexeme, a term of single meaning not predictable from its morphology or syntax, as the basic element of study, it is possible to systematically elicit some meaning systems by inserting the lexeme of interest into frames, such as: "An *x* is a kind of what? What kinds of *x* are there? Are there any kinds of *x* other than *y*?" The forms thus elicited are then grouped into semantic fields representing groups of contrasting terms differing from each other in meaning, but at the same time defining each other in their contrasts. In general, a semantic field can be defined by determining that its terms are closely substitutable in a distributional sense, that is, by measuring the extent to which lexemes can substitute for and replace each other in similar elicitation frames. Techniques for eliciting the meaning systems that are implicit in natural languages are described in Black and Metzger (1965), Conklin (1962, 1964), Frake (1961, 1964), and Metzger and Williams (1962, 1963a, 1963b). Stefflre, Reich, and McClaran-Stefflre (1971) extended these eliciting procedures to describe patterns of multiple meaning of individual words and interrelations among semantic fields. In addition, they provided a quantitative base for describing meaning systems and thus were no longer restricted to working with a few informants; this work made it possible to forge a link between ethnographic semantics and large-scale research.

Mauser (1972) provides an example of an approach that incorporates ethnographic semantic procedures as an integral part of a survey. The study concerned political preferences in California for 1968 presidential candidates (see also Stefflre, 1972, 1979). The goal of the approach is to develop a model of the patterns of competition among candidates or parties in a political contest. The eliciting procedures of ethnographic semantics were used as an aid in the conceptualization of patterns of competition. Underlying Mauser's approach is the hypothesis that political candidates compete with each other to the extent that they are seen by the electorate as resembling each other. If this hypothesis is correct, it should be possible, simply by mapping voters' perceptions, to predict the extent to which any new candidate (or party) who entered the contest would draw votes from any other candidate. Before presenting his findings, we describe in detail the general approach.

The approach consists of a series of four interlocking stages, beginning with relatively intensive questioning of a small number of respondents to

provide the foundation for the study and then proceeding through the other stages to large-scale research:

1. defining the semantic field of candidates;
2. mapping perceptions;
3. mapping preferences; and
4. large-scale survey.

Each of the stages, which progressively increases in sample size, is designed both to provide information for subsequent stages and to empirically verify the findings of previous stages.

The objective of stage 1 is to determine the full range of potential candidates in the election. The success of subsequent stages depends upon having a representative and comprehensive sample of potential candidates. Failure to achieve this objective could result in the investigator's overlooking crucial problems or issues in the election. The two steps in this stage are based on ethnographic semantic procedures. First, from a sample of 10–20 respondents, names of potential candidates and prospective appeals are identified. Then, based on a different sample of 30–50 respondents, judgments about which candidates are seen as making which appeals are collected. To elicit candidate names and appeals, respondents are asked to list the individuals whom they see as likely or desirable candidates for the contest. Then for each of the candidates so named, respondents are asked to describe their anticipated appeals, probable campaign themes, or supporting groups. Typically, it is found that this first step produces 40 to 50 candidates and the same number of appeals. The second step uses statistical clustering and classification techniques to identify classes of candidates with similar appeals (Hartigan, 1975).

Stage 2 examines voters' perceptions of candidates in greater detail. A list of 20–30 of the candidates identified in stage 1 is presented to another sample of 40–60 respondents. A respondent is asked first to judge which candidates are similar to each other and then to identify the features or characteristics shared by candidates who were judged as similar. Similarity judgments are obtained by asking respondents to pick out three or four candidates whom they see as being similar to each of the candidates on the list. The basis for those judgments is obtained by asking respondents to state for each pair of candidates judged as similar why they see them as similar. The similarity judgments are analyzed using multidimensional scaling techniques (Kruskal and Wish, 1978). The multidimensional scaling analysis yields a spatial model of the set of potential candidates, in which the dimensions of the model represent the principal features of the candidates. The dimensions must be interpreted by the investigator, but those interpretations are informed by data from the respondents; content analysis

is used to determine the features elicited from respondents in order to find major classes of perceived features.

The goals of stage 2 are to identify the major classes of candidates, to develop hypotheses about which features govern the position of candidates in the perceptual structure (that is, their location as represented in the spatial model) and to develop hypotheses about which issues will be the most important in determining patterns of political preferences. In order to increase the chances of understanding voter preferences, it is important to generate a wide variety of hypotheses—about 20–40—at this stage. These hypotheses, which will be tested in stage 3, take the form of the construction of "hypothetical candidates" from bundles of features, such as issue positions and personality characteristics, generated and identified as important in this stage and the prediction of the location of such hypothetical candidates in the perceptual structure.

Stage 3 examines voters' preferences for potential candidates by drawing yet another sample, this time of about 200 to 400 respondents. Respondents are asked to rank order the 20–30 candidates scaled in stage 2 plus the 20–40 descriptions of hypothetical candidates developed in that stage. Multidimensional scaling techniques are used to analyze voters' preferences for candidates and appeals in a manner parallel to that followed for the similarity data. The hypotheses generated in stage 2 are tested in stage 3 primarily in two ways. The investigator observes whether the hypothetical candidates are positioned in the perceptual structure where predicted. For example, hypothetical candidates identified by bundles of liberal features should be located in the spatial model near candidates well-known as liberals. And the investigator observes whether the hypothetical candidates draw votes from other potential candidates as predicted. (During the past several years Mauser has found, across a variety of elections in the United States and France, a close relationship between the patterns of voters' perceptions as elicited in stage 2 and their patterns of preferences as elicited in stage 3, with Spearman *rho* correlations ranging from .68 to .87 [Mauser, 1979, 1983].)

Stage 4 consists of a large-scale representative sample of 1,000 to 1,500 respondents. In this stage, mock elections are presented to respondents, including actual and hypothetical candidates, in order to test hypotheses developed from earlier samples about the extent to which candidates draw votes from each other and why. The most important hypothetical candidates are reintroduced at this stage to confirm the results of stage 3 and, in addition, new hypothetical candidates are introduced to clarify the preliminary results of stage 3 concerning which features govern positioning in the perceptual structure.

For 1968, Mauser's study showed that the patterns of vote-splitting for third-party candidates could be predicted from their perceived similarity to

the major candidates in the election. For example, George Wallace was seen by California voters as more similar to Richard Nixon than to Hubert Humphrey and thus, contrary to the intention of the Wallace campaign to split the Democratic vote, Wallace drew more votes from Nixon than from Humphrey. Based on voters' perceptions of candidates, this particular research was able to predict this pattern of vote-splitting more than a year in advance of the election.

The ethnographic semantic approach addresses the same problems of understanding survey research data that Lazarsfeld (1937, 1944) attacked many years earlier, namely, those of determining whether a response has one or multiple meanings, of developing an appropriate set of response alternatives, and of ascertaining the difference in meaning between one response and another. Procedures based on ethnographic semantics add heuristics and algorithms to Lazarsfeld's conceptualizations. The entire process of research involves coordinated investigation of meaning systems at different levels, using the small-scale elicitation strategies of ethnographic semantics to serve large-scale survey research in its aim of predicting choice behavior by a sample of respondents. The solution to the potential problem of a difference between an investigator's and respondents' systems of meaning is to investigate the meaning systems of respondents first and to organize the investigation in order to preserve the properties of the systems that have been discovered.

Notes

1. To be sure, individual respondents can act as informants about their communication roles and habits, and their answers can serve as useful surrogates for observation (Lazarsfeld, Berelson, and Gaudet, 1944; Katz and Lazarsfeld, 1955), but they are surrogates. And while survey sampling frames often consist of clusters of contiguous households, this aspect of survey design is seldom exploited as a source of information about communication networks among neighbors.

2. For a more extreme view advocating the rejection of formalized survey research in favor of an alternate research strategy rooted in ethnography, linguistics, and cognitive science, see Cicourel (1974).

8

The Role of the Respondent

The social context can constrain or facilitate the expression of attitudes and opinions, and therefore the feelings expressed in different social situations are not necessarily consistent. For example the question "How are you?" is likely to elicit different answers when asked by a friend in casual conversation or by a psychiatrist in a therapy session, and appropriately so; the feelings that are real and relevant in the two contexts are different.

The influence of social context on attitude measurement is illustrated by an early analysis of public opinion in a small rural community by Schanck (1932). He found that churchgoers when questioned in public settings almost universally adhered to conservative norms about card playing, but just as universally expressed more liberal views privately. The fact that people were inhibited about expressing their private views to anyone but intimate friends implied that the "entire group [maintained] a public position in contradistinction to the private attitudes of the majority" (Schanck, 1932:101). It also implied that different procedures were required to elicit public versus private attitudes. Schanck used the simple technique of measuring public attitudes in public settings, or suggesting to respondents that their views would be made known to others.

NOTE: Throughout this chapter we follow the example of the Survey Research Center (1976) *Interviewer's Manual* and refer to the interviewer as "she" and the respondent as "he" for the sake of clarity. Actually, a vast majority of interviewers and a slight majority of respondents (in most cross-sectional surveys) are women.

As was true for Schanck in 1932, modern-day survey researchers must take care that the context of measurement matches the real-world context to which survey findings are to be generalized. As Hyman (1949:38) put it, "If our aim is to predict a given kind of behavior in a social setting, we should design our tests so that they incorporate the fundamental aspects of the setting into the test."

In terms of social context, we do not know much about how the role of respondent in a face-to-face interview is similar to or different from other social roles.[1] There has been relatively little systematic analysis of how respondents interpret the survey interview and their role in it. (For useful discussions, see Back and Cross [1982] and Biderman [1975].) The interview setting is not a familiar one for most people, and evidence suggests that respondents do not have a clear understanding of what they are supposed to do. Based on reinterviews with a sample of respondents from a Health Interview Survey, Cannell, Oksenberg, and Converse (1977:308) conclude

> The respondents do not know what is expected of them; they understand neither the general goals of the survey nor specifically what they are supposed to do on particular questions. They tolerate the interview but do not have a clear understanding of their task.

The uncertainty of the situation leaves the respondent free to interpret the meaning of the interview in a variety of ways, some appropriate and some inappropriate. Because of uncertainty, a respondent is likely to be attentive to situational cues as to what behavior is appropriate. This implies that there is no single or standardized respondent role: respondents may adopt a variety of roles depending on various contextual cues in the interview. We consider four factors that, we hypothesize, influence the way a respondent defines his role vis-à-vis the interviewer and the survey.

8.1 The Interviewer's Approach to the Respondent

The relationship between an interviewer and a respondent is initiated when the interviewer requests an interview.[2] The interviewer's request may be granted and the interview successfully completed, or the request may be refused, but in either case the interviewer and the potential respondent continue to interact, perhaps over a series of encounters, until the issue is resolved. In the course of their interaction, the interviewer and the respondent are in effect defining the situation and their relationship.

In this process, the interviewer is in some ways at an advantage. She has a more or less well-formulated plan for carrying out the interview while the respondent did not plan the encounter and probably has no firm ideas about how (or whether) to be a respondent. The interviewer's experience and the respondent's ignorance and confusion make it more likely that she will succeed in imposing appropriate (from her point of view) definitions of the situation and the respondent's role in it. Thus, the Survey Research Center (1976:7, 9) instructs interviewers:

> The most successful interviewer is one who is able to size up the situation quickly on the basis of what little information is available and to act accordingly. Approach each person who answers the door as if he were friendly and interested. . . . If your approach is uncertain or uneasy . . . this feeling will be communicated to the respondent and he will react accordingly.

In other respects, the interviewer is at a disadvantage because in order to complete an interview she must go beyond the bounds of what is in most circumstances considered polite behavior by a stranger. She seeks entry into the privacy of the respondent's home; she wants an hour of the respondent's time (though this request may not be explicit); and she will, if permitted, ask a series of questions that in many contexts are considered personal or rude.[3]

Because the interviewer's requests exceed what would usually be asked of a stranger, securing the designated respondent's cooperation is a delicate business which requires tact and confidence on the interviewer's part. Different survey organizations emphasize different aspects of the survey in order to instill confidence in interviewers. SRC stresses the intrinsic rewards for respondents: "In many homes you will be welcome because you represent a change in the day's routine. Most people enjoy being interviewed and many develop interest and insight into matters which they have not thought about before or thought about in the same way." The National Opinion Research Center (NORC) tells interviewers that "The majority of the public are flattered by the opportunity to make their voices heard." And the Census Bureau, for the Current Population Survey (CPS), emphasizes the value of the data: "You as an interviewer have the duty and responsibility for stressing to respondents the need for and importance of these data."

There are also organizational differences in how interviewers are instructed to approach respondents and persuade them to grant interviews. SRC advises an interviewer to proceed in stages: first, establish her identity; second, get inside the door (because "It is easier for the respondent to say 'no thank you' at the door than it is in the living room"); and third, "convince the person of the value of his cooperation." In this process, an SRC interviewer is advised to provide the respondent with a minimal amount of

information. The SRC cautions, ". . . the doorstep introduction should be just enough to get you inside the house," and "Remember not to be too specific about the interview in introducing yourself and the survey to the respondent. It is important that you avoid introducing a bias into the interview. . . ."[4] Nevertheless, an SRC interviewer is instructed to give "correct and courteous" answers to any questions the respondent might have. An interviewer must be sensitive to the respondent's concerns if she is to be successful, and SRC advises her to "vary your approach in accordance with your intuitive feelings about the person. Some respondents will be willing to be interviewed with only a brief explanation of your purpose; for others you will need to go into some detail"—for example, about how the respondent was picked, why the interviewer is interested in his opinions, what the survey is about, or what good it will do. In the case of a respondent who questions the value of the survey, the tone of the SRC manual (and, we assume, of the interviewers who take their cue from it) is almost apologetic:

> This is one of the most difficult questions to answer. While a survey which adds to our knowledge about problems and concerns in society is valuable, a particular study will probably have little or no direct effect on individual respondents. This is a good time to bring out [newspaper and magazine] clippings from past studies. . . .

Interviewers for the CPS are similarly instructed to vary their approaches for different respondents, but, in contrast to SRC's instructions, there is a stronger and more forceful emphasis on the social importance of the data as a means of convincing respondents to participate.

> There will be a few respondents who want additional information, are reluctant to give information, or who refuse to be interviewed. It is your responsibility, as a Census Bureau representative, to sell the CPS program. . . . To convert reluctant respondents, try to identify the respondent's specific objection(s) and tailor your answer accordingly. A thorough understanding of the survey is the key to an appropriate explanation.

CPS interviewers are encouraged to explain the purposes and uses to be made of the survey results, and the CPS manual provides convincing explanations of the importance of collecting unemployment data and their use in policy making and planning.

NORC interviewers are instructed ("when appropriate") to "Tell respondents what you are doing. Have information about the study clearly in mind so you can explain it in such a way that it stimulates his interest." Neither CPS nor NORC place quite as much emphasis as does SRC on not telling respondents the purpose or nature of the interview. Indeed, these factors are the major selling points for the CPS interviewer.

These organizational differences in emphasis and approach reflect real differences among organizations in the content of their surveys, the uses to be made of the data and, perhaps, organizational philosophy. Because of these differences, and because there are differences in whether and how the survey organizations are known to the public, respondents are likely to raise different questions to interviewers representing different organizations. This likelihood is implicit in the fact that the three interviewer manuals differ in the attention they give to various questions or concerns that might be raised by respondents. For example, the CPS manual devotes more attention to questions about confidentiality and the voluntary nature of response than either NORC or SRC. The NORC manual instructs interviewers what to tell respondents who have never heard of NORC; this is evidently not a common problem for CPS or SRC interviewers, since their instruction manuals do not mention it. (CPS and SRC interviewers identify themselves as employees of the U.S. Bureau of the Census and the University of Michigan, respectively, in their opening statements and ID cards.)

It is important to note that, in her approach to a respondent, the interviewer is attempting to persuade him to cooperate with the survey. The fact that she successfully influences him at this point may interfere with her ability to avoid influencing him during the actual course of the interview. There is a certain conflict between an interviewer's roles as persuader and as neutral communicator. It is likely that most interviewers know at which stages of the interview they are to be persuasive and at which stages they are to be neutral, but we wonder whether respondents make this distinction so readily. The fact that the interviewer succeeds in her persuasive role early on may prime the respondent to be influenced in subtle ways during the rest of the interview.

In addition, the amount and kind of information that respondents receive prior to the interview are not constant. In contrast to their training in virtually every other aspect of interviewing, interviewers generally are encouraged not to standardize the opening lines they use with each respondent but to be responsive to the requirements of the situation. This implies that the respondent's first impressions of the interview situation, and his role in it, are dependent on particular questions he raises and the particular answers given by the interviewer. There is reason to expect individual variability in how respondents react to an interviewer's request for an interview. Since a respondent is unfamiliar with the situation, it is almost inevitable that he will initially adopt some more familiar social situation as a model for appropriate action before and during the interview. A respondent has several sources of guidance. First, some people have actually been interviewed in a survey, so the situation is not completely novel.[5] Even without this experience, it is likely that media coverage of election polls and other surveys has created a fairly widespread, if vague, idea of what it means to be

interviewed. That is, some concept of the role of respondent is familiar to a substantial and growing number of Americans. Second, everyone has been asked questions or talked about himself in various situations, and one of those situations may provide a model for appropriate behavior during a survey interview. Finally, an interviewer is prepared to teach a respondent how to behave, although there are strict limits on the kind of instruction she is supposed to provide. Ideally, she may not instruct the respondent too directly, but should attempt (gently) to correct inappropriate behavior on his part. In large part, she attempts to influence a respondent to select appropriate models for their respective roles (she's not a salesman, more like a friend, perhaps) and then modifies his expectations appropriately (she's not quite a friend, either, though, because she can't give him her opinions nor listen indefinitely).

8.2 The Respondent's Concept of His Role

With the interviewer's guidance, a respondent evolves his own role. The expectations held and communicated by the interviewer, and those held or developed by the respondent on the spot, are likely to vary among interviewers and among respondents. Let us consider several different meanings that an interview may have for the respondent.[6]

THE INTERVIEW AS INTIMATE CONVERSATION, CONFESSION, OR THERAPY SESSION

The interviewer's assurances of confidentiality (if they are believed), her friendly and attentive demeanor, and the fact that she is a stranger who is not involved in the respondent's life may invite a respondent to explore problems and feelings that he does not feel free to discuss in other contexts. As Converse and Schuman (1974:62–63) note, the interview offers a respondent an opportunity that is rare in ordinary conversation—the chance to talk freely about personal experiences and concerns and be listened to by a person who is tolerant and genuinely attentive. Accounts of successful studies of extremely sensitive topics sometimes suggest that the success of the project is due to the fact that respondents took advantage of the opportunity for intimate conversation. For example, Wish and his colleagues report that their 94 percent response rate in a study of Vietnam veterans "might have been attributable to the veterans' welcoming an audience for a recounting of the hardships of their Vietnam experience and the difficulties they were experiencing since return" (Wish et al., 1978:3).

Although people who approach the interview situation as if it were an intimate conversation are likely to be cooperative, they are also likely to deviate from the role of respondent by, for example, asking for the interviewer's opinions, or straying from the topic of the question to talk about other matters. (In the 1971 Detroit Area Study, for example, interviewers reported that 15 percent of the respondents "occasionally" or "often" asked their opinions.) In such cases the interviewer is usually advised to be humane but task-oriented. For example, Converse and Schuman (1974:63) suggest:

> When a respondent strains for information about the interviewer, delay is the recommended procedure: the interviewer explains that [she] is not to express [her] own opinions until after the questionnaire is over. . . . Digressions can be shortened by interviewer cues of inattention or even dexterous interruption, but some rumination or reminiscence may be essential for the respondent's continued concentration, interest, and sense of the interviewer's acceptance.

Of course, it is sometimes true that an interviewer also feels the pull toward intimacy with a person whom she finds congenial. Indeed, an early survey of interviewers conducted by Hyman found that 10 percent regarded overinvolvement with respondents as their chief failing as interviewers (Hyman et al., 1954:75).

THE INTERVIEW AS CITIZEN REFERENDUM

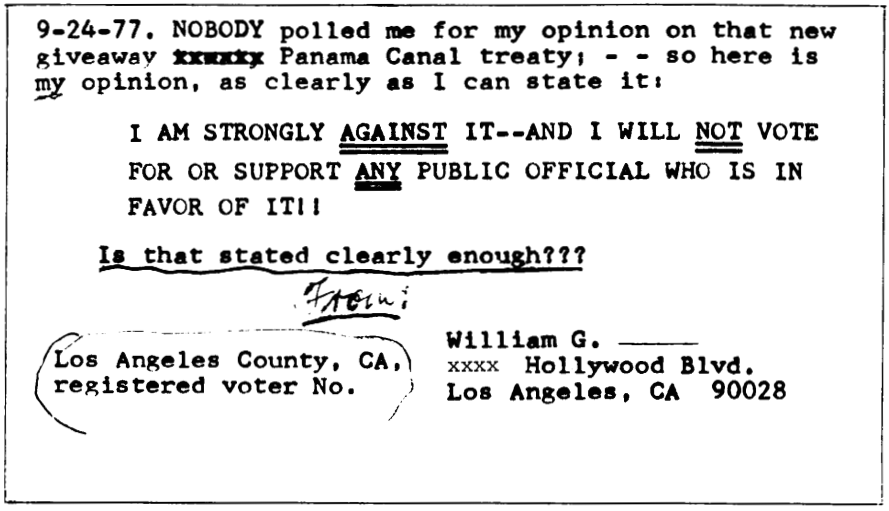
Surveys and polls have become important vehicles for transmitting public opinion to public officials and, via the mass media, to the American public. Hence many people may treat their responses to the interviewer as something akin to votes.

It is perhaps significant that pollsters receive mail from people who want their opinions to be counted. The postcard shown in Figure 8.1 was received by the Louis Harris Data Center and has been altered only to protect anonymity (with which the "respondent" was obviously unconcerned). Clearly, this man regards being interviewed as one mode of political expression, like voting or writing a public official. It seems unlikely that, as a respondent, he would treat the interview as an intimate conversation between two people; indeed, his desire is to be heard and counted in a larger political context. An interview would probably be meaningful to him only if he believed that it would enable him to reach the people he wants to speak to.

THE INTERVIEW AS AN EXERCISE OF FILLING OUT A FORM

To many respondents, an interview is not meaningful as a way of publicly or privately expressing their opinions and sentiments. The content of the interview is not as important as the fact that there is a form to be filled out for

FIGURE 8.1
Postcard received by Louis Harris Data Center.


9-24-77. NOBODY polled me for my opinion on that new giveaway xxxxxx Panama Canal treaty; - - so here is my opinion, as clearly as I can state it:

I AM STRONGLY AGAINST IT--AND I WILL NOT VOTE FOR OR SUPPORT ANY PUBLIC OFFICIAL WHO IS IN FAVOR OF IT!!

Is that stated clearly enough???

Thanks!

Los Angeles County, CA, registered voter No.

William G. ———
xxxx Hollywood Blvd.
Los Angeles, CA 90028

the benefit of a bureacratic third party. Thus, Converse and Schuman (1974:43) supply the following quote from a woman who was trying to persuade her niece (the designated respondent) to cooperate with the interviewer: ". . . it doesn't matter what you say as long as the young man checks off something or other. He just has to fill in some box or line."

Even if a respondent begins an interview in an interested and thoughtful frame of mind, the questions themselves may seem so boring and meaningless to him that he begins to respond mechanically, as Converse suggests (1970:177):

> . . . I would hypothesize that such a phenomenon as *test fatigue* is itself a direct consequence of pressures, felt by the subject to search for faint or nonexistent bits of affect to fulfill the requirements of the attitude questionnaire. . . . One outcome of such harassment is fatigue; another is a more or less conscious recourse to some response set touched off more by question form than question content.

A respondent who treats the survey as a meaningless exercise of filling in the boxes is unlikely to pay much attention to the questions or his answers; the likely result of such an interview is a pattern of random responding that suggests nonattitudes. Sometimes this occurs not because a respondent lacks strong opinions, but because he fails to make the connection between the numbers, boxes, and scales into which he is asked to fit his responses and the opinions and sentiments he expresses to the interviewer. That is, he may take his opinions seriously, but the notion of fitting them in boxes

is not meaningful, or is perhaps insulting to him. This is most likely to be true when the questions or response categories do not capture what a respondent is trying to say. (Poor questions are also likely to alienate the interviewer who must ask them as well as the person who answers them. In such cases, both the interviewer and the respondent are likely to reject the premises and validity of the questions.)

A respondent who regards an interview as a meaningless exercise often consents to be interviewed only as a personal favor to the interviewer, as implied in the preceding quote from a respondent's aunt. In such cases, the respondent sometimes treats the interviewer as if they were two people in league against the people who designed these foolish questions that the interviewer must ask. Converse and Schuman (1974:55) suggest that an interviewer who must ask a question that she finds distasteful or meaningless also "dissociates [herself] from it (and the anonymous souls who think up such things)—ever so slightly, as best [she] can, with language that leagues [her] subtly with the respondent." Sometimes an interviewer's collusion with the respondent against the investigator is not so subtle, as suggested in one interviewer's comment about her respondent: "He hated the closed questions; that's what I get for having an intelligent respondent" (quoted in Converse and Schuman 1974:67).

When the questions on a survey are regarded as meaningless by the interviewer or the respondent, or both, the inclination to get through the interview as quickly as possible may be quite strong. The desire to shorten the interview may lead interviewers to encourage, and respondents to give, short, cursory answers that fulfill the requirements of the questions minimally (or not at all). When a long series of followup questions is made contingent on a respondent's prior report of (for example) a health problem or an incident of victimization, there may be an incentive for interviewers and respondents not to report the incident in order to shorten the interview. Gibson et al. (1978) report that when a special supplement of questions about recent acute conditions was added to the Health Interview Survey, reports of such conditions dropped by 20 percent and then returned to previous levels when the supplement was eliminated from the questionnaire. Similarly, when a supplement of questions to be asked of discouraged workers (those who are technically not in the labor force but want or intend to look for work) was added to the CPS, there were substantial and atypical changes in labor force classification: the number of people classified as discouraged workers dropped. This occurred despite the fact that labor force classification was determined early in the interview and interviewers were instructed not to change it.

THE INTERVIEW AS A TEST

Many respondents lack confidence in their ability to provide adequate answers to survey questions. Interviewers are instructed to reassure anxious

respondents that the respondent's opinions are of interest no matter how ill informed he may feel himself to be. Thus, SRC (1976:11) advises its interviewers:

> You should avoid creating the impression that the interview is a quiz or cross-examination; be careful that nothing in your words or manner implies criticism, surprise, approval or disapproval either of the questions you ask or of the respondent's answers.

Nevertheless, there are many aspects of the interview that are strongly reminiscent of a test-taking situation. First, the interviewer is usually instructed not to help the respondent by explaining difficult words or ambiguous questions; rather, it is up to him to provide interpretations—"whatever it means to you"—as in a test. For example, the SRC manual states (1976: 12–13; the NORC and CPS manuals are similar):

> Occasionally . . . a respondent may misunderstand or misinterpret what is asked. When this happens, the best technique is to repeat the question just as it is written in the questionnaire. . . . If the respondent asks for an explanation of a word or a phrase, you should refrain from offering help and return the responsibility for the definition to the respondent . . . . If . . . the respondent does not know the meaning of the word . . . do not try to define or explain it but go on to the next question.

Second, despite the interviewer's reassurance, interviews frequently do include questions with right or wrong answers that are intended to test the respondent's knowledge. And even questions that do not have factually right answers may implicitly represent a test of knowledge. When respondents are called upon to express opinions about, say, Nixon's trip to China, or fringe benefits for union members, or the virtues of various political candidates, the implication is clear that they ought to know about the topic in question. Converse and Schuman (1974:68) quote an interviewer who empathized with her respondents' discomfort at being questioned on remote topics about which they knew little: "Most respondents seemed to think that I, the interviewer, probably knew the 'answers' to these questions while they were being forced into displaying their own ignorance." Test anxiety on a respondent's part appears to be relatively common, as suggested by interviewer reports that 17 percent of the respondents in the 1971 Detroit Area Study "occasionally" or "often" wanted reassurance that their answers were "adequate or correct or 'good ones'." (The figure ranged from 27 percent of respondents who did not finish high school to 8 percent of those who went to college.)

Duncan and Duncan (1978) present evidence suggesting that the person who lacks confidence in his ability as a respondent is also likely to lack

confidence in his opinions generally. In an analysis of the 1971 Detroit Area Study data, they find that respondents who agreed that "Sometimes politics and government seem so complicated that a person like me can't really understand what's going on" received lower ratings from interviewers on their general understanding of questions. Duncan and Duncan (1978:19) also found:

> Respondents who lack confidence in their ability to understand [politics]. . . more often solicit the interviewer's opinions, more often seek reassurance about their answers, have a greater desire to please the interviewer; and the interviewers less often say that they liked interviewing these persons a great deal.

The case in which a respondent treats the interview as a test of his knowledge can be considered as one of a general class of situations in which a respondent accepts the interviewer or survey questions as authoritative and sees his own role as one of trying to comply with what is expected of him. Survey researchers have long been concerned that respondents—particularly those with lower educational or social status than the interviewer—will be too acquiescent or deferential to the interviewer and the questions she asks. Lenski and Leggett (1960) proposed that deferential respondents tend to acquiesce to simple agree/disagree propositions even when acquiescence leads to contradictory responses. They claimed that this yea-saying tendency is more pronounced the greater the social distance between interviewer and respondent, and they supported that claim with data showing that agreement with item reversals was most frequent among respondents with little formal education and more frequent among black than white respondents. Their interpretation of the latter result was that traditional norms governing relations between blacks and whites led black respondents to be more acquiescent to white interviewers. For several reasons, Lenski and Leggett's interpretation of their findings as evidence for respondent deference to the superior status of the interviewer may be questioned (see for example Jackman [1973], Fischer [1974], Martin [1980], and Schuman and Presser [1981] for other interpretations and analyses of yea-saying). Nevertheless, Lenski and Leggett's point that ". . . interviews inevitably bring into being a relationship between respondent and interviewer in which the behavior of each is influenced by his perception of the situation and the appropriate social norms" (Lenski and Leggett, 1960:463) remains valid, although the effect of that relationship on answers given by respondents remains uncertain.

A second bias, which is believed to result in part from a respondent's desire to please and be liked by the interviewer, is due to social desirability. Survey researchers have long believed that respondents try to present themselves in the most favorable light to interviewers (and to themselves)

and hence falsely claim positive or desired traits, attitudes, or actions and deny undesirable ones. Thus, respondents overreport events that reflect well on them and underreport events that are embarrassing or undesirable. In one study, Weiss (1968) found that New York welfare mothers were more likely to report socially desirable events (for example, voting, their children having passed in school) when they had not actually happened if rapport was high between interviewer and respondent, and they were more likely to express socially desirable attitudes (for example, high aspirations for a child, negative feelings about being on welfare) when the interviewer was similar in status to the respondent. Although plausible, the hypothesis that a bias due to social desirability results if respondents are too eager to please interviewers has not been conclusively demonstrated. A review of the literature (see DeMaio, Chapter 9, Volume 2) reveals that much of the research on social desirability is conceptually flawed and poorly executed.

THE INTERVIEW AS SUBTERFUGE

Potential respondents may be suspicious of the motives of an interviewer or of the organization she represents for several reasons: respondents may suspect that the interviewer wants to rob them, sell them something, or obtain damaging information about them.

House and Wolf (1978) argue that a result of the widely reported increasing crime rate during the 1960s was heightened fear of crime and suspicion of strangers. One consequence was an increase in refusal rates in national surveys conducted by the Survey Research Center between 1956 and 1972, particularly in large urban areas and especially central cities. House and Wolf found that the strongest predictor of the refusal rate obtained in an area was the crime rate, and they concluded that rising crime rates led to a suspicion of strangers, including interviewers, which made people increasingly reluctant to grant interviews. Even when they grant interviews, suspicious respondents may withhold information. For example, the 1971 Detroit Area Study survey found that mistrustful respondents were more likely than others to refuse to give a phone number to the interviewer (Fischer, 1974).

A respondent may also suspect that the interviewer is a salesperson, and for good reason, since sales pitches and fund-raising promotions are often presented in the guise of opinion surveys. Over half of the Opinion Research Corporation's interviewers report having been refused because the respondent thought the interview was a sales pitch (Rugg, 1971–72, cited by Converse and Schuman, 1974). A 1963 Roper poll found that 27 percent of respondents claimed they had at some previous time been approached on a "survey" whose real purpose was to sell the respondent something (Baxter, 1964).

A respondent's suspicions may be aroused by the fact that the methods by which an interviewer gains entry are similar to a salesperson's technique (as we described previously, SRC interviewers are advised to get inside the door as quickly as possible and to give minimal information to the potential respondent). Because the tactics are similar, it may be difficult for a potential respondent to discriminate bona fide interview attempts from sales pitches.

In other ways, and for many survey organizations, recommended interviewer procedures are subtly evasive and indirect and, in some cases, manipulative, which may contribute to respondent suspicions. For example, it is doubtful that most interviewers give precise and direct answers to respondents' questions about how long the interview will take (see Converse and Schuman, 1974:42, and NORC and SRC manuals). CPS interviewers must advise respondents that the survey is voluntary, but this fact is deemphasized: it is stated in the "Respondent" letter, but is not part of the recommended opening statement (see the CPS manual) in which the interviewer introduces herself as an employee of the U. S. Bureau of the Census. Although she is required to ask respondents if they have received the "Respondent" letter and to provide them with a copy if they have not, she is evidently not required to ensure that respondents really understand the voluntary nature of the study. Finally, the NORC manual recommends flattery as an interviewer tactic: "A compliment about the beautiful flower garden, cuddly baby, interesting painting, etc. will often break the ice and help establish rapport."

These practices are not regarded by interviewers or the legitimate organizations they represent as deceptive, particularly since the interviewer's intent is benign—she only wants to gather information which will not be used to harm respondents. Nevertheless, some of the practices might be questioned on the grounds that a respondent is not fully informed before he consents to be interviewed. In addition, it seems possible that an interviewer's evasiveness at the door may in some cases arouse suspicions of a deception that is worse than is actually the case for a bona fide survey. It may be argued that the long-term effect of evasive and deceptive interviewer practices, even those that seem innocuous, is harmful both to respondents and the survey enterprise because they make it difficult for potential respondents to discriminate legitimate requests for interviews from other activities. In this regard, it is worth noting that Singer (1978) finds that fully informing respondents about the purpose of the interview results neither in lower response rates nor in less complete responses to sensitive questions.

Finally, respondents may suspect that the information that the interviewer requests and records will not be kept confidential, despite assurances. In a study of confidentiality, a National Research Council panel (1979) found that 80 percent of both black and white Americans did not believe that

census records are kept confidential or did not believe that confidentiality could be maintained. Only 5 percent believed that other government agencies could not obtain census records "if they really tried." Valentine and Valentine (1971) also noted the mistrust of the Census Bureau auspices by many inner-city respondents, who believed that any information they provided would be available to other agencies and would be used against them. Survey practices may sometimes exacerbate such fears. NORC, for example, provides police escorts to interviewers in high-crime areas, a practice that may make it difficult to convince a suspicious respondent that the information he provides will be held confidential and used "for statistical purposes only."

One apparent consequence of respondent mistrust is distorted information about black urban households. Valentine and Valentine (1971) compared results obtained by an intensive ethnographic study of an inner-city neighborhood with Census Bureau interviews conducted at the same households. (Census Bureau interviewers did not know that independent observations were made.) Their study confirmed the Bureau's undercount of adult black males, 61 percent of whom (according to this study) were unreported to Census Bureau enumerators for the small number of households studied. According to the ethnographic study, the reason for the omission was that, for a variety of reasons, the presence of men in black households was concealed from the enumerators. The result was distorted information about the characteristics of the 25 inner-city households in the study. For example, census data would have led to the inference that 72 percent of the households were headed by females, while the ethnographic observers concluded that only 12 percent were actually headed by women.[7] Nationally, black men are underrepresented in the census, with estimates of the proportion missed as high as 20 percent in some age groups, and 9 percent for all black males in 1970 (see Bureau of the Census, 1973, for details).

The potential respondent who, for whatever reason, doubts that the interviewer's stated purpose is the real one is likely to refuse to be interviewed, although many people grant interviews despite lingering doubts about the interviewer's motives. In one survey, interviewers' reports indicated that 26 percent of respondents were "very" or "somewhat" suspicious before the interview began (National Research Council, 1979). We do not know how many continued to be suspicious throughout the interview, and we do not know how respondents answer questions when they are in a suspicious frame of mind because they believe that the interviewer has an ulterior motive in asking them. It seems likely that respondents will be protective about information that they believe could harm them, and that they may therefore lie; admissions made in focused group interviews suggest that this is so (see National Research Council, 1979).

8.3 Social Expectations

The social characteristics of an interviewer and a respondent, such as age, race and sex, are significant during their brief encounter: different pairings have different meanings and evoke different cultural norms and stereotypes that influence the opinions and feelings expressed by respondents.

In particular, the effect of the interviewer's race upon responses is well documented. Both black and white respondents give more racially liberal responses to black than to white interviewers (see Chapter 5; Hyman et al., 1954; Williams, 1964; Schuman and Converse, 1971; Hatchett and Schuman, 1975–76; Schaeffer, 1980; Campbell, 1981). Sudman and Bradburn (1974) conclude that the race-of-interviewer effect is restricted to questions pertaining to race, although Schuman (1974) finds that nonracial political items that tap symbolic allegiance to the country or its institutions may also be affected. The most reasonable explanation for race-of-interviewer effects is that the respondent wishes to "avoid offending a polite stranger who has come into his home for a legitimate reason" (Schuman and Converse, 1971) and therefore mutes his expressions of racial hostility to an interviewer of the opposite race. As Schuman and Converse note (1971:58):

> To accept a guest into your house and then proceed to explain that you neither trust nor feel friendly toward people of [her] race probably takes more chutzpah than the average respondent possesses.

The effect of interviewer's race is presumably mediated by the respondent's expectations of what sentiments and attitudes can be expressed without offending interviewers of the opposite race.[8] Interviewers are also likely to modify their behavior in accordance with their expectations about respondents. Hyman et al. (1954) suggest that interviewers are guided by their expectations about the general nature of attitudes and their stereotypes or expectations about respondents with particular social characteristics. According to Hyman and his colleagues, different expectations lead interviewers to adopt different roles, and "the differing roles that interviewers define for themselves with respect to probing, rapport building, recording, etc., will account in part for differences in the results they obtain" (Hyman et al., 1954:57).

There is evidence to suggest that an interviewer's subtle cues can influence respondents' answers to survey questions. An interviewer's voice inflection, for example, has been found to influence respondents' reporting of health events (Barath and Cannell, 1976), although the evidence is mixed (Blair, 1977–78). The hypothesis that different expectations are expressed in systematic differences in interviewer behavior has not really been explored in the survey context. However, some indirect support for the hypothesis of

Hyman and his colleagues is provided in an experimental study by Word, Zanna, and Cooper (1974). They conducted a two-part experiment designed to explore the role of interviewer stereotypes upon respondent behavior in a simulated employment interview. In the first stage, they found that naive white interviewers displayed more intimacy toward white than toward black interviewees. (Intimacy was measured in terms of nonverbal cues such as physical distance and the length of the interview.) In the second stage of the experiment, interviewers were trained to nonverbally express intimacy or distance, and the effects of the differential treatment on the performance of naive white interviewees was measured. Interviewees who were treated more distantly by interviewers reciprocated by behaving distantly themselves, liked the interviewer less, and gave less adequate job-interview performances (as measured by judges' ratings). Word, Zanna, and Cooper (1974) interpret their findings to mean that white racial stereotypes lead to differential treatment of white and black interviewees in ways that adversely affect the performance and attitudes of the latter.[9]

Their findings are not directly applicable to survey interviews, since their interview was a simulated employment interview, the respondents were college students, and there were no survey-type questions. Nevertheless, the same processes are likely to play a role in the survey setting, and experimental investigations might provide insight into the social dynamics involved. For example, Price and Searles (1961) and Pettigrew (1964) found that black respondents were more often able to identify political leaders when questioned by black than by white interviewers. Schuman and Converse (1971:59) suggest that this effect is due to the "suppression or repression of knowledge by black respondents when confronted by white questioners." It is thus possible that white interviewers communicate negative expectations that adversely affect the performance of black respondents, in much the same way that Word, Zanna, and Cooper (1974) found in their studies.

8.4 The Presence of Others During the Interview

Although we have analyzed the relationship between respondent and interviewer as if they were the only two people involved in the interview, this is frequently not the case. In the 1971 Detroit Area Study, for example, 45 percent of the interviews were conducted in the presence of other people, most commonly the respondent's spouse (27 percent). The presence of other people often distracts a respondent and disrupts the interview, some-

times influences a respondent's answers to the questions, and in rare cases may alter an interview situation dramatically (Converse and Schuman [1974] provide some vivid examples). The presence of others thus represents an additional source of uncontrolled variability among interviews. (The NORC manual advises interviewers to conduct interviews in private, but neither the SRC nor the CPS manuals mention whether interviews should be conducted with other people present.)

Using data from the 1971 Detroit Area Study, Duncan and Duncan (1978) analyzed the effects of the spouse's presence on a respondent's answers. They found that the presence of a husband was more likely to be disruptive than the presence of a wife during an interview. Of those interviews conducted in the presence of spouses, 41 percent of the interviews with wives and 23 percent of the interviews with husbands were affected by the spouse's presence (according to the interviewer's rating). Additional analysis of these data shows that wives who were asked how satisfied they were with different aspects of their marriage expressed greater satisfaction in response to three of the four questions if husbands were present than if they were not. Similar findings are reported by Taietz (1962).

We note, however, the difficulty of determining whether wives expressed more satisfaction with their marriages when their husbands were present because of response distortion, or because the husbands who were present really were more satisfactory. The aspect of marriage that shows the greatest variability in satisfaction according to whether the husband was present or absent while the questions were asked is companionship, which is linked in an obvious way to a husband's presence or absence more generally. Because situational factors are likely to be related to real sources of variability in attitudes and behavior, their effects could only be unambiguously determined by experiment.

8.5 Implications

Our discussion of the social dynamics of the interview situation suggests several conclusions. First, variability in the social aspects of the interview situation results in variability in respondents' role expectations and behavior during interviews. We have discussed four possible sources of variability in the respondent role from one interview to the next: different expectations communicated by interviewers; different expectations and preconceptions held by respondents or developed on the spot; varying characteristics of interviewers and respondents, such as sex, race, and so forth; and the presence of other people during the interview.

Second, it is quite likely that variations in a respondent's role expectations and behavior influence the answers to survey questions. This would occur because the types of errors or distortions in responses are likely to be quite different, depending on the respondent's frame of mind and how he conceives his proper behavior during the interview. For example, respondents who are eager to please seem less likely to refuse to answer an interviewer's questions and more likely to distort their answers to present themselves in a favorable light to her. Even if respondents' answers are error-free, they may vary according to how they conceive their role as respondent. For example, we have identified two good respondent roles (the intimate and the citizen roles), in the sense that such a respondent is actively interested in the interview and attentive to the answers he gives. But even in this case, we might expect systematic differences in a person's answers depending on whether he treats the interview as an intimate conversation or a citizen referendum. In one setting, he may respond as an individual revealing private thoughts and feelings, and in the other, as a citizen voting his public attitudes and preferences for the society as a whole. Thus, on topics for which private sentiments and public opinions are discrepant (perhaps matters pertaining to race, sex, and morality, for instance), we might expect differences between a survey that encourages respondents to adopt an intimate role and one that encourages the citizen role.

We have only begun to understand how the interview, viewed as a social relationship, influences responses to survey questions. We do know that the variability attributable to interviewers may be very great. (See, for example, the interviewer variance study conducted by Bailey, Moore, and Bailar [1978] in connection with the National Crime Survey.) It is likely that some, but not all, variability among interviewers occurs because interviewers differ in the social relationships they establish with respondents and the role expectations they communicate to them. Of course, interviewers also vary in how well they perform the particular tasks of asking questions and probing and recording answers (see Chapter 9).

Before situational variability can be adequately controlled or eliminated, we need more systematic investigation of (among others) the four sources of variability previously discussed. Such investigations require, at a minimum, that surveys routinely record information on each interview situation, such as the presence of other people, the interviewer's characteristics, and so on. (It is distressing to discover, for example, that the National Crime Survey does not record the presence of other people during the interview, even though the respondent is asked to report incidents of victimization potentially involving family members or friends as offenders.) Considerable investigation would be required to adequately understand respondents' varying conceptions of their role and to develop procedures by which inter-

viewers may more effectively communicate appropriate role expectations to respondents. (For some evidence on respondents' poor comprehension of the task requirements of the interview, see Chapter 9; and Cannell, Oksenberg, and Converse, [1977; 1981]. Cannell and his colleagues also offer some promising approaches to the problems.)

We believe that investigation of these (or other) sources of variability should not assume that situational variables can be analyzed in the abstract, as if an interview is a sterilized measurement device isolated from the cultural context in which it is used. Rather, it seems fruitful to approach the interview holistically, as a situation that has meaning to its participants, in order to develop a more adequate understanding of the influence of the survey process upon survey results.

The goals of a more intensive and holistic analysis of the situational determinants of survey response are twofold. First, a better understanding of the interview situation might be used to improve survey procedures so that situational variability is eliminated, controlled or, at a minimum, recorded for further analysis. At present, the social interaction between interviewers and respondents is subject only to loose and crude attempts at standardization. (For example, many survey organizations attempt to match the race of respondent and interviewer.) Efforts to monitor, understand, and control the social aspects of surveys interviews could, and should, be more rigorous than they are at present.

Obviously, different problems will call for different remedies, and a technique that is helpful in one situation may be harmful in another. For example, Henson, Cannell, and Lawson (1973) found that a personal (rather than impersonal) interviewing style sometimes led to better, and sometimes to worse, reporting by respondents. A personal style of interviewing resulted in more complete and accurate reporting among respondents who had little education, but worse reporting among highly educated respondents. Their interpretation is that a personal style eased the anxieties of the former group (who perhaps treated the interview as a test) but was resented in the latter group, who took it as a cue that the questions were not to be taken too seriously.

Monetary payment has also been found to have inconsistent effects on respondents' willingness to participate in surveys (see, for example, Ferber and Sudman, 1974). It is plausible that the effectiveness of a monetary incentive to elicit respondent cooperation depends (among other factors) on how it influences a potential respondent's definition of the interview situation and his role in it. In some circumstances, the offer of money may arouse suspicion, as was evidently the case in an SRC study that requested information about respondents' savings accounts: response rates were substantially lower when respondents were offered compensation than when they were not (Lansing, Ginsberg, and Braaten, 1961). Respondents who

are already highly motivated in the role of respondent (because they approach it as an intimate conversation or a citizen referendum) may well be insulted by the offer of a monetary reward, since payment essentially robs the interview of the particular meaning they had given it, and their interest and effort may fall off as a result.[10] On the other hand, payment can be helpful when respondent motivation is low because the interview requires effort at a task that is not seen as intrinsically rewarding. Ferber and Sudman (1974:330) note that the "effects of compensation are especially clear in obtaining higher response . . . if sample members are requested to keep written records of expenditures." Thus, in some circumstances, money can induce respondents to take the task seriously.

When situational factors are better understood, controlled variations in, say, respondent role expectations will be able to provide information about the nature and generalizability of the attitudes measured in surveys. A systematic discrepancy between answers given in public versus private modes, for example, would not simply be an error to be eliminated; surely such a finding would tell us something about the inconsistency or ambivalence with which attitudes are expressed in the larger social context. The fact that an interview, as a particular social situation, is inseparable from the cultural context can be informative about the larger society in ways that have not been systematically exploited to date. (See Schuman [1982] for a discussion of this point, and Williams and Cantril [1945] for an early attempt to interpret "artifacts" as meaningful differences.)

More generally, it is important not to lose sight of the fact that the interview setting is itself distinct from other settings in which attitudes are expressed, and hence we should not expect to find complete congruence between attitudes expressed in interviews and in other social contexts. A better understanding of the factors that differentiate survey interviews from other contexts of social expression would enhance our ability to interpret answers to survey questions. (See, for example, Zanna, Higgins, and Herman [1982] for a social-psychological analysis of the situational determinants of actions and attitudes.) Some hints concerning the ways in which interviews are different from other social settings are offered by Hyman (1949) and by Converse and Schuman (1974), who suggest that a bias of rationality characterizes survey interviews. Several factors effectively dampen the intensity of feeling expressed by the respondent during an interview. First is the formalism of the questions themselves. A respondent must fit his reactions into a set of neat formal categories, which constrains and structures the way he expresses himself and may impose a distance between his feeling and his response. Questions themselves are usually worded neutrally, toned down so as not to elicit too violent a reaction from respondents. Perhaps most important, an interviewer's tone and manner are very reasonable and impartial. In Hyman's words (1949:39), "an atmosphere of permis-

sive indifference is provided." Overall, the interview setting seems designed to encourage detachment and thoughtfulness on a respondent's part and therefore may create a bias in favor of reasonable, rational answers to the interviewer's questions.

A second bias inherent in the interviewing situation might be termed a bias of individuation. The interview ideally occurs in a personal one-to-one interaction, during which a respondent is asked to voice *his* opinions—not what neighbors, spouse, or friends would think. This, and the fact that the interviewer is a stranger, serve to isolate the interview from the respondent's normal life and especially from the normal social pressures and constraints that may inhibit or distort his expression of opinion. If he doesn't already believe it, the interviewer tries to convince the respondent that his opinions are important. Indeed, the interviewer's effort to secure an interview may itself impress the respondent. As Converse and Schuman note (1974:45):

> . . . the interviewer has made some sort of personal behavioral demonstration that this particular respondent is valuable and important to [her] . . . —calling back, making appointments, and accommodating the respondent's busy life.

(See also Abelson's [1982] discussion of individuated and deindividuated social contexts.) Attitudes elicited in an individuated context may be predictive of behavior in other, similar individuated contexts, such as voting. However, one would not necessarily expect the attitudes expressed in survey interviews to be consistent with actions or attitudes in, say, a highly charged group setting. If we wish to predict the latter, we must heed Hyman's deceptively straightforward advice to incorporate the fundamental aspects of the setting into the test.

Notes

1. In this chapter, we draw mainly on results from face-to-face interviewing. Obviously, the increasing reliance on telephone interviewing represents a potentially important change in the interview setting. Although there are sometimes differences between results obtained by the two methods, (for example, respondents are apparently more uneasy in telephone interviews), the differences to date do not appear to be as great as might have been expected (see, for example, Groves and Kahn, 1979; Bushery, Cowan, and Murphy, 1978; and Woltman, Turner, and Bushery, 1980). However, there is not yet enough evidence to draw any firm conclusions about differences between personal and telephone interviews.

2. Throughout Chapters 8 and 9, interviewing manuals prepared by the Survey Research Center (1976), the National Opinion Research Center (1972a, 1972b), and for the Current Population Survey (Bureau of the Census, 1980b, 1981) are used as sources of information

about the norms that govern interviewers' behavior. For our purpose, these manuals are useful primarily because they codify common lore concerning proper interviewer practice. However, it should be noted that few survey organizations have interviewer manuals at all, and that even when manuals are available, interviewers may not follow their instructions. Thus, the interviewer role as described in this section is an ideal, from the point of view of the survey organizations, which may not be met in actual practice.

3. Indeed, it is because they must violate norms of privacy that neophyte interviewers frequently experience embarrassment and discomfort. Although SRC advises its interviewers to "Remember that you have a right to ask a person for information, just as the respondent has a right to refuse to answer," it is initially difficult for an interviewer to overcome the feeling that the mere asking is an invasion of the respondent's privacy. Such attitudes on the interviewer's part are not without consequences. For example, CPS interviewers who believe that it is improper to ask about the respondent's income obtain slightly higher rates of nonresponse for that item (Bailar, Bailey, and Stevens, 1977).

4. The bias referred to is that of respondent self-selection on the basis of their positive or negative reactions to the topic of the survey. It is important that potential respondents accept (or refuse) an interview for reasons that have nothing to do with the content of the survey.

5. A 1976 survey by the National Research Council (1979) finds that 45 percent of respondents report having been interviewed at least once in the previous 4 or 5 years. Being interviewed is thus increasingly a part of American culture. See Biderman (1975) for a discussion of the implications of this trend.

6. The following discussion has benefitted from Converse and Schuman's (1974) perceptive account of interviewing from the interviewer's point of view. Observers' descriptions of actual interviews done for the Census Bureau are also informative. See, for example, Rothwell and Tracey (1966).

7. The numerical estimates from this study should be treated with caution. The enumerated area was small, and a number of men had ties to households outside the area. Many of them would have been counted by census enumerators as residing at those households, and thus the 61 percent "missed" is almost certainly an overestimate. In addition, because the ethnographers' observations were conducted over an extended period of time while the census enumerators typically visited each household once, it is questionable whether the residency definitions employed by the two methods could be truly comparable. Nonetheless, this study yields important insights into some of the reasons for underenumeration of minority populations.

8. The effect (discussed in Chapter 5) of interviewer's race on the race of the entertainers mentioned by respondents as their favorites is probably not due to respondents' fear of offending the interviewer.

9. We note that their study can be criticized because they never tested the effect of differential interviewer treatment on naive black interviewees.

10. There is an extensive experimental literature (outside of survey research) on the effects of monetary payment on intrinsic interest in a task (see, for example, Deci, 1975).

9

The Question-and-Answer Process

Although question wording and question order sometimes have substantial effects on response distributions, survey researchers often cannot predict when such effects will occur or explain them when they do occur. This lack of knowledge suggests that the process by which respondents interpret survey questions and formulate answers to them is not yet well understood. Furthermore, it is probable that many survey artifacts represent systematic psychological phenomena that do not exist only in surveys. We therefore consulted relevant psychological literature in order to formulate hypotheses about psychological sources of bias in the question-and-answer process. In this chapter we review four analytical approaches to bias, each of which suggests different sources of error in survey responses.

At the most elementary level, it is apparent that respondents cannot give meaningful answers if they do not understand a question. This can happen if a question is ambiguous or incomprehensible to respondents, or if the interviewer fails to communicate the meaning and intent of the question. The first section is concerned with the problem of communication between interviewer and respondent.

A second source of bias, discussed in the second section, is the influence of a question on a respondent's subjective state. A question is itself a source

of information that sometimes alters the way a respondent classifies or labels the event in question, and a respondent's answer represents a verbal commitment that, despite its seeming lack of consequences, can reinforce the respondent's belief or intention. Response variability can occur if the process of answering a question changes the subjective state to which the question refers.

A third source of bias is due to contextual influences within sets of survey items. In a survey, prior items or item sets can influence responses to a subsequent item. The psychological evidence for consistency and contrast effects is discussed in the third section.

Finally, survey questions can be approached as cognitive problems the respondent is asked to solve. Although the questions asked in surveys present respondents with a diverse and complex set of problem-solving tasks, relatively little is known about the structure of these problems or how respondents go about solving them. Some evidence on two different cognitive tasks posed to respondents in surveys, recall and introspection, is briefly discussed in the fourth, and final section of this chapter.

9.1 Rules for Asking and Answering Questions

It is the interviewer's job to communicate the meaning of the questions to the respondent and to establish a set of rules that provide the framework within which the respondent formulates answers to them. The interviewer is typically instructed to ask each question exactly as it is worded and in the order specified in the questionnaire. She should not explain questions to respondents, but should probe incomplete or irrelevant answers in order to help the respondent focus upon the specific content of the question. She must communicate the same stimulus (that is, question) to all respondents so that their answers are comparable, and respondents must be encouraged to provide specific, complete responses that satisfy the objectives of the questions.

In the case of closed questions, respondents are encouraged to answer in terms of the response categories printed in the questionnaire. When a respondent says "I don't know" or gives an irrelevant answer, the Survey Research Center (SRC) interviewers, for example, are instructed to repeat the question, pause expectantly, or probe the response in a neutral manner (for example, How do you mean? Which would be closer to the way you feel?) to encourage an appropriate response. The National Opinion Research Center (NORC) interviewers are similarly instructed to "Always try at least once to obtain a reply to a 'Don't know' response, before accepting

it as the final answer," but to "Be careful not to antagonize a respondent or force an answer if he says again he doesn't know."[1] Interviewers should not help a respondent fit his answer into one box or another; SRC warns that "the answers recorded in the questionnaire should reflect decisions made by the respondent, not by the interviewer."

In the case of open questions, it is more difficult for an interviewer to recognize when a respondent's answer fails to meet the objectives of the question. SRC interviewers are asked to "study the instruction book thoroughly before starting to use the questionnaire," because "it is only through a complete understanding of the objectives that you can recognize when and where probes are needed." Similarly, the Current Population Survey (CPS) manual provides detailed instructions and explanations for each question in the CPS interview.

The rules of interviewing are based on the premise that the survey questions themselves are worded so that most people can understand them and give meaningful answers in terms of the fixed categories. This assumption is not always true, and when questions are poorly formulated, interviewers may be tempted to rephrase or explain a question. Perhaps in part for this reason (as well as interviewer carelessness), interviewers frequently do not follow the rules for questioning respondents, according to research conducted by the Survey Research Center and by the Bureau of the Census. In an experimental evaluation of the performance of CPS interviewers in tape-recorded, mock interviews, Rustemeyer (1977) found that interviewers were more often correct in their written work than in the oral part of their task; nearly one-half of all errors were related to how well interviewers asked questions. CPS interviewers, regardless of their experience, asked only about 60 percent of the questions exactly as they were worded. Similarly, an analysis of actual tape-recorded interviews conducted by SRC professionals shows that 36 percent of the questions were not asked as written, with 20 percent altered sufficiently to destroy comparability (Cannell, Lawson, and Hausser, 1975). In the latter study, 19 percent of the probes were directive, introducing the possibility of response bias.

Moreover, experience does not reduce the likelihood of some of the most serious types of interviewer errors. For example, Rustemeyer (1977) found that experienced CPS interviewers were more likely than inexperienced interviewers to alter the scope of questions, to change the sequence of questions, and to probe directively. Perhaps the reason that experience does not enhance all aspects of interviewer performance is that some interviewer errors are extremely difficult to detect. Rustemeyer estimates that only 19 percent of all the interviewer errors found in her study could have been detected in an office review. There can be little feedback provided to interviewers concerning these "invisible" errors and little incentive for interviewers to improve performance. These findings imply that office edits

and reviews are not sufficient to ensure low rates of interviewer error, and that it is almost impossible to detect and control many serious interviewer errors without systematic, direct observation of interviewers' performances. As Rothwell (1979) shows, the errors made by the CPS interviewers are neither random nor trivial. In the CPS mock interview study, interviewers' errors tended to be highest for cases in which questions were ambiguous or rules contradicted common sense.

An interviewer is supposed to provide feedback to respondents to encourage them to supply complete and accurate reports that fulfill the requirements of the questions. Marquis and Cannell (1969) found that SRC interviewers frequently gave feedback, but not in a fashion that led to better reporting by respondents. Interviewers gave positive feedback indiscriminately for both good and poor respondent behavior; indeed, some of the worst respondent behaviors (such as refusing to answer a question) were most likely to elicit positive comments by interviewers. Cannell, Oksenberg, and Converse (1977:309) conclude:

> Feedback, properly used, is an effective means of influencing behavior in desired directions. If the behavior of interviewers in this study [by Marquis and Cannell] can be taken as a guide, the technique is not being used by survey interviewers in ways that promote good respondent behavior. Rather, interviewers appear to use it primarily to reassure the respondent (and themselves) by indicating receptivity and approval.

In fact, the quality of data can be improved by feedback that is contingent on respondent performance. Vinokur, Oksenberg, and Cannell (1977) found that more complete health information is obtained if respondents receive positive feedback when they provide such information.

In the light of the aforementioned findings, perhaps it is not surprising to find that respondents often do not have a clear conception of what is expected of them. In a reinterview of 412 respondents one day after their participation in a Health Interview Survey, Cannell, Fowler, and Marquis (1968) found that respondents had very vague notions of the survey itself, why or by whom the information was collected, and what sort of information was required in response to the questions. For example, despite the fact that the questionnaire was set up to obtain detailed and complete reports about specific health conditions and events, only about half of the respondents said that the interviewer wanted exact information; the rest thought general ideas were sufficient. Three-fourths of the respondents thought the interviewer wanted a report of everything, while one-fourth thought that only more important things should be mentioned.

Belson (1968) carried out an experiment to discover how often respon-

dents understood survey questions as the researchers had intended. Respondents were interviewed on fairly straightforward topics regarding media use and then were debriefed. Belson found that the best understood questions were interpreted exactly as the researchers had intended by about 50 percent of the respondents. A substantial portion of the questions were correctly understood by fewer than 20 percent of respondents. Respondents usually had interpreted the questions so as to make them more applicable to themselves or more easily answered. Belson recommended more rigorous pretest debriefing in survey research in order to ensure that questions are meaningful to respondents.

More generally, we may question whether survey researchers have devoted sufficient attention to the problems of communication between interviewer and respondent. Perhaps we too frequently assume that respondents understand the meaning and purpose of survey questions and therefore we do not provide interviewers with an adequate set of tools or procedures to test and resolve confusion on the part of respondents (see Chapter 7 for examples of such methods).

9.2 The Individual Survey Question as a Source of Bias

It is clear that respondents do, and should, give different answers to the different questions posed in surveys. Yet, as we reviewed in Chapter 5, answers sometimes vary even when questions differ in ways that appear substantively trivial. Such variations may occur because the verbal labels and grammatical structure of a question subtly influence how a respondent thinks about an issue. The influence of a question is likely to be particularly great if the subjective state or event in question is not well defined or clear in a respondent's mind. In that case, the information that the respondent derives from the question may dominate whatever information is available from introspection, so that the respondent's report is influenced more by the former than by the latter.

Loftus and Palmer (1974) provide an excellent experimental example. Subjects viewed films of automobile accidents and then answered questions about details of the accidents. The authors found that the question, "About how fast were the cars going when they smashed into each other?" elicited significantly higher estimates of speed than when "smashed" was replaced by "collided," "bumped," "contacted," or "hit." That is, the subjects' perceptions of the event were influenced by the way it was labeled in the

question. Loftus and Palmer note that people are generally inaccurate in reporting numerical details, such as time, speed, and distance, and one would expect labeling effects to be stronger when the event in question is ambiguous and difficult to judge, as in this case. Furthermore, such labeling effects are not transitory. Subjects questioned a week later about the film still showed the effects of the label used in the question. For example, subjects were significantly more likely to remember that the film of the accident had depicted broken glass (which it had not) if they had earlier been asked the "smashed" question than if they had been asked the "hit" question or no question at all. Loftus and Palmer suggest that their subjects constructed a single, integrated representation of the event based both upon their own direct observation and the additional information (the verbal label) supplied by the question.

Changes in the grammatical structure and form of a sentence—changes that appear quite trivial—can also subtly alter its meaning. Indeed, some linguists (see especially Bolinger, 1977) argue that any alteration in form necessarily changes the meaning of a sentence. For example, Bolinger analyzes the semantic distinction between active and passive constructions in English and asserts that passive voice is only appropriate if "the subject . . . is conceived to be . . . genuinely affected by the action of the verb" (1974: 67). Evidence from psychological experiments also supports a difference in meaning between active and passive constructions. Pryor and Kriss (1977) presented subjects with a series of statements in either active (for example, "John likes the car") or passive voice (for example, "The car is liked by John"). The grammatical structure of the sentence influenced both memory and inference: persons were relatively more salient, and seen as more causal, than objects when the sentence was in the active rather than the passive voice. According to studies from several disciplines, it is probable that the structure of survey questions influences the answers given to them, because the logic and structure of a question provide a way of organizing one's thoughts and feelings about the events in question.[2] Linguists, psychologists, and survey researchers sometimes appear to work on similar topics unaware of each others' work. More communication among these three disciplines might prove fruitful in solving common problems (see Recommendation 11, Chapter 10).

A second way that a question may influence a respondent's thinking is as a persuasive message. A question may contain an implicit premise that a respondent is led to accept, or the question may otherwise induce the sentiment or opinion that it is intended to measure. As an example of the former, Loftus (1975) found that subjects who viewed accident films and were asked questions which varied only in the form of the article (for example, "Did you see *a* broken headlight?" versus "Did you see *the* broken head-

light?") were more likely to give a positive response to questions of the latter type, both when the event had actually appeared in the film and when it had not. Loftus argues that the use of the definite article suggests to subjects that the event in question did, in fact, occur.

Similarly, questions with strong affective connotations may suggest the sentiment or mood they are intended to measure. For example, a respondent who is asked to agree or disagree with a series of statements such as "It's hardly fair to bring children into the world with the way things look for the future" and "These days a person doesn't really know whom he can count on," may find himself feeling gloomier at the end of the series than he did at the beginning. In this regard, it is interesting to note that one technique that has been used (see Velten, 1968) to induce unhappy moods in experimental subjects is to have them read a series of gloomy, self-referential statements (for example, "I'm feeling very lonely, isolated, and depressed") that are similar, in both form and content, to survey items used to measure alienation and other negative feeling states. The fact that this mood-induction technique is effective in experiments suggests that respondents in surveys may be similarly affected when presented with a series of such statements. Indeed, we wonder whether the use of self-referential statements in surveys may not generally be a more intrusive form of measurement than questions.

Although we might expect mood changes induced by negatively (or positively) worded questions to be transitory, they may influence reponses to subsequent questions in the survey. For example, Smith (1981) and Turner and Krauss (1978) speculated that a prior series of negatively worded questions about the government in a 1976 Harris survey may have lowered respondents' evaluations of their confidence in institutions. Their hypothesis received partial support in an experimental test (see Turner, Chapter 7, Volume 2). However, the results of an earlier experimental variation in the order of political alienation items do not support a mood-induction hypothesis (see Duncan and Schuman, 1980). Other, nonobvious measurements may also be influenced by induced mood changes. For example, Bower (1981) reviewed evidence suggesting that subjects in whom happy or sad moods have been induced have differentially higher recall for events that are congruent with their mood.

Survey questions sometimes have more important and lasting effects on a respondent's sentiments and opinions than just his responses on the survey. A respondent who has no interest in politics and no confidence in his political opinions may develop more interest and confidence after being interviewed on political topics—especially since the interviewer is so eager to interview him and listens to and records his answers so diligently. According to several studies, pre-election interviews stimulate higher rates of po-

litical participation. Clausen (1968–69) suggests that one source of the systematic discrepancies between official and SRC postelection survey estimates of voter turnout is that respondents in the postelection survey were stimulated to vote by pre-election interviews. For example, the SRC postelection survey estimate of voter turnout in the 1964 presidential election is 15 percent higher than the official estimate. Although restrictions on the SRC sample, nonresponse bias, and overreporting by respondents explain part of the difference, Clausen concludes that part of the difference is due to the effect of SRC respondents' having been interviewed before the election. Clausen's hypothesis is confirmed by several subsequent investigations by Kraut and McConahay (1973), Yalch (1976), and Traugott and Katosh (1979).

Finally, responses to many survey questions represent verbal commitments for or against policies, principles, or candidates. Of course, a respondent's choice is not very consequential, and he will not be publicly identified with the stand he has taken; nevertheless, the mere fact of having voiced an opinion, pro or con, may strengthen his intention or belief. Indeed, experimental research suggests that it is precisely in low-pressure, free-choice situations (such as a survey interview) that a statement pro or con may strengthen one's original inclination (see, for example, Kiesler, 1971).

The possibility that respondents' attitudes, beliefs, and intentions may change as the result of being questioned about them is extremely important for panel surveys. Since the same respondents are interviewed more than once in such surveys, it is possible that prior interviewing alters their responses in subsequent interviews. The phenomenon of "rotation group bias" is well documented (see Bailar, 1975); for example, the number of respondents classified as unemployed in the Current Population Survey decreases according to the number of prior interviews. Although there are a number of explanations for this (including selective retention), it is also possible that the interview itself stimulates changes in job-seeking behavior in much the same manner that questions on political matters may stimulate respondents to vote.

Apparently, interpreting the meanings of a survey question and an answer to it is more subtle and difficult than might be supposed. Changes in wording that seem trivial at first glance may prove on further analysis to imply subtle and significant changes in meaning. One partial solution is to preserve the wording of a question exactly if results from different surveys are to be compared, although meanings can change over time. The problem is also compounded since a question is never asked alone, and its meaning may change according to the context created by preceding questions. The next section deals with artifacts due to the order and context of survey questions.

9.3 Contextual Influences Within Sets of Survey Questions

The issues involved in order effects are very subtle. Consistency of response across two or more related items in a fixed order is not in itself evidence for artifact. Respondents with very stable, well-structured attitude systems, or whose attitudes derive from a stable social network giving currency to a packaged set of opinions (Abelson, 1979), are likely to give consistent responses to a variety of items in the same domain. It is important to distinguish conceptually between revealed consistency reflective of the coherence of underlying prior attitudes, and constructed consistency developed on the spot under pressure of the interview situation. Even this distinction is not adequate, however, as there is an intermediate case in which consistency is constructed for the first time in an interview, yet reflects the underlying dispositions of the respondent. This might occur when a set of questions gives a respondent the opportunity to think through the value implications and interrelations of incompletely considered positions, revealing previously undiscovered consistencies. Alternatively, superficial consistency might be produced as a transient response to some passing and irrelevant influence in the interview, unconnected to underlying dispositions. By superficial consistency, then, we mean an apparently coherent pattern of responses to a set of items, which would not have occurred if the items had not been simultaneously present in the same interview and which will not survive long after the interview.

The psychological literature on consistency pressures may help to put the relevant survey phenomena in perspective. In conducting such a review, it is also important to be alert for tendencies opposed to consistency, where subjects for some reason respond differently to related items. A few isolated survey findings suggest such an occasional tendency, and it is useful to try to understand possible reasons for it. Our review of factors disposing respondents toward or away from consistency is necessarily speculative because many of the phenomena are interrelated in a complex manner, and because many issues are controversial or as yet unstudied.

FACTORS DISPOSING RESPONDENTS TOWARD CONSISTENT RESPONSES

A logical first question to ask is the extent to which consistent responses are given when related items are asked at the same time. No simple answer can be made to such a broad question, but general guidelines may be given. Many observed correlations between logically or ideologically connected public opinion items are surprisingly low. In a review more than a decade ago of the extent to which public attitudes on different issues are meaning-

fully clustered, Sears (1968:331) concluded that "abstract organizing principles are hard to find, and policy attitudes cohere in only a vague manner." Sears based his judgment heavily, but not exclusively, on the findings by Converse (1964) that several items that might be considered representative of various prototypical liberal/conservative issue stances had an average correlation of only .16 in the general public (see review by Smith, in Chapter 8, Volume 2).

Since then, there has been some controversy among political scientists as to whether interissue coherence did or did not increase during the turbulent 1960s (Nie and Andersen, 1974; Bishop et al., 1979; Sullivan et al., 1979) and to what degree the public can be considered ideologically oriented (Converse, 1975; Nie, Verba, and Petrocik, 1979). Bishop, Tuchfarber, and Oldendick (1978:266) argue that apparent changes in ideological consistency may be artifacts of changes in question wording, so that "had . . . Converse (1964) had access to survey data from 1956 and 1958 that used the 'improved' 1964 items, . . . [his] basic conclusion might not have changed, but it would have been tempered."

Even with a tempered conclusion, however, one still comes away from this literature impressed by many notably small correlations between seemingly ideologically related issues. Carmines and Stimson (1980) argue that these findings result because there are "hard issues," whose implications are adequately processed only by politically sophisticated people, and "easy issues," which have become accessible to most people by long public exposure and (often heated) discussion. Some of the relatively higher interissue correlations found by Nie and Rabjohn (1979) seem to involve "easy issues," such as civil rights, abortion, and the proposed Equal Rights Amendment. Thus, a public opinion analyst may tend to overestimate the degree to which the public draws connections between different issue positions. That opposition to nuclear power ought to imply support for alternative energy sources or for conservation measures may be logical to activists and to academics, but for many people these various opinions may be embedded in disparate mental structures: for example, nuclear power may be associated with nervousness about its effects, alternate sources of energy may be connected with doubt about their effectiveness, and conservation may relate to personal unwillingness to change one's life style. This and other examples suggest that what really constitutes consistency between opinions is often slippery and depends on point of view. A survey respondent may not perceive what is supposed to be consistent with what; in that case, the operation of consistency pressures would be moot.

Even though consistency pressures do not seem to automatically exist between conceptually related items, it is possible that such pressures can be generated by factors in the response situation. Anything that highlights conceptual relationships between items may induce respondents to respond

more consistently. The social psychological literature on relationships between attitudes and behaviors is helpful in suggesting ways that such highlighting might occur.

This attitude-behavior literature addresses a very striking problem in response consistency: namely, that measures of attitude and behavior are often not consistent (Abelson, 1972; Schuman and Johnson, 1976; Wicker, 1969). Thus, attitude toward religion is only weakly correlated with church attendance; attitude toward blacks is not a very good predictor of behavior toward a specific black in an experimental laboratory encounter; attitude toward energy conservation is a mediocre clue of a person's zeal in lowering thermostats; and so on. Some analysts (for example, Ajzen and Fishbein, 1977) have argued that weak relationships arise when the objects of attitude and behavior do not "correspond": religion as an attitude object is an abstraction, while church attendance is a very specific behavioral act.

To predict specific behaviors, it is argued, the researcher should elicit attitudes toward those specific behaviors. To predict church attendance, for example, attitude toward churches would serve better than attitude toward religion and still better would be attitude toward church *attendance*. When attitude and behavior objects are merely conceptually related rather than conceptually identical, there are too many intervening links that may be weak. People can assert strong religious belief and manifest spotty church attendance without necessarily being inconsistent.

Research has shown, however, that if a person's attention is called to a behavior-attitude discrepancy, change of behavior (or of attitude) may result. Indeed, the results of several recent empirical studies support the straightforward idea that focusing a person's attention at the time of attitude or behavior measurement will tend to close the attitude-behavior gap (Wicklund, 1982). Snyder and Swann (1976) show that simply asking people to think over the practical implications of their beliefs about a general issue (in this case, affirmative action) for a few minutes before filling out an attitude questionnaire substantially improved the correlation of that attitude measure with a related later behavior (a decision in a concrete affirmative action case). Snyder (1982) notes that the "action component" of an attitude —a person's preparedness to translate belief into appropriate action—can by no means be presupposed to be automatically operative; people must develop such a component and be reminded of its applicability. In a somewhat similar vein, Schwartz (1977) has explored reasons why people who have abstract altruistic values do not necessarily behave helpfully in emergencies. One important extra condition is that the abstract altruists must become aware of the consequences of their intervention or lack of intervention—they must have the imagination to relate quickly their value systems to the potential fates of victims in situations suddenly encountered.

Just as the connections between attitudes and behaviors often seem to

need spelling out before people are able to act consistently, so too the connections between different attitudes may need to be spelled out before survey respondents give consistent responses. In studying consistency effects in surveys, therefore, one should ask what factors in the interview situation might cause respondents to focus on conceptual relations among items and so adjust their responses to be more consistent. (It is an open question whether one would wish to interpret such adjustments as artifacts or as legitimate products of self-reflection by the respondents about their own values, but one would certainly want to be able to distinguish those cases in which it occurs from those in which it does not.)

Mere contiguous presentation of implicitly related pairs of items does not seem adequate to prompt subjects to adjust their responses in accord with that relationship (except in special cases, noted in the next paragraph). Some research has investigated the so-called "Socratic effect," in which premises sufficient for an unstated logical implication are elicited for agreement by subjects, who then tend to endorse the implication more than subjects not given the premises (McGuire, 1960). An example of two related premises might be "Ocean water near inhabited land has become more polluted this year" and "Swimming beaches are likely to be closed when the water is polluted." The implication is "Ocean swimming beaches are more likely to be closed this year." In contrast to formal logic, in which the implication is instantly entailed by the premises, the psychological process of getting the point takes time. Laboratory findings on the Socratic effect show that the likelihood that the implication will be endorsed is greater after 10 minutes than it is immediately after exposure to the premises (Watts and Holt, 1970).

In general, subjects in psychology experiments and respondents in survey interviews seem to operate very sluggishly in the face of what may seem very obvious logic to the eye of the researcher. The drawing out of inferences is something that takes concentration and time to think. Indeed, explicitly providing subjects with time to think about an experience they have just had or a stimulus they have just seen produces marked and systematic effects on their judgments about it. The major experimental finding is that the opportunity to think about an object or event results in the polarization of attitudes toward it (Tesser, 1978). In other words, objects that are moderately favorably evaluated when a subject has no opportunity to think about the answer become more favorably evaluated given a few minutes' thinking time. Similarly, moderately negatively evaluated objects become more negatively evaluated after brief thought. This polarization effect seems to be greater when experimental subjects have a fair degree of prior knowledge about the objects being judged. Tesser and Leone (1977) refer to this phenomenon in terms of the "schematization" of the attitude

object. When there is a strong schema concerning an object, thinking about it tends to recruit evaluatively consistent cognitions, increasing the polarization of an attitude. When there is a weak schema, a subject's thought is more diffuse, and the evaluative consequences are not so clear. In one experiment, male subjects showed the thought polarization effect for judgments about football plays, and female subjects showed thought polarization when considering dress fashions—but not vice versa. Tesser and Leone contend that this finding showed there is a strong schema about football for males in general and about fashion for females in general.

The notion of schema is somewhat vague in social psychology although currently much in vogue (see Taylor and Crocker, 1981). The general intent of the idea is clear enough, however: if provided with clear and compelling general interpretive structures (variously called "schemas," "scripts" [Schank and Abelson, 1977], or "frames" [Goffman, 1974; Minsky, 1975]), people will tend to process subsequent information in terms of them. How an issue is framed, in other words, will influence how it is responded to if the frame is powerful enough (Fischhoff et al., 1981; Tversky and Kahneman 1981). For example, in studies about risk, it has been shown that if a smaller certain loss is called "insurance," it is much more likely to be preferred to a larger uncertain loss, than if the same terms of risk are offered without this label (Hershey and Schoemaker, 1980).

Many possibilities for framing effects in the wording of individual survey items come to mind (for example, the use of the phrase "forced busing" on the issue of school busing). What concern us in the present section, however, are schematic or framing effects in which one item creates a conceptual context within which later target items are interpreted. While it is impossible to give an explicit catalogue of instances in which this will happen, we can outline four general factors governing the extent of the phenomenon: (1) the relevant schema is obvious and familiar to the general public; (2) the schema has been well elaborated within the respondent's social group or by the respondent; (3) there is opportunity for the respondent to think through the implications embodied in the schema (especially when the first two factors are weak); and (4) the target item is ambiguously stated, thereby inviting interpretations from the surrounding context.

Certain known examples of nonsampling survey errors seem to be explainable in terms of these factors. (Some of the examples of order effects given in Chapter 5 are pertinent.) The norm of mirror image fairness illustrates a strong schema (factor 1): for example, having first endorsed the right of American reporters to have free access to news in Russia, respondents are subsequently more likely to accord Russian reporters the same right in America than if respondents encounter the Russian reporters item first (see Table 5.5).

FACTORS DISPOSING RESPONDENTS AWAY FROM CONSISTENT RESPONSES

Another set of psychological factors governs cases when the answer to one question differs to an unusual extent from the answer to a previous question. The term "contrast effect" denotes the tendency for a response to show a marked disparity from some available benchmark. In this section, we discuss four possible types of contrast effect, giving them names to help distinguish among them: perceptual contrast; response contrast; interpretive contrast; semantic contrast.

Perceptual Contrast. Perceptual contrast effects are very well known in the field of psychophysics. (Psychophysics deals with how people subjectively appraise the intensity of various objective physical stimuli.) In one common kind of experiment on magnitude estimation, a subject assigns numbers to the intensities on a particular dimension of each of a series of stimuli (for example, the brightness of a light). Such response assignments depend not only on the objective stimulus, but also on several factors in the experimental situation (Poulton, 1968). A description of one such study will give the flavor of such research and its implications for survey design.

One design parameter is how much of the range of possible stimuli to sample; with brightness, for example, one could go from the threshold of visibility to the threshold of eye damage or restrict oneself to the middle range. Research shows that as the range of stimuli presented expands, so does the range of responses that people use, although it does so at a slower rate. As a result, when a larger stimulus range is used, people appear less responsive to any given absolute change in the intensity of the stimulus. Both the relative judgments given to two stimuli and the absolute judgment of each of them will be affected by a change. One cannot avoid this issue by presenting only one stimulus because it is likely to be judged in the context of some implicit range of possible stimuli (Poulton, 1977). Nor can it be avoided by using a verbal rather than a numerical response scale, for those scales, too, have implicit endpoints for the highest and lowest response an individual will use—and these endpoints will change more slowly than the endpoints of the stimulus range.

The range of stimuli available for comparison influences not only judgments of physical intensity, but also judgments about social and political objects (Sherif and Hovland, 1961). For example, the judged seriousness of moderately serious moral offenses is diminished when they are embedded in a context of highly serious moral offenses (Upshaw, 1969a; Pepitone and DiNubile, 1976). The reference stimuli against which the given stimulus contrasts might even be implicit rather than explicit. For example, one might see two children as quite different when viewed in the context of a

school class, but as quite similar when placed in the context of a cross-section of the population (see, for example, McGuire and Padawer-Singer, 1976). Or, if you ask a respondent how much he likes a poor political candidate from the party he favors, he might respond differently if his implicit reference group is the set of scurrilous other-party opponents or the illustrious alternatives from his own party. Thus the research showing that the stimulus-range effects are powerful and pervasive indicates one parameter of perceptual contrast that deserves special attention. Survey investigators must take care to ensure that the range of possible responses presented in a survey matches the range in the real-life context for which they wish to predict behavior.

It is obviously more difficult to control for implicit reference stimuli than for explicit ones, but in both cases, the type of preventive measure that should be taken against perceptual contrast in survey responses seems clear. Whenever feasible, respondents should be explicitly introduced to the full range of relevant stimuli (if one can be defined) before asking responses to any of them.

Response Contrast. We have labeled the phenomenon of contrast against a biased stimulus background perceptual contrast, but it has been alternatively explained as having to do with the way subjects adjust their response scales so as to use the full range of responses available to them (Upshaw, 1969b), which can be called response contrast. In surveys, one especially blatant form of response contrast can occur if respondents are asked to give a series of "Yes" or "No" responses, and they are eager to find an opportunity (for example) to say "No" after one or more occasions of feeling compelled to say "Yes". It is hard to be precise about the conditions under which this would happen, but the central idea is that in a series of part items on a given topic, some aspect is varied in such a way that it is unreasonable to give the same response to all parts. Under such conditions respondents may tend to switch responses earlier than the item content would warrant.

Interpretive Contrast. There are other ways that contrast tendencies may be provoked unintentionally. Here our analysis is particularly speculative because not much research has been done. Our point of departure is a somewhat puzzling order effect on a pair of abortion items, one specific and the other general (Schuman and Presser, 1981). When a general abortion item is preceded by a specific question about abortion under a highly justified circumstance (danger to the mother's health), the general item is less readily endorsed than if it is not preceded by the specific item. It is as though the justified circumstances are exempted from the second question. Schuman and Presser call this a "subtraction effect," which can apply to

part-whole questions: that is, when a part of the issue is queried before the whole issue is raised, the whole-issue question is reinterpreted to exclude the specific part. This we label an interpretive contrast. Why this contrast occurs is unknown, but it may be due to the respondent's sense that it is the intent of the questions or they would not have been posed that way.[3]

Perhaps a stark (albeit hypothetical) example will clarify the point about interpretive contrast. Suppose the following two questions were asked in the following order:

1. How sympathetic do you think the president has been toward blacks?
2. How sympathetic do you think the president has been toward minority groups in general?

Our intuition is that in question 2 "minority groups" will tend to be interpreted as groups other than blacks, or at any rate, as less oriented toward blacks than it would have been without question 1 first. The presence of question 1 changes the meaning of question 2, and thus responses to question 2 might be contextually variable (and especially if the president is seen as differentially favorable to blacks and nonblack minorities).

Semantic Contrast. Semantic contrast refers to the case in which the meaning of two items seems on the surface so different that a response of "different" is made, even though such response is not reasonable on closer inspection. King and Klockars (1982) give an interesting laboratory demonstration in which subjects were asked to rate pairs of concepts mutually on given scales. The critical pairs were concepts that were obviously different in major respects, but did not differ on the given scale. For example, subjects were asked to rate "ocean" and "puddle" on a scale of wetness. A noticeable tendency appeared for ocean to be rated wetter than puddle, but this inappropriate tendency disappeared when there had been a previous opportunity to rate ocean and puddle on a scale of size. It appears that relevant acknowledgment of conceptual differences obviates the apparent pressure to express such differences irrelevantly.

POSSIBLE IMPLICATIONS FOR SURVEY DESIGN

It may be hard to anticipate the appearance of some of the four types of contrast on survey instruments. They may be less frequent and less sizable than possible consistency effects. Nevertheless, it is the better part of wisdom for survey designers to be alert to the possible contextual influences that highly related items may exert on one another. Part-whole items and items arranged in steps along a continuum are especially suspect in terms of contrast effects. Item sets that evoke strong irrelevant schemas or provoke an unusual amount of thinking are suspect in terms of consistency effects. It

is natural, of course, for an investigator to want to query several aspects of each issue. (Indeed, there are considerable analytical opportunities for investigating systematic interrelations within item clusters, as discussed in Chapter 6.) However, perhaps there are ways to mitigate superficial consistency and contrast tendencies within item sets. Distributing related items across segments of an interview would be one way. If such redistribution is found to make no difference in responses, the chances are that there was little or no consistency or contrast artifact in the first place (see Martin, 1980). The advantages and disadvantages of such approaches should be explored and more analytical and empirical attention should be paid to the whole complex area of response (in)consistency.

9.4 The Tasks Posed by Survey Questions

Sections 9.2 and 9.3 have been concerned with how the wording of a question and its context within the interview can influence the way respondents organize their thoughts about the issue in question. Respondents' thought processes also vary according to the type of intellectual problem posed by a question. In this section we briefly consider evidence on two cognitive tasks commonly encountered by survey respondents: those of recall and introspection.

We take as our starting point the conclusion reached by Cannell, Oksenberg, and Converse (1977:307) that "our findings do not support the frequently stated assumption that some respondents are accurate reporters and others are not. Rather, the evidence shows that most respondents report some information accurately and some poorly." Each survey question presents respondents with one or more cognitive tasks which they are asked to solve. Different types of problems are likely to be characterized by different types of errors which require different solutions, and the psychological literature can illuminate these issues. To the extent that survey questions represent problem-solving tasks, then factors relevant to good task performance become relevant to surveys. Thus, respondents must be motivated to expend the necessary effort to provide complete and accurate answers to difficult questions. Questions must be carefully worded and structured so as to make the task manageable for respondents. Finally, respondents may need time and freedom from distraction in order to concentrate on the task at hand.

The more difficult or ambiguous the problem posed by a survey question, the more likely respondents are to reinterpret or simplify the question so that they can answer it in the allotted time. There are some questions

which cannot be answered adequately in survey interviews. Including them in a survey can have the pernicious effect of undermining respondents' motivation and effort on other questions. If a respondent is confronted with a complex and difficult question which he is expected to answer quickly, without much thought and in terms of a single, simple response category, he might reasonably infer that the survey investigator is not really interested in serious answers to the questions posed. Evidence suggests that the cognitive tasks of recall and introspection are not handled particularly well in surveys. (The discussion in this section is based on Appendix H, which presents a heuristic classification and analysis of survey questions according to their task characteristics.)

RECALL TASKS

Respondents in surveys are frequently asked to report whether or not some event or condition (such as being victimized by crime or hospitalized for illness) has happened to them, and if so, when it happened. Evidence now accumulating suggests that the task of recalling if and when events occur is far more difficult than survey researchers have typically assumed. Real world events appear to be forgotten rather rapidly and at a fairly constant rate (see Linton, 1975). This appears to be true even for those events that one intuitively would expect to be memorable. For example, one study in which victims known to police were subsequently interviewed showed that the proportion of crime victimizations correctly reported in the actual month of occurrence (according to police records) varied from .70 for events 1 to 3 months prior to the interview to .46 for events 10 to 12 months earlier (Penick and Owens, 1976:23). If remote events are recalled, they tend to be remembered as having occurred more recently than they did. Both phenomena—the tendency to forget remote events and to report them, if they are recalled at all, as occurring more recently than they did—contribute to what is known as recency bias. Recency bias is well documented in both laboratory studies and in surveys. (See Biderman and Lynch [1981] and Cannell, Marquis and Laurent [1977a] for evidence from surveys of victimization and health, respectively; see Baddeley and Hitch [1977] and Baddeley [1979] for a review of experimental evidence.)

The problems of describing and explaining the various sources of recall bias and error are complex, and we do not treat them here. (The interested reader can find useful discussions in the references cited in this section.) Based on current knowledge, we can suggest some obvious, if partial, steps to improve accuracy and some promising possibilities for survey methods that would help respondents recall events and place them in time.

First, if reporting accuracy is an important objective of a survey, respondents usually should be supplied with a limited reference period for which

they are asked to report events. The National Crime Survey, for example, asks respondents to report only about incidents of victimization that may have occurred during the 6 months prior to the interview; even a 6-month reference period may be too long for measuring victimization (see Lepkowski, 1981). For this reason alone, questions such as, "Have you ever been a victim of crime?" are worthless if the answers are to be treated as factual. (Of course, questions such as these may be useful for other measurement objectives.)

Second, the task of recalling and placing events in time probably requires more time and effort on the respondent's part than surveys typically allow for. Often, the date an event occurred is the first piece of information a respondent is asked to supply. (This is true, for example, in the National Crime Survey.) At the least it seems advisable to ask respondents to place events in time after, rather than before, they have been asked about other, more easily remembered details about an event; such details can provide contextual information which helps respondents reconstruct the timing of an event. In addition, when the intent of a survey question is to obtain complete and accurate information, it should be made clear to respondents what is expected of them. For example, although accuracy is stressed in a letter to respondents for the National Crime Survey, the importance of providing complete and accurate reports is never explicitly mentioned to respondents during the course of an interview for that survey (unless an interviewer deviates from the printed schedule and instructions). For the Health Interview Survey, failure to provide appropriate guidance for respondents probably accounted for the finding reported in Section 9.1 that only half the respondents thought that exact information was required (Cannell, Fowler, and Marquis, 1968). Since then, Cannell and his colleagues have tested different ways of clarifying a survey task for respondents and motivating them to work at it. Their experiments show that by manipulating factors such as instructions to respondents, interviewer feedback, and respondent commitment, one can significantly improve the amount and precision of information given by respondents (see, for example, Cannell, Miller, and Oksenberg, 1981).

In the long run, new and innovative survey techniques are probably needed for eliciting reports about past events. In one study, Baddeley, Lewis, and Nimmo-Smith (1978) found that date was one of the details about an event that was least accurately recalled; they also found considerable variation in accuracy according to the recall strategy or cues used by respondents to place events in time. These findings suggest that temporal accuracy might be improved by asking questions which prompt respondents to use more efficient recall strategies. Along this line, Loftus and Marburger (1983) have found that using a landmark event as a temporal anchor for respondent reports can improve accuracy. (Their study, conduct-

ed on the West Coast, used the first eruption of Mount St. Helens as the benchmark event.) We also note that there have been promising efforts to use concepts and procedures from cognitive psychology as a basis for designing improved survey measurement techniques, but there has not been a great deal of actual research to date. (For a report of a workshop applying concepts and findings from cognitive psychology to survey measurement problems, see Biderman and Moore, 1980.)

INTROSPECTION

There is a fundamental assumption in survey research that respondents can give valid reports of their own subjective states. That is, one usually assumes that respondents are aware of and can report feelings and attitudes. Often respondents are asked to report their motives or reasons for thinking, feeling, or acting as they do, as well.

It might be assumed that when questioned about their motives and other subjective states, respondents base their reports on introspective self-examination. However, some psychologists question the validity of introspective reports, arguing that people do not necessarily have privileged knowledge of their own attitudes, motives, or the causes of their behavior. Nisbett and Wilson (1977) review 31 attitude change experiments and conclude that even when attitudes change (as evidenced by behavioral or physiological indicators), verbal reports of them may not. In a majority of studies, the experimental manipulation influenced behavioral or physiological measures of attitude, but not verbal reports. In five studies for which information was available, verbal reports were uncorrelated with behavioral indicators.

Nisbett and Wilson also conclude that people do not have privileged knowledge of their own motives or the causes of their attitudes or actions. In some attitude change studies, investigators probed subjects to find out more about their thought processes during the experiment. For example, experiments show that a severe initiation or hazing increases subjects' liking for a group (Aronson and Mills, 1959), but subjects did not report this factor as influential and, in fact, denied its influence when it was suggested to them (see Nisbett and Wilson, 1977). In a study of social judgment, Nisbett and Bellows (1977) find that subjects not only deny the importance of factors that can be shown to have influenced their judgments, but also assert the influence of factors that did not influence them at all. Nisbett and his colleagues suggest that such discrepancies occur because people asked to report on the causes of their attitudes or behavior do not use introspection to determine their own motives; rather, they rely on judgments of plausible causes and effects based upon explicit rules and implicit theories of human nature.

Zajonc (1980) offers a somewhat different explanation for the fact that people often cannot verbalize the reasons for their likes and dislikes. He suggests that the salient and notable features of an object are not necessarily the same features that feelings are attached to. Presumably, when one is asked to account for one's preference, features that are salient are most readily called to mind, whether or not they are the true basis for the preference (see also Tversky and Kahneman, 1973).

A second evaluation of the validity of introspective reports reaches more sanguine conclusions than Nisbett and his colleagues do. Ericsson and Simon (1980) assess subjects' ability to verbalize their cognitive processes while solving problems. Although these verbal reports are quite different in character from the verbal reports elicited in surveys, Ericsson and Simon raise several pertinent points. First, they note that introspective reports are less valid if subjects are asked to report about internal states that they do not normally consider (for example, motives); if they are asked to produce general accounts that require them to interpret and generalize across a series of experiences; or if they are asked retrospective questions that require them to reconstruct feelings or thoughts in the past. In such cases, according to Ericsson and Simon, a subject is in effect forced to speculate about his own mental processes. In order to secure more valid reports of actual thoughts and feelings, questions should refer to immediate and particular incidents. Ericsson and Simon note that in areas of applied psychological research in which verbal questioning has a long tradition, subjects are asked about specific events rather than for general information or conclusions.

Other psychologists have argued that subjects provide more meaningful reports if their attention is focused upon themselves rather than the environment. Pryor et al. (1977) and several studies reviewed by Wicklund and Frey (1980) and Wicklund (1982) find that induced self-awareness enhances correlations between verbal self-reports and behavioral indicators of attitude (see also Section 9.3). The phrasing of a question can also influence respondents to direct their attention inward or outward. Salancik (1974) used a sentence completion task to measure motives and attitudes with respect to a college course (for example, "I attend class *because* . . ."). The type of reason given by subjects varied according to the wording of the sentence. When "because" was replaced by "in order to," subjects were more likely to give instrumental reasons instead of intrinsic motives. In addition, attitude toward the course was correlated with different factors depending on how motives were measured.

The findings previously reviewed suggest that respondents are best able to report their actual thoughts and feelings when questions focus upon current, specific, real events in respondents' lives. A more extensive review of literature in psychology and other disciplines would undoubtedly yield fur-

ther insights into the question of how, and under what circumstances, valid reports of subjective states can be elicited from respondents. Here we have only touched on the relevant literature. (The evidence on attitude-behavior consistency is also pertinent.)

Caution should be exercised when interpreting respondents' reports of their own reasons or motives. When asked to explain themselves, respondents may give reasons that seem plausible or socially acceptable, or they may report salient features of a situation or object, without really knowing why they act, feel, or think as they do. These accounts certainly provide valuable data on the perspective or frame of reference that respondents adopt when answering a question, but it is suspect to accept self-reported reasons as valid measures of underlying motives without other evidence.

Lizotte and Bordua (1980) represent an example of a survey analysis that makes questionable use of respondents' reported reasons for their behavior. They used answers to the question, "For what purpose do you own firearm(s)?" to classify respondents as members of a "sporting culture" of gun owners or of a group that owns guns for protection. (About 18 percent of their Illinois respondents claimed to own a gun for sport and 6 percent for protection.) Lizotte and Bordua's conclusions about determinants of subcultural patterns of gun ownership are based entirely on analyses using responses to this question as the dependent variable. They take respondents' stated purposes at face value, and offer no corroborating evidence for, or assessment of, their validity. (For example, do people who claim to own guns for the purpose of sport also have the hunting licenses that would be necessary to carry out that purpose?) We suggest that uncritically accepting respondents' stated purposes or motives as valid and basing a full-scale analysis on them is a risky strategy, at best.

More generally, we hope the reader is convinced that recalling events in one's past and being introspective about one's motives are difficult tasks for people, and so respondents often provide answers that are erroneous or misleading. If respondent reports are to be at all meaningful, measurement procedures must be thoughtfully selected or designed, and the resulting data carefully scrutinized to identify sources of error and bias.

Notes

1. As we discussed in Chapter 8, there are systematic differences between survey organizations in the degree to which respondents give substantive (versus "don't know" or other) responses, perhaps due partly to organizational differences in training and instructions for interviewers (see also Converse, 1976–77).

2. The frame of reference implied by the response categories offered as part of the question undoubtedly plays an even more important role in structuring a respondent's thinking about an issue. Respondents are far more likely to choose among response categories that are explicitly offered than to offer one of their own, so the range of response categories essentially establishes constraints on the range of acceptable responses. (See Chapter 5.)

3. We note, however, that this result is contrary to a similar case for the "happiness" questions (discussed in Chapter 5).

PART V

IMPROVING SURVEY MEASUREMENT OF SUBJECTIVE PHENOMENA

This collection of rules will perhaps appear needlessly complicated if one compares it with the procedures in current use. All this apparatus of precautions may seem very laborious for a science which, up to this point, scarcely demanded from those who devoted themselves to it more than general and philosophical training. It is certain, indeed, that the practice of such a method cannot have for its effect the popularization of interest in sociological matters. When . . . we ask men to discard the concepts they are accustomed to apply to an order of facts, in order to reexamine the latter in a new way, we cannot expect to recruit a numerous clientele. But this is not the goal toward which we are heading. We believe, on the contrary, that the time has come . . . to spurn popular success, so to speak, and to assume the exacting character befitting every science.

EMILE DURKHEIM

This chapter is a collective product of the panel. It presents our recommendations and is a direct result of our deliberations.

10

Recommendations

We come now to the portion of our report in which we must translate our work into advice. We cannot offer a series of dramatic and novel recommendations; too many capable critics and practitioners have been over the ground before us for this to be a realistic aspiration. Moreover, almost every recommendation must carry a tacit disclaimer: this advice has not been adequately tested. But we can offer a series of recommendations that could significantly improve the survey enterprise.[1]

Our recommendations are intended to advance three goals:

—To create an improved climate for the conduct of polls and surveys;

—To upgrade current survey practice; and

—To advance the state of the art of survey measurement and the scientific use of survey data.

Most of the recommendations we offer pertain to more than one of these goals, but we use this list as a device for organizing the subsequent presentation. We begin, however, with our single most important recommendation, which applies equally to all three goals.

RECOMMENDATION 1: TAKE SURVEYS AND POLLS SERIOUSLY

Survey researchers should take surveys seriously in much the same way that physicists take particle accelerators seriously, astronomers take tele-

scopes seriously, and space scientists take space vehicles seriously. There are, of course, recreational aspects to these tools. Using them for entertainment to a limited degree is no doubt benign. Anyone who wants to can "poll" any others who are willing, and little harm may be done. But, just as one would not endorse the idea that just anyone can play doctor, administering any and every drug, one should not endorse the conceit that anyone can do or interpret a poll or that anybody's poll is necessarily as good as anyone else's. And we protest in the strongest terms when commercial, political, and social action groups carry out bogus surveys as a way of ingratiating themselves to their prospective customers or converts.

How would life be different if this recommendation were generally accepted? There would be fewer polls and surveys that are manifestly inadequate to their stated purpose, ones carried out merely for the sake of a ritual or to postpone the making of a decision. One would not read extravagant or frivolous interpretations of poll results, presented so that a critical reader has no opportunity to make a contrary interpretation. Polls and surveys would be applied to serious social purposes and not used solely for the sake of public relations or the legitimation of decisions already taken. Moreover, the public would be supportive of competent and careful exercise of the art of surveying or poll taking and capable of discriminating between standard and substandard performance. And, finally, resources would be husbanded so that fewer surveys and polls would provide more and better information.

We have not addressed the issue of how to implement the recommendation to take surveys seriously. But the reader should not jump to the conclusion that regulation, coercion, or other forms of ordering and forbidding are the only way. We intend, more deviously, to change people's minds and habits by insinuating better practices as models for the conduct, reporting, and use of surveys. In our subsequent recommendations, we will propose measures that amount to taking surveys more seriously and count on competition to eliminate, in the long run, the inferior practices.

10.1 Improving Public Understanding of Surveys and Polls

In most of our recommendations we are seeking better surveys and polls that will be more effectively used and therefore enjoy greater public confidence. In this section, however, we are concerned with the climate within which these improvements must come about. There are, of course, many interested parties: the various sectors of the survey industry itself (govern-

ment, commercial, and academic); the many kinds of users, (journalistic, political, governmental, scholarly, and others); and the diverse groups addressed by the users and the industry. We neither anticipate nor desire an undifferentiated approval of or participation in all surveys. On the contrary, there must be vigorous debate, constructive criticism, and responsive actions concerning issues of purpose, content, and method in surveys. What we would hope for, however, is an elevation of the plane of discussion. Fortunately, we can be quite concrete about what we are seeking, given the plentiful supply of published commentary on surveys and polls.

We note first a good example of an attempt to educate the public, an unsigned article entitled "Opinion polls are [ ]accurate, [ ]slanted, [ ]don't know," in *Changing Times* magazine (March 1980). The article's lead reads: "As the political polling season swings into high gear, keep this rule in mind: Don't believe everything you read. Be especially suspicious of polls by groups with a definite point of view." The message is that the reader must avoid "uncritical acceptance of a single study." The article identifies sources of bias as the survey sample, the sequence of questions, and the questions themselves (as, for example, leading questions). The tenor of the article is suggested by this excerpt:

> To make a decision on the merits of a poll, you need certain basic information: who sponsored the survey; when and how the interviews were conducted—by telephone, by mail or in person (face-to-face polls are the most reliable, mail polls the least so); the exact wording of the questions; what part of the population was surveyed (registered voters or adults over 18, for instance); how big the sample and the subgroup were; which results were based on a subgroup (for example, all those aware of an issue, versus all those surveyed) and what the sampling error was.

We note also other efforts in this general direction (Gallup, 1972; Yankelovich, 1980; Cantril, 1980) including an important attempt to provide guidance to journalists in evaluating polls (Wilhoit and Weaver, 1980). While some of these could no doubt be improved, they are examples of the kind of analysis that we hope will find its way into general circulation magazines and newspapers.

Another generally good example of journalism dealing with this subject is a series of articles by E.J. Dionne, Jr., in the *New York Times* in 1980. Somewhat more technical and detailed than the item just quoted, these articles cover a wide range of topics including the effects of question wording on the results of abortion polls, the effects of poll results on political campaigns, the growth of the industry, and the (limited) use of polls in the Soviet Union.[2] In contrast to these constructive appraisals, which are by no means uncritical, we take strong exception to the irresponsible approach

taken in some other widely circulated articles.[3] While making some sound points—such as the fallibility of election forecasts, the problem of nonresponse, and the overreliance of political candidates and office holders on poll results—these articles take an extreme position in advocating that the public refuse cooperation and provide "phony replies" to political polls. Refusal to participate in a poll or survey is, of course, a right that people can and do exercise. However, indiscriminate and malevolent supplying of "phony replies" would damage all types of research *regardless of its merit* and might ultimately compromise important national measurement progams, such as the monthly surveys that produce national statistics on employment and unemployment.

We hope that in the future irresponsible criticism will be supplanted by frank and more frequent public discussions by survey practitioners of the problems and progress of their profession. Our second recommendation, therefore, is directed to survey practitioners.

RECOMMENDATION 2: ENCOURAGE MORE COMPETENT AND INFORMED PUBLIC DISCUSSION OF THE FAULTS AND MERITS OF POLLS AND SURVEYS

Survey practitioners have a responsibility to educate the public and to disclose fully their methods and findings, so that informed criticism and debate are possible. Indeed, it is in the interest of reputable survey practitioners to encourage such debate in order to protect themselves from irresponsible or false allegations and to prevent the survey enterprise from being undermined by inferior or blatantly dishonest uses of surveys and polls (see Chapter 3). The recent example of the dubious use of a phone-in "poll" conducted by ABC News after the 1980 presidential debate—a "poll" that was criticized by ABC's own pollster (Sanoff, 1981)—is unfortunately not an isolated case. We recommend that the professionals who conduct surveys and work with survey data contribute a significant portion of their professional effort toward the enlightenment of the public.

Such enlightenment would be advanced by the increased discussion in public by survey professionals of the faults and merits of particular polls and surveys. Critical symposia might have a salutary effect if the professional debates in this area were more frequently made public. As we noted, journalists also have an important role to play as critics. The public should be educated on the importance of asking a range of questions on an issue and not relying on any single item as a measure of public opinion.

To further progress toward public enlightenment, our third recommendation is directed to the relevant professional organizations, in particular the National Council of Public Polls, the Council of American Survey Research Organizations, the American Association for Public Opinion Re-

search, and the survey methods section of the American Statistical Association.

RECOMMENDATION 3: SECURE THE AGREEMENT OF THE VARIOUS AGENCIES THAT PUBLISH SURVEY FINDINGS, AND ESPECIALLY OF THE PRESS AND SCHOLARLY JOURNALS, TO ADHERE TO APPROPRIATE STANDARDS OF REPORTING AND DISCLOSURE

Reports of surveys and polls in the media often do not provide adequate information to allow informed evaluation by readers. Our exploratory review (see Chapter 3) indicated that even such elementary information as who sponsored or funded the survey and the wording of the survey questions is often lacking. Another reviewer of poll reports in the *New York Times* concluded (Paletz et al., 1980:505):[4]

> . . . with the . . . exception of polls conducted by [its own survey] unit, the *Times* stories inadequately detailed the technical complexities or the possible deficiencies of most of the polls printed. . . . it is our impression that the way methodological information about polling is reported in the media tends more to reassure than alert the audience about the possible defects of polling data.

Professional codes of practice require that pollsters include in their reports a range of information—including the sponsorship of the survey, dates of interview, sample sizes, complete question wordings, and so forth—but the news media do not in turn report such details to their readers. Yet such information is essential for an informed evaluation of poll reports.[5] We note that several British newspapers have entered into informal agreements setting minimal standards for the reporting of election polls. However, our analysis of poll reports in Great Britain indicates that these agreements do not appear to have had an effect on the reporting of other (that is, nonelection) polls; the British poll reports were woefully incomplete in the majority of cases we examined.

Consequently, our general recommendation has two specific implications: (1) survey researchers in the United States should make collective efforts to secure the agreement of the American media to disclose appropriate details of published polls; and (2) extant agreements in Great Britain (and any other nations having them) to disclose polling methods should be applied to all polls and should be carried out. We note that such agreements presuppose that survey methods and findings are fully documented. This is a presumption that applies to all scientific endeavour and certainly ought to be met in any poll or survey that wants to make a proper claim to public acceptance.

Both survey researchers and journalists must bear responsibility for establishing a climate of opinion that makes such details newsworthy. It

presumes little technical sophistication to point out that a newsworthy element of a report that, for example, "*X* percent of the public supports expansion of the nuclear energy industry" is knowing whether the poll was commissioned by a pronuclear or antinuclear lobbying group. Yet such obviously relevant facts are not always reported to the public. Just as a competent reporter would regard the conflict of interest of a public official as worthy of note, so too, poll reports should provide sufficient information to allow an evaluation of the integrity of polls.

In tandem with the foregoing we address our fourth recommendation to the public.

RECOMMENDATION 4: DISCRIMINATE AMONG SURVEYS ACCORDING TO THE QUALITY OF THEIR RESEARCH PROCEDURES

Users of surveys—those who sponsor, use, publish, or simply read about them in newspapers—should be far more careful than they have been about the sorts of practice they will accept; they should become more selective consumers. While precise knowledge at the frontiers of methodological research is sometimes uncertain, there is a set of elementary norms about how surveys should be conducted. For example, sampling should be designed to guard against unplanned selectiveness, question wording should be carefully examined for special sensitivity or bias, interviewers should be thoroughly trained, questionnaires should be pretested, data analysis should be competent and clear, and survey methods and data should be documented and made available for independent examination.

Given some fairly well-accepted practices and professional review (see Recommendations 5 and 6), users can evaluate, or have others evaluate, the quality of the procedures used to conduct a survey and thus the dependability of its reported findings. When available, consumers should use such information and expertise to improve their critical judgments. Consumers should be skeptical of any survey organization not willing to make its procedures and relevant data available for independent review. Indeed, the potential consumer of the findings from such undocumented surveys ought to consider them suspect. Only by exercising their full discriminatory powers and taking advantage of available professional resources can survey consumers be assured that survey organizations will take all of the steps necessary to maintain practice at the highest level.

10.2 Upgrading Current Survey Practice

There are two distinct desiderata for improving survey practice. One is to elevate the state of the art, so that there can be better surveys in the future

than are possible today. The other is to upgrade current practices, moving their quality closer to the limits imposed by the current state of the art. In fact, the present state of the art is a great deal more advanced than one would imagine from reading the widely published criticisms of surveys of recent years. Authors of many of these criticisms may have missed one of their stronger rhetorical points in failing to hold up the better surveys as exemplars against which to compare the poorer ones.

No doubt some criteria of excellence are in part incompatible with other criteria, so that laying out specific regulations would be a thankless if not hopeless task.[6] But that is only to say that no set of prescriptions can usurp the role of considered judgment in balancing alternative means and ends. We believe that a knowledgeable working party of six or eight people could produce a document specifically oriented to surveys of subjective phenomena comparable to the booklet, "Standards for Discussion and Presentation of Errors in Survey and Census Data" (Gonzalez et al., 1975), based on practice in the Bureau of the Census and other government agencies engaged in fact-finding surveys. To be helpful such a document need not have the force of law; the cited document has only the force of example in regard to nonfederal survey organizations.

Preparation of such a document would be a logical sequel to the work of our panel. The reader could ask why we did not include that task on our agenda. The answer would have to be that we had a much broader mandate that left neither time nor resources to do the painstaking detailed work contemplated in this proposal. But standards are implicit in the suggestions and conclusions of all our chapters, and the literature is not devoid of supporting technical material.

In the next several recommendations, we focus on what may be done to encourage practitioners to upgrade the quality of individual surveys and then consider improving the state of the art. In order that the public may more effectively evaluate the quality of surveys, we offer two parallel recommendations to survey researchers.

RECOMMENDATION 5: IMPROVE THE REVIEW AND EVALUATION PROCESS FOR POLL REPORTS

Survey data are presented in various forms to various publics and the review process is thereby fragmented. Many data are collected for private use only (for example, in-house marketing research and political polls by the candidates). Since those data are not directly presented to the public, there is no external review process, although everyone feels the impact of the marketing and political decisions that are partly shaped by those data.

Of the data that are publicly released, a large body is collected by or reported in the mass media, including the regular privately conducted polls

by Gallup, Harris, and others. Although it appears (see Chapter 3) that public review is spotty and usually lacking in depth, one does, nonetheless, encounter some examples in which poll results are disputed, reanalyzed, and reinterpreted.[7]

A second major vehicle for the dissemination of survey data is scholarly publication. Data collected for academic purposes (for example, the national election studies of the Institute for Social Research [ISR] and the General Social Survey of the National Opinion Research Center [NORC]) are evaluated and reviewed by the agencies that funded the data collection. There are various scholarly guidelines on preferred survey practices and methods, but no authoritative set of standards for design, collection, or analysis of survey data. Of course, journal articles undergo editorial and peer review before acceptance and are subject to general academic scrutiny after publication. There are instances of vigorous exchanges, such as that between Nie and Rabjohn (1979), Bishop et al. (1979), and Sullivan et al. (1979), over the impact of question wording on conclusions about the belief systems of the public (Converse, 1964), and the recent debate between Caddell (1979) and Miller (1979) over the American "crisis of confidence." Social scientists, however, have generally been insufficiently attentive to survey methods. Because of this, even scholarly survey research often does not get as thorough a review as it should (see Presser, 1980; and Chapter 3, Volume 2).

Finally, surveys conducted by the federal government are usually subject to rigorous internal review and evaluation. The Census Bureau in particular carries on a significant program of methodological research designed to improve its measurement procedures. And when published, government survey reports are often subject to scholarly evaluation and criticism. We note, nonetheless, that a recent review (Bailar and Lanphier, 1978) of a small sample of federally sponsored surveys suggested that research practice in this domain is still far from adequate and that considerable improvements should be sought.

In addition to these usual procedures, there are ad hoc evaluations. Occasionally an interested party is so agitated by a survey (or its results) that a special evaluation is launched. Mayor Meier of Milwaukee, for example, had three groups of outside consultants examine a pair of polls conducted by the *Milwaukee Journal* (Gimbel, 1979); Serge Lang launched an extensive letter writing campaign against the 1977 survey of American professors (Lang, 1981); and the *Minneapolis Tribune*, unhappy with the results of its own Minnesota poll, had an evaluation conducted by associates of the Institute for Social Research (Campbell, Heeringa, and Miller, 1979). Each of these is an example of valuable ad hoc evaluations and reviews.

There exists, in sum, a scattered system for evaluating survey data. Often

this system works very well. Careful reviews are carried out and useful criticisms are made. But the patchwork nature of this system also implies that it often fails completely or works only belatedly. And, unfortunately, the book review columns even of scientific journals seem hopelessly wedded to literary criteria.[8] Few surveys per se are reviewed, and there are practically no reviews of polls or surveys that are not in published books. Thus, a stream of survey reports issues from government agencies and such major data producers as the Survey Research Center without undergoing formal and critical external review.

We suggest that the survey research methods section of the American Statistical Association organize a forum for the publication of reviews of polls and surveys. There are interesting precedents. The *American Statistician* now reviews computer programs and teaching materials, and *Contemporary Sociology* reviewed the *Cumulative Codebook of the General Social Survey*. The *Public Opinion Quarterly* and *Public Opinion* magazine routinely present compilations of survey and poll results. These precedents might be expanded to include critical reviews of surveys and polls. Of course no single journal can review all survey reports. But even a highly selective coverage of current output would be valuable. Review articles, as distinct from mere compilations of survey results, would also be very useful.

To assist in upgrading the quality of current surveys, the following recommendation is addressed to the producers of survey data.

RECOMMENDATION 6: MAKE INDEPENDENT PEER REVIEW A PART OF THE DESIGN, ANALYSIS, AND REPORTING OF SURVEYS. IF VOLUNTARY REVIEWS ARE NOT AVAILABLE, THEN A PORTION OF THE RESOURCES OF THE SURVEY SHOULD BE ALLOCATED TO OBTAIN PAID REVIEWS[9]

Peer review is a feature of all science. However, the applied nature of survey research and the institutional structure of the industry may make competent reviews difficult to obtain. Thus, many surveys are not reviewed at all within the scientific community. Such public commentary as is generated often comes in political arenas, in which the participants may be less concerned with the quality of measurement than with the political implication of a survey.

One alternative is to have independent review of a survey be made a standard budgeted expense. Review should come both early in the process, so that the design of the survey can be improved, and late, so that the most honest reporting possible will occur. A survey that followed such a procedure might receive some special designation (as do those that abide by

standards of disclosure of the National Council of Public Polls). This might encourage those who sponsor surveys to undertake the additional expense.

Although there are obvious problems with ensuring the independence of reviewers who are on the payroll, some reasonably sound arrangement might be made with the help of existing institutions or the independent center whose establishment we recommend, (Recommendation 13). In any case, we believe that more vigorous institutionalized self-criticism within the scientific and survey research community can be an effective way to regulate the quality of survey research.

Election polls are in particular need of continuing review, because of their potential significance to candidates, voters, and the survey research community, and we therefore address our next recommendation to the relevant sectors of the professional survey research community.

RECOMMENDATION 7: ESTABLISH A PANEL OR COMMITTEE TO EVALUATE THE PERFORMANCE AND METHODOLOGY OF ELECTION POLLS

The history of election polling reveals periodic public criticisms of the accuracy of election forecasts (see Chapters 1 and 2). This dates back at least to the *Literary Digest* fiasco of 1936. In 1948, a public outcry followed the erroneous "election" of President Dewey by the polls (the Gallup and Crossley polls predicted Truman would lose with 44 percent of the popular vote, but Truman won with almost 50 percent of the popular vote). This led to a specially commissioned study by the Social Science Research Council (Mosteller et al., 1949). This example has not, however, been repeated, even though the raw errors in predictions have exceeded in some cases the discrepancies between the polls and results in the 1948 presidential race. For example, in the 1978 elections, several governors and members of Congress "elected" in the polls went down to defeat in the election. Similarly, in the 1980 presidential election, Ronald Reagan's substantial margin of victory over President Carter (51 to 41 percent) was not anticipated by most of the public polls (Mitofsky, 1981).

We were not charged with reviewing election forecasts, and we have not, therefore, devoted detailed attention to this topic. We believe, however, that a regular review of the accuracy of such forecasts could be of use both to the survey industry and to the public. We note in this regard the example set in Britain, where election post mortems have been held under the auspices of the Royal Statistical Society. Such evaluative activities can upgrade survey practice and public understanding of survey results. A committee appointed to evaluate American election forecasts might well function under the institutional auspices of the center whose establishment we propose below (Recommendation 13).

Allocating Resources. Many measures that might result in the improvement of surveys are technically feasible and economically in reach without major new expenditures for survey research. Of course, such expenditures might be justified on other grounds, primarily the need for more information or information of a different kind than is now available. We have not tried to assess such a need, but we do strongly feel that increments to the quantity of survey research must not be purchased with decrements to its quality. Improvements in the quality of surveys do not inevitably require increases in total costs. In this regard, we address our next recommendation to survey sponsors and survey organizations.

RECOMMENDATION 8: MAKE A BETTER ALLOCATION OF RESOURCES FOR SURVEYS

It is quite possible there are too many surveys. Certainly there are too many surveys of mediocre and inferior quality that do not accomplish anything useful. There are too many that report on matters of only the most transitory or parochial interest. There *may* be too many in the light of their burden on respondents, but we do not know. If any of these situations exist, we join others in deploring the state of affairs, without having any general remedy to offer other than relentless publicity for substandard—and superior—practice in the professional and public media. Respondents must regulate their own burden by choosing not to respond to some or all surveys. Evidence suggests that respondents' refusal to participate in a given survey is associated with their (stated) perception of the value of surveys in general rather than their perception of the value of the particular survey in which they are asked to participate (Frankel and Sharp, 1981:110). It is thus to the advantage of the survey profession to decrease the frequency both of substandard polls and of sales and direct mail solicitations that masquerade as surveys.

More specifically, our recommendation implies that each survey organization should re-examine the ways it spends its money. In addition, each agency that contracts for or funds surveys can adopt policies to improve the pattern of resource allocation. In particular we urge survey organizations to consider the following propositions:

1. There is too little methodological work of the right kind;
2. There is too little investment in pilot studies and auxiliary investigations.

Or, to state the point in a different way, there are now many small-scale studies with a few hundred respondents that are manifestly inadequate for making serious estimates of important quantities and relationships but that are splendid as pilot studies. The problem is that they are not treated as pilot studies but are regarded as completed investigations. Apart from can-

didate preference polls and single-question referenda, serious surveys are seldom interested in merely estimating a single proportion. The interest is in comparisons of population strata and in multivariate relationships. The samples used for these purposes are often much too small—appropriate enough for pilot studies but not for the supposed substantive purposes of the research.[10] In academic research, one of the few instances in which this has been acknowledged is NORC's General Social Survey. Its annual sample size is only 1,500, but most of its questions are repeated year after year; hence, the sample size accumulates gradually so that serious analyses ultimately become feasible.

In addition to too little methodological research and too little investment in pilot studies, we also assert that:

3. There is far too little analysis of the typical survey, at least from the point of view of basic research, whatever may be said concerning applied uses.

It is well known to investigators with a major commitment to survey research that most of the data collected are never closely and searchingly examined. The ratio of analysis to collection is too low, and this is particularly distressing because analysis is relatively inexpensive compared with the costs of data collection. One obvious cause of this misallocation within an organization has to do with the organizational necessity of continually undertaking new surveys in the interest of maintaining cash flow. How this can be overcome is more than we can say, but it is a matter requiring serious attention.

From a social point of view, too many resources and too much of the public's time is spent collecting survey responses and too little are spent in finding out what the responses mean. This panel does not agree with one commentator's observation on the preliminary report of a survey that: "Sociologists will be spending the next year or so determining what these and other apparent trends really mean. But the raw data give the layman a pretty good idea right now."

Our whole report speaks to the point that it is a lot more difficult than it seems to turn raw data into warranted conclusions. The task, however, is worthy of our best efforts, and hence we make the following recommendation to all survey organizations.

RECOMMENDATION 9: BUILD INTO VIRTUALLY ALL ONGOING PROGRAMS OF SURVEY DATA COLLECTION AND RESEARCH ASSESSMENTS OF ERRORS AND COLLECTION OF DATA RELEVANT TO METHODOLOGICAL ISSUES

We distinguish carefully between this recommendation and the usual one that calls for more methodological research. A report by Storer of a confer-

ence on survey applications of social psychological questions, for example, noted "the pervasive tendency among social scientists to specialize in substantive areas, which tends to lower their enthusiasm for methodological research" and went so far as to suggest (Storer 1969:6):

> . . . [it might be] efficient to divide organizational responsibility for . . . various types of social research, either through upgrading private, nonprofit organizations by providing stable financial support, or through the establishment of another separate survey research agency in the Federal government. In the short run . . . the essential need is for increased support of methodological research through whatever organizations are willing and able to conduct it.

The suggested division of responsibility is precisely what we think is *not* "efficient." It is, on the contrary, a recipe for ensuring that the results of methodological research are never used and, in particular, are not used to enhance the meaning and usefulness of actual surveys. An organization interested in cutting corners can too easily assent to the solution of letting someone else work out the methods and then find, with a show of regret, that the methods they developed would never work.

The separation of method from substance in the social sciences is intellectually stultifying as well as economically costly. It has resulted in a self-sustaining "methodological" literature read only by methodologists discussing methods that have never been shown to work because their sponsors have never been responsible for solving actual substantive problems. Successful methodological development has almost always involved intimate interaction between methods experts and substantive investigators (sometimes an interaction going on within a single cranium).

Of course, some methodological questions will require intensive and expensive investigations that might well be carried out separately, perhaps under the auspices of the center recommended below. However, we protest the idea that methodological work is in general a costly overhead or luxury for the survey organization. If anyone today were to suggest that calculation of sampling errors is not a part of a respectable survey but should be left to separate "methodological studies," he or she would not receive a sympathetic hearing. Among professional survey researchers, it is taken for granted that a survey is not complete until estimates of sampling variation have been computed. Sooner or later survey researchers and survey users must accept the equally compelling argument that estimates of nonsampling sources of variation are indispensable for a competent survey, and then the bifurcation of substance and method will be seen as equally absurd.

To be concrete: If there is an issue as to the wording of a question, with several plausible alternatives, the obvious way to resolve the issue is with

experiments trying the alternatives. The outcome can be either that the wording does not matter (that is, essentially the same results are obtained with several different wordings) or that the wording does matter. If the former, the inclusion of both wordings in the main survey (not in some separate methodological experiment) strengthens the conclusions. But if the wording does matter, then surely a reliable estimate of the wording effect is required, not merely for some "methodological" purpose but to make a complete report on the substance of the investigation.

Methodological experts should be asked to provide something they are perfectly capable of providing, to wit, survey designs[11] that build in internal estimates of nonsampling as well as sampling variation without compromising the substantive goals of the study (see Chapter 5 for examples of these techniques). We note that such procedures may become quite inexpensive to implement in the future as more survey organizations adopt computer-assisted telephone interviewing. We also point out that the results of such factorial designs can be conveniently handled using the modern survey analysis procedures we recommend (see Recommendation 14).

We do not question the need for preliminary and auxiliary studies to develop new questions or procedures, but we believe that such studies will be most useful when they are part of the process of designing serious substantive surveys. Even more important, they cannot be relied upon by themselves to provide the needed estimates of sampling and nonsampling variation. Those estimates must come from the main survey itself. If making them is a methodological activity, then our new motto must be: "Every survey a methodological inquiry."

Our next recommendation is also directed toward the improved understanding of the factors, particularly nonsampling ones, that cause variation in survey results.

RECOMMENDATION 10: PROMOTE EFFICIENT COOPERATION BETWEEN SURVEY ORGANIZATIONS

Just as routine methodological investigation should be a normal part of the internal working of any survey, so should cooperation among agencies be seen as a normal means of improving the general level of performance and state of the art. We are not, of course, advocating any mandated sharing of proprietary information or any legislated uniformity in topics studied. But when different organizations have different approaches, it is in the public interest to show how those differences matter. Over the long term, a modest program of between-house experiments could produce "item banks" of questions with known properties. We are here drawing on the experience of the Social Science Research Council's (SSRC) Center for Coordination of Research on Social Indicators, which convened a working group composed

of representatives of academic, commercial, and federal survey agencies. That group reached a consensus on a core set of items that provide adequate definition of important background variables with minimal investments of interview time and coding effort. The results are in a manual, *Basic Background Items for U.S. Household Surveys* (Van Dusen and Zill, 1975). We do not suppose that agreement concerning subjective items will be so easy, and our point is not to aim at uniformity of practice. It is rather to secure *calibration* among organizations when, for sufficient reason, they find it advisable to go their separate ways. The SSRC experience suggests that modest expectations of cooperation are not unreasonable. The panel's own experience with small-scale pilot studies involving planned between-house comparisons is further evidence to this effect (see Chapter 5). Similarly, the recent effort to coordinate research on computer-assisted telephone interviewing is a precedent for the sort of cooperative venture we are proposing. The development of coordinated measurement programs also finds ample precedent (for example, Youden and Steiner, 1975) in attempts by physical and chemical scientists to bring their decentralized analytical measurements under statistical control.

There is room for considerable action by those who commission surveys as well. Organizations that contract for large-scale surveys of subjective phenomena can both build into such surveys methodological investigations (of the type noted previously) and split the field work contracts for surveys between two or more commercial houses. The variability in the resultant measurements can provide a useful insight into the effects of uncontrolled differences between organizations in such things as interviewer staffs, training, coding, and so on. They can thus provide useful information on the range of variation (above sampling error) that might be anticipated to be produced by fluctuations in organizational factors when there is no true change in the population—a necessary criterion for inferring change when surveys are subsequently replicated.

Cooperative ventures such as these sorts might well hold an important place on the agenda of the center we propose (Recommendation 13).

10.3 Advancing Survey Measurement and the Scientific Use of Survey Data

Our assessment of the present state of the art of survey measurement and analysis is that it is ripe for improvement. If survey researchers take advantage of the growing body of scientific knowledge and recent advances in statistical methods in order to improve their measurement and analytical

techniques, significant improvement in the quality of survey research can be expected. We also find that the field is on the threshold of a far more thoroughly scientific use of survey data that can advance theoretical understanding of the dynamics of public opinion and public sentiment.

In the following recommendations we first consider what should be done to improve the state of knowledge and next suggest steps to advance the scientific use of survey data. The recommendations in this section are addressed to all those who are now involved in the survey enterprise and to some whom we invite to join it.

RECOMMENDATION 11: BROADEN THE BASIC SCIENCE COMPONENT OF SURVEY DESIGN AND PRACTICE

A great deal of what is done in a survey organization is based on tradition and on informal prescriptions for getting the job done. Alternatives are not often systematically considered, nor are there explicit justifications for many practices. Only some of the disciplines with potential contributions to make to survey research are actually involved in an intimate way; others need to be involved and exploited (for example, cognitive science and anthropology).

What can happen when a fundamentally scientific approach is taken is illustrated by past achievements in survey sampling, which during the 1930s and 1940s began a sustained evolution toward a coherent and powerful body of theory translated into codified practice (Neyman, 1934, 1938; Deming, 1950; Yates, 1949; Cochran, 1953; Hansen, Hurwitz, and Madow, 1953; see Stephan [1948] for a history of early use). Other parts of the survey enterprise have, to be sure, been subject to careful study. Just a few monographs are noted here (Hyman et al., 1954; Bailar, 1975; Bradburn et al., 1979; Groves and Kahn, 1979; Schuman and Presser, 1981), but there is not yet a body of knowledge on other topics comparable in scope and rigor to the works on survey sampling.

We do not believe, moreover, that such a body of knowledge will be developed if reliance is placed exclusively on the efforts of the political scientists, social psychologists, sociologists, and statisticians who are providing most of the basic scientific input to survey work at the present time. We expect anthropologists, cognitive scientists, linguists, and perhaps scholars in various branches of the humanities, among others, to make novel and productive contributions. As a first step, we would like to see the serious and formal involvement of some such mix of scholars in the conduct of a sustained program of survey research on a topic of common interest.

It should be emphasized that benefits from such cooperation would not flow in only one direction. The survey environment provides many unique opportunities for scientists. For example, the survey context could provide

experimental psychologists interested in the study of memory and cognition with a rich real-world laboratory complete with large random samples of the adult population (rather than the college students usually used as subjects in such experiments).

We have an implicit agenda in calling for this program. If successful results were forthcoming from such a collaborative research program and were reported, among other places, in disciplinary journals, then other scientists might well be stimulated to follow suit. If a good example were initially set and the field fertile, in time many similar endeavors might bloom from this initial seed.

RECOMMENDATION 12: ORGANIZE A SYSTEMATIC LONG-TERM STUDY OF ALL PHASES OF THE SURVEY PROCESS. THIS SHOULD INCLUDE AN EXTENSIVE INTERDISCIPLINARY INVESTIGATION OF SUBJECTIVE ASPECTS OF SURVEY QUESTIONS.

Even for most people engaged in survey research, the interview is something of a black box—there is the printed questionnaire and interviewer's instruction manual and the codebook and data tape, but little else by way of systematic documentation of how the data were actually generated. A deeper understanding is needed of how this instrument, the survey interview, works. Survey research should have greater intellectual input from other social and cognitive scientists. In Chapters 7 through 9 a preliminary reconnaissance is presented of what some of these fields might contribute to our understanding of the interviewing process. We find much material—including empirical results, methodological innovations, and theoretical constructs—that could be exploited to improve survey measurement, and in particular to improve understanding of the nature of the interaction between a respondent and the interviewer and between a respondent and the survey questions themselves. Collaborative investigations of interviewing that draw upon ideas and methods from all relevant scientific disciplines can deepen understanding of the structure of the errors affecting survey data and can suggest innovations in concepts and methods of survey research.

We offer several possible approaches, without prejudging which (if any) will prove fruitful. First, we have presented (see Chapter 9 and Appendix H) a heuristic classification of the ostensible tasks that survey questions pose for respondents. The analysis of such a classification, informed as well by evidence from psychology, might provide a starting point for studies of how, and how well, these tasks are executed.

A second approach considers the interviewer. Carefully designed studies of the interviewer's contribution to survey error demonstrate that variability due to interviewers can be extremely high (see, for example, Bailey,

Moore, and Bailar, 1978). However, we have little understanding of the interview dynamics. (In Chapter 8 we offer a preliminary discussion of the dynamics of the interaction between respondent and interviewer, which might be expanded into a more systematic analysis of their mutual roles and the effects of role expectations upon the quality of survey data. The inquiry might include the interviewing of interviewers or the systematic use of debriefing.) Our hunch is that interviewers have more wisdom about some of the problematic aspects of surveys than is usually taken advantage of, although they also have the biases of their particular perspectives. Of course such biases are also worthy of investigation.

A third approach is to develop error profiles and models of total survey error (see Chapter 4 in this Volume and Lessler, Chapter 13 in Volume 2).[12] Finally, we note the extreme importance of developing a systematic framework for the interpretation of the effects of changes in the wording of survey questions. Wording (and other) changes sometimes have a substantial effect on response distributions and relationships among survey variables, and yet in general one cannot accurately predict when or explain why such effects occur (see Chapter 5). We believe that theoretical development, drawing on ideas and methods from other scientific disciplines is required in order to better understand the sources of meaning and frames of reference that respondents employ in answering survey questions.

RECOMMENDATION 13: ESTABLISH A CENTER FOR THE STUDY AND EVALUATION OF SURVEY METHODOLOGY

We believe that methodological research on subjective measures will be most fruitful when carried out as part of significant substantive investigations. However, there will be times when concerted methodological work may be called for. Much of the early understanding of interviewing comes from the large-scale project reported by Hyman et al. (1954); appreciation of question wording problems was greatly stimulated by Cantril (1944); and concern for interviewer variance in attitude measurement was heightened by the early work of Kish (1962).

An independent center for survey methodology (or a consortium of cooperating centers) could provide a locus for future interdisciplinary research on survey methods (see Recommendation 11) and for long-term intensive investigations of the sort suggested in Recommendation 12. Such a center might also facilitate cooperation among survey organizations (see Recommendation 10). In addition, a constructive role in improving the review process (see Recommendations 5 and 6) might be played by an independent center set up for the study and evaluation of survey methods.

One of the most effective rebuttals to flawed or biased polls is the re-

porting of the results of less flawed (or less biased) polls. We have noted (see Chapter 3) the salutary example of a North American Newspaper Alliance poll conducted in rebuttal to a *New York Times*/CBS poll claiming a political realignment in the United States. Simiarly, we note examples of criticism (Schuman and Presser, 1980) in the popular media of a poll on support for a constitutional amendment to ban abortions; of a critical review of polls on nuclear energy (Mitchell, 1980a, 1980b); and of a critique of polling on the second Strategic Arms Limitation Treaty (Lanouette, 1980). An independent center for the study and evaluation of survey methodology might serve as a forum for debate on the merits of particular polls and surveys, and it could provide institutional resources for the commissioning of both independent polls and critical reviews such as those discussed previously. The dissemination of such work might foster more informed public discussion of the faults and merits of polls (see Recommendation 2).

RECOMMENDATION 14: MODERNIZE SURVEY ANALYSIS

In various parts of this report and in our volume of supplementary studies we have furnished illustrations of models and statistical methods that offer improvements over conventional models and methods in survey work. The branch of statistical methods most relevant to the data collected in surveys —those dealing with categorical data—was markedly underdeveloped up to a decade ago.[13] But times have changed. The major textbooks of Bishop, Fienberg, and Holland (1975), Goodman (1978), and Haberman (1978–79) record a development almost as remarkable in its way as the evolution of sampling methods that we noted previously. The dissemination of these tools to the survey practitioner, together with a heightened awareness of the need for such multivariate strategies, would be a significant advance. Moreover, these analytic methods have the coincidental advantage of encouraging the use of factorial designs for the study of wording, context, and other effects. By providing a convenient vehicle for their analysis, these techniques facilitate a forthright confession by survey practitioners of the fundamental truth that responses often depend on how a question is asked. This fact, which is ordinarily hidden from public view, can and ought now to be routinely incorporated in survey design.

In addition to increased use of categorical analysis techniques, we suggest graphical analysis and exploratory data analysis techniques (Tukey, 1977; Mosteller and Tukey, 1977). The use of these techniques may often lend additional support to the interpretation of results, but they also may point out anomalies in the data. The use of these techniques may yield insights not provided by other analytic procedures.

These domains of statistical methods continue to be active, with impor-

tant new work being reported regularly. If surveys fail to improve, the fault will not lie with the statisticians, for they have given the survey craft some very powerful tools indeed. If practitioners continue to suppose that they can get along with homemade implements, the survey community will have itself to blame for falling far short of its potential.

We offer our next recommendation to all users of surveys and poll data.

RECOMMENDATION 15: EXAMINE PERIODICALLY THE CONCEPTUAL BASIS FOR THE USE OF POLL AND SURVEY DATA IN SCIENTIFIC DISCIPLINES AND APPLIED FIELDS MAKING HEAVY USE OF THEM

This recommendation is not as bland as it sounds. The panel struggled in various ways with the question of what surveys, especially those dealing with subjective phenomena, are good for. We find that the conceptual foundations for much ongoing survey activity are not firm. A great deal in the way of "psychologizing" seems to be involved in the interpretation of survey data, but the psychological theory involved is almost wholly implicit and often incoherent. This is not to say that the theory should be faulted for poor application. We struggled with the ways in which psychological theory might be more effectively brought to bear upon the survey process (see Chapter 9), and this indeed was one of the motivations for our previous recommendation that a long-term collaborative study of the survey process be organized.

The substantive scope of surveys, however, is exceedingly broad, ranging over matters dealt with by all the social sciences. But for many of these sciences, surveys are on the periphery and their conceptual problems do not command attention. We believe that public opinion research has become overly preoccupied with questions that may be answered solely with survey data to the neglect of theoretically central issues calling for other kinds of data or for systematic supplementation of surveys with auxiliary information (see Chapter 7).

One area for consideration is social indicators development, as recorded, for example, in *Social Indicators III* (Bureau of the Census, 1980a), a publication issued too recently for thorough assessment by this panel. But we note that the effort to define a niche for subjective indicators within the broader context of social indicators has attained at least temporary success. In *Social Indicators III*, subjective measures are prominently displayed and editorially justified. We thus have at least a first approximation to what Campbell and Converse (1972:9) asked for: "some direct monitoring of key social-psychological states of the population." The question now is how to assess this accomplishment and to discover what the information is good for.

One must ask whether the tangibility or meaningfulness of such subjective measurements can be substantiated. Too often it is assumed that anything to which people will respond constitutes a measurement. This, however, begs the question of whether there exists something to be measured. It is surely difficult to demonstrate that such measurements fit into some coherent pattern of orderly (theoretical or empirical) relationships. It is not, however, beyond the reach of our current abilities.

Validation provides a fundamental challenge for all who would take subjective measurements seriously, and it is not an inherently unattainable goal. To assist in meeting this basic challenge of scientific measurement we make two recommendations related to external validation of surveys.

RECOMMENDATION 16: PUT SURVEYS IN CONTEXT WITH APPROPRIATE AUXILIARY DATA

Most of our report focuses on matters internal to a survey. But one could argue that the context of a survey is equally important to its effective use. An illuminating discussion by G. Carlsson (1970) takes note of the predilection of sociologists (in contrast with economists) for studies that "deal with variation between units at one point in time, cross-sectional variation, or what might be called differential behavior"—in short, with the one-time survey. Carlsson goes on to note the small likelihood of success in building systematic theory or in understanding historical change on the basis of this research strategy. In a study of shifts in the agenda of public opinion, MacKuen (1981) argues the necessity of a dynamic modeling approach and effectively exploits time series of survey data and indexes of media content in demonstrating the role of the media in bringing about such shifts. A general methodological statement was offered by Beniger, Watkins, and Ruz (1978:118), whose study of trends in opinion on abortion was intended to illustrate the need for social reporting:

> . . . to include measures of discrete events, mass media coverage and public opinion, analyzed both longitudinally and cross-sectionally, in the context of other national problems, and that monitoring the interrelationships among these elements of [social] change, is a central task of effective social reporting.

(Further examples of this work may be found in MacKuen, Chapter 14, and Beniger, Chapter 15, Volume 2.)

We assume a growing stock of survey data (as discussed in the next recommendation) and a carefully planned program of increments to this stock. But even this is not enough, according to the authors just cited. We therefore suggest the "outrigger principle" (which their work adumbrates), in which there are two supports for the main vessel. The main vessel is the

program for repeated cross-sectional surveys to ascertain changes in response levels and differential changes for population subgroups. It is supported, on one side, by studies of media output, organizational activities, political debate, and so forth that provide indicators of the everchanging complex of stimuli that may move public opinion and commitment. On the other side, the main vessel is supported by special-purpose, in-depth surveys of leadership cadres, members of relevant groups, communities where effects may be greater or appear sooner, and of panels, which provide supplementary data on the process of change, calibrated by items shared with the main surveys.

Further supplements to traditional surveys ought to be considered. We note proposals (for example, Zeisel, 1980) for entirely different sorts of surveys that would involve exposure of random samples of the population to extensive arguments about a topic prior to the elicitation of an opinion. We also note the innovative use of laboratory studies as a supplement to surveys of topics that are not frequent concerns of respondents (for example, the risk consciousness in residents of flood-prone and earthquake-prone areas; see Kunreuther et al., 1978). In this work, primitive causal theories as to what led people to buy insurance against such risks were developed from cross-sectional survey data. The theories were then tested in a series of laboratory experiments that afforded greater opportunity for the control of relevant variables in (mock) decision-making tasks. These studies were then complemented by econometric analyses of actual market data.

We do not endorse the particulars of any specific approach as a panacea, but the external validation of surveys will be greatly facilitated by tying the attitudes and attitude changes measured in survey data to their causes—wherever they may lie. Such auxiliary data permit causal analysis of the origins of attitude formation and change. One might then be in the position of examining methodological questions such as intersurvey discrepancies in estimates from the advantageous perspective of models that make predictions (tested by past data) as to the expected behavior of a given time series of measurements. It is possible that some of the discrepant survey measurements that now confound us will appear regular and lawful in the light of substantive theories describing the processes that generate changes in the underlying phenomena that a survey is trying to measure.

In making the recommendation that surveys be placed in context, we are not proposing an enterprise that can be casually undertaken. It is not our intent to further encourage offhand observations that, for example, a dip in presidential popularity coincides with some particular presidential action. (Rigorous testing of the sorts discussed above might, however, turn such speculations into warranted inferences.)

Our next recommendation is aimed at facilitating implementation of Recommendation 16.

RECOMMENDATION 17: MAINTAIN, ENLARGE, IMPROVE, AND EXPLOIT ARCHIVES OF SURVEY AND AUXILIARY DATA

Archives of poll and survey data have burgeoned in recent years; the Roper Public Opinion Research Center, the Inter-University Consortium for Political and Social Research, and the Louis Harris Data Center are prominent examples. Secondary analysis has become an important modus operandi of a large sector of social research and has even been studied as a movement and codified as a technique (Hyman, 1972). The usefulness of archival data, particularly for the analysis of social change, has been advanced in recent years by the publication of indices of repeated questions (Southwick, n.d.; Center for Political Studies, 1976; Martin, McDuffee, and Presser, 1981).

We favor this kind of work, but we also ask for something more, namely, the feedback of secondary analysis into current surveys that might be deliberately designed to illuminate issues arising in the archival record. For example, in June 1969, the Gallup poll asked the following question:

> President Nixon has ordered the withdrawal of 25,000 United States troops from Vietnam in the next three months. How do you feel about this—do you think troops should be withdrawn at a faster rate or slower rate?

Of the respondents, 42 percent answered "faster," 16 percent "slower," 13 percent had no opinion, and 29 percent volunteered the response "same as now." Three months later the Harris poll asked respondents:

> In general, do you feel the pace at which the president is withdrawing troops is too fast, too slow, or about right?

The Harris poll obtained responses of 28 percent "too slow," 49 percent "about right," 6 percent "too fast," and 18 percent no opinion. Unfortunately, we do not know how much of the difference between the Gallup and Harris poll results is due to events occurring in the interval between the polls, and how much is due to the differences between the two questions (Schuman and Duncan, 1974:233). Are the two questions really getting at the same issue? Are they actually measures of the same latent attitude, or are they dealing with different though related dimensions? What is required to answer such questions is the design of new measurements so that such issues, which routinely arise in the archival record, could be studied by experimentally varying the question wording in a future survey.

As a second example, measurements by Louis Harris and NORC of Americans' "confidence in national institutions" have been found to be discrepant even though the wording of the questions and the survey dates were closely comparable (see Turner and Krauss, 1978; Smith, 1981; and

Turner, Chapter 7 in Volume 2). In this case, the discovery that different survey organizations asking the same questions obtained different results (indeed, even the estimates of trends in confidence were not the same) sparked a debate that was part of the motivation behind the present study. The existence of archival records of such surveys makes intensive scientific study possible and allows us to learn from the experiences of the past, even when the significance of such experiences is not recognized at first exposure.

Much broader questions of this same kind can easily be proposed. Answers to such questions, narrow and broad, would both enhance the capital value of the stock of survey data and improve understanding of data now being collected. Analysis of archived data would be even more informative if survey organizations were to routinely coordinate their measurements as we propose in Recommendation 10. (We note that the General Social Survey [GSS] has been responsive to investigators' concerns about the meaning of its questions, most of which are copied from earlier surveys. In the eight years since its inception, GSS has conducted approximately one dozen split ballot experiments, about half of these at the instigation of data users.) We wish to encourage a more vigorous interaction between the analysts and producers of survey data. The product of this interaction would benefit both parties and might result in a general improvement in the state of theory and survey practice.

We hope that more survey organizations and polling enterprises will deposit their data in archives. The example set in the commercial sector by those organizations, such as Harris and Gallup, that routinely release most of their surveys to public archives deserves to be more widely emulated. Concerns about commercial and proprietary issues could in many cases be met by establishing an appropriate embargo period prior to public release of the data. When survey results have been publicized, we think that fair play dictates that the data themselves be made available at the same time. We take this to be, in part, the spirit of the rules on public disclosure recently adopted by the National Council on Public Polls.

We also urge the development and expansion of archives of auxiliary data. We are favorably impressed with efforts to explain and interpret trends in public opinion in the larger context of social and cultural events. We have included empirical demonstrations of such analyses in Volume 2, and we have in Recommendation 16 proposed more extensive linkage of opinion and other time-series data. To this end, we also propose the archiving of data on environmental and media events that will provide the frameworks for understanding the dynamics of opinion formation and change. The Vanderbilt Television News Archive provides one example of such an endeavor. The scope of the auxiliary data that will be useful, however, extends far beyond that derived from television news, and substantial work

is required to transform such raw data into a form that will be useful in the analysis of public opinion. Central tasks in such an endeavor will be the careful assessment of the coverage and reliability of coding in extant data bases. Our experiences with the *New York Times* data bank, a computerized index to archived news stories, and the services offered by publicly available clipping services, suggest caution in evaluating any claim of comprehensive cataloguing of the media's daily product. Severe and selective undercoverage is probably typical (see Appendix C). However, new technologies, such as the automated retrieval of news items from computer formatted wire services, may alleviate some of these problems.

Indexing, however, is a more substantial intellectual problem. What is ordinarily desired for analysis are counts of the frequency (or relative density) of news coverage of particular topics. The definition of topics and the coding of subject matter are problems that are neither new nor intractable, but they are difficult and crucial. The development of appropriate techniques for the organization and indexing of archival data will be necessary for their effective use.

Our final recommendation turns to the broadest perspective possible, to the question of surveys and society.

RECOMMENDATION 18: STUDY THE ROLE OF POLLS AND SURVEYS IN MODERN SOCIETY

Here we broach the notoriously treacherous subject of *Wissenssoziologie*, the study of knowledge as a social product and social instrument, how it is produced and how it is used. On the far side of such inquiry are the epistemological issues that first fascinated the German scholars who delineated this domain of research (for example, Mannheim, 1931/1936). But much can be done to produce enlightenment about surveys on the nearer side of such issues. As a first step, someone should write the chapter on polls and surveys that is missing from an eminent economist's monograph on *The Production and Distribution of Knowledge in the United States* (Machlup, 1962). Our preliminary skirmishes with this task (recorded in Chapters 2 and 3) indicate some of the difficulties and pitfalls. The effort serves to convince us of the utility of a "survey of surveys," surely a task at which the industry would excel if it chose to undertake it.

Our own exploratory study of the use of polls by the media provides some preliminary lessons. First, a veritable avalanche of poll reports is published by the American and British press. We have been able to do little more than cite illustrative cases of abuse that result from lax standards of reporting, but a sustained effort could provide reliable knowledge about the use of polls both as a method of social communication and as a method for the self-education of the society about its own views. A more detailed and

richly textured understanding of the ways in which published polls are used —by whom, under what circumstances, and for what purposes—would be of value. In the light of the recent research on the influence of the media upon the public (see Volume 2, MacKuen, Chapter 14; Beniger, Chapter 15; Marsh, Chapter 18), it would be interesting to know more about the feedback effects, if any, that occur when the public is told what the majority thinks about its presidents, the malaise of the national spirit, and so forth. Our explorations suggest that this will be a rewarding topic for those willing to undertake careful and systematic studies.

The second lesson is that the content analysis procedures required in such work must involve a series of successive approximations to develop a reliable and meaningful coding scheme. Our own attempt has, we hope, been described with sufficient detail to suggest problems that will be encountered. Future work may be facilitated by creative applications of automated retrieval systems to computer-readable data banks of news stories (see, for example, Erbring, 1980) *not* summaries or indices, which may hide errors of coding and coverage.

Finally, we offer a speculation—fueled by the experience of some of our members who coded and read a large sample of newspaper poll stories—that polls and surveys *can* play (or are reported as if they do or should play) a role in the decision making of public officials. Thus, the potential exists for the abuse of surveys to serve the interests of participants in a political struggle (see examples in Chapter 3). The scope of this problem and the actual (as opposed to potential) impact of such abuses should be investigated in an intensive study of the ways in which public opinion is reified, represented, and taken into account in settling conflicts in society.

The role of surveys and polls in contemporary society and the effect they have had in fixing public attention on numerology rather than the process of public debate is suggested by the divergence between the following two perspectives:

> The following pages tell the story of a new instrument which may help to bridge the gap between the people and those who are responsible for making decisions in their name. The public opinion polls provide a swift and efficient method by which legislators, educators, experts, and editors, as well as ordinary citizens throughout the length and breadth of the country, can have a more reliable measure of the pulse of democracy. . . . [Gallup and Rae, 1940:14–15].

> Having been polled as a representative of the public, [the citizen] can then read reports and see how he looks. As polls become more scientific and detailed —broken down into occupations, counties, income groups, religious denominations, etc.—the citizen can discover himself (and the opinions which he "ought" to have or is likely to have). . . . Public opinion—once the public's expression—becomes more and more an image into which the public fits its expression. [Daniel Boorstin, quoted in Gollin, 1980:449].

We offer to social scientists the challenge of exploring the sources of these contrasting views as an entry point to a broader study of the impact of polling on the processes of public debate, political conflict, and public decision making.

Notes

1. The panel's procedure for approving recommendations was as follows. All recommendations were submitted in writing, and each was voted on separately by members of the panel, including the chair and staff. The inclusion of staff in the voting reflects the fact that the distinction between members and staff of the panel was less meaningful than may usually be the case; one member of the staff, for example, was at the outset a member of the panel. However, no recommendation was considered approved if 6 or more of the 11 current members of the panel voted in opposition. All persons voting in opposition to an approved recommendation were afforded the option of having their names recorded in dissent, and here, as elsewhere in the report, separate statements were permitted. Three such statements appear at the end of this report.

The recommendations themselves were voted on in the form shown in the text. The accompanying text and supporting materials were discussed by the panel and directions were given to the editors. Unlike the recommendations, the text materials were subject to subsequent editorial changes without the requirement that individual changes be resubmitted to the panel for voting.

2. The titles of the individual articles convey their flavor: "Pollsters do their number on what black America thinks" (January 20); "Opinion polling by the Russians is oblique art" (February 3); "1980 brings more pollsters than ever" (February 16); "Polls, once scorned, gain new esteem" (April 5); "Experts find polls influence activists" (May 4); "The business of the pollsters" (June 29); "Abortion poll: not clear-cut; wording of a question makes a big difference" (August 18).

3. Two such articles by D. Greenberg appeared in the *Washington Post*: "Polls and other superstitious rituals" (September 9) and "The plague of polling" (September 16).

4. Obviously, the evaluation by Paletz et al. would be unfair if most *New York Times* stories reported *Times*/CBS News surveys. Based on our assessment, this does not appear to be the case.

We used the results of a search of the *New York Times* Information Bank (as current on January 15, 1980) to compile a list of *New York Times* stories published in 1978 whose abstracts contained the name of one or more of the polling and survey organizations listed in the Information Bank's index. We thus located 104 stories that (according to their abstracts) reported poll results. (We excluded biographical pieces on pollsters, stories on polling as an enterprise, and so on.) In the abstracts of these 104 stories we found 112 references to specific polls or surveys (some stories referred to more than one poll). Only 30 of these 112 references were to *New York Times*/CBS News polls. The remaining 82 references were made to surveys by: Gallup (43 references); Harris (13); ABC/Harris (4); Roper (3); Yankelovich, Skelly and White (3); Opinion Research Corporation (2); Conference Board-National Family Opinion (2);

California Poll (2); Audits & Surveys (2); and—with one reference each—Cambridge Survey Research; Mervin Field; National Education Assocation poll of members; CBS news poll of delegates; Associated Press/NBC; Opinion Research Centre of U.K.; Market and Opinion Research International of U.K.; and *Los Angeles Times* poll. Unfortunately, we have no convenient method of judging the relative space of these stories, and, as we note later, the Information Bank is not an entirely reliable index. Nonetheless, it appears that the Paletz et al. evaluation does cover a significant and noteworthy segment of poll reports in the *New York Times*.

5. We note and applaud the fact that some news media polls, for example, CBS/*New York Times* and ABC/*Washington Post*, produce regular summaries of their polls that include additional information not included in their news stories. Such documentation serves as a useful supplement to details reported in the news stories, and some combination of a report of essential details (including sponsorship, wording, sampling error, and so forth) in the story plus further technical information in a publicly available memo (for which a charge could be made) might provide the most flexible means of ensuring adequate reporting.

6. In Chapter 3, we review briefly five documents proposing or promulgating standards for polls and surveys. Dodd's (1947) proposed code to govern certification of opinion survey organizations; the Code of Professional Ethics and Practices of the American Association for Public Opinion Research (1979–80); the Principles of Disclosure of the National Council on Public Polls (1979); the Code of Standards for Survey Research of the Council of American Survey Research Organizations; and Circular A–46, "Standards and Guidelines for Federal Statistics," used by the Office of Management and Budget in reviewing proposed federal surveys (see Executive Office of the President, 1976). These documents differ in many ways, in particular in regard to their relative emphasis on technical versus ethical issues and on subjective versus objective variables.

7. See, for example, the analysis and reinterpetation of presidential popularity measures by Sussman (1978) in the *Washington Post*, which is discussed in MacKuen and Turner, Chapter 16 in Volume 2, and articles by E. J. Dionne, Jr., and others, discussed in this chapter.

8. For example, in most issues of the *Journal of the American Statistical Association*, the book review columns do not have any coverage of actual survey reports, even though the "publications received" section lists numerous such reports. (Of course, books that include survey results are reviewed.)

9. Lester Frankel, Howard Schuman, and Tom Smith dissent; see statement in Appendix A.

10. The idea that a national sample of 1,500 respondents is adequate for a survey may trace back to someone's calculation from the binomial distribution that with $p = .5$, two standard errors of the estimated proportion are approximately .026, so that the 95-percent confidence interval implies an "error" of between 2 and 3 percentage points. (When p is larger or smaller than .5, this "error" is less.)

11. This principle is well known in experimental studies, for which the idea of factorial design is now commonplace, startling as it was when Fisher introduced it more than a half-century ago.

Experimental designs can provide estimates of (specific) components of nonsampling variation if these components are identified beforehand and the relevant factors (for example,

question wording, interviewer training, question context, and so on) are experimentally varied in the design of the survey.

12. Estimating the parameters of such models would require the use of factorial designs to assess the magnitude of the errors introduced at each step in the production of survey data.

13. To be sure, the Lazarsfeld school of survey analysis had developed a set of procedures corresponding to a well-articulated "language of social research," which featured cross-classification as a device for testing causal hypotheses (see, for example, Rosenberg, 1968). But the same group of research workers rather deliberately eschewed formal procedures of statistical modeling and inference.

APPENDIXES

Primary responsibility for preparation of these appendixes was assumed by DeMaio (B, C, D), Marsh (B, C), Martin (H), Smith (F, G), and Turner (B, C, D, F, G). See Note on Authorship, page vi. Separate statements (Appendix A) are the individual responsibility of the named authors. The material presented in Appendix E is reprinted with the permission of Union Carbide Corporation.

A

Separate Statements

Statement, Otis Dudley Duncan

An important topic is not covered in the foregoing recommendations because the panel was unable to formulate a position commanding majority assent. I state my personal view in the interest of stimulating consideration of this topic by readers of our report.

The topic is the role of federal agencies. Collection of data on subjective variables by federal statistical agencies should, in my opinion, be limited. This report acknowledges, indeed, emphasizes, that there are subjective aspects to many questions that ostensibly deal with matters of fact. So there is no possibility of producing social statistics without getting into subjective phenomena to some extent. But this concession should not be allowed to become the foot in the door whereby government statistics are expanded to cover the whole range of subjective variables. There are two considerations, one technical and the other political, the former being much less important than the latter.

On the technical side, I write with the experience of some three decades of working with government statisticians in various ways and of attempting to do research in both the fact-finding and subjective modes. My observation is that the skills and sensitivities needed for good fact-finding research are, for the most part, complementary to rather than coincident with those required for first-rate research focused on subjective variables. Storer

(1969:6) may have recognized this state of affairs in proposing "the establishment of another separate survey research agency in the Federal government." No doubt such an agency could recruit survey experts knowledgeable about subjective phenomena and thereby overcome any present technical limitations on the expansion of government statistics into the subjective area. Hence there is no need to dwell upon or overemphasize any such limitations.

The main concern is political. There are large parts of peoples' lives for which they should not be accountable in any way to the government, including what they think, feel, and believe. A panel that preceded us found that: "A substantial proportion of respondents expressed negative feelings about the value and accuracy of surveys, how interesting they are to respondents, the confidentiality of survey records, and the integrity of survey takers" (National Research Council, 1979:39). One can argue that a "substantial proportion" of the public is mistaken about these matters. Nonetheless, people have the right not to be badgered by their government into participating in an activity they do not believe in. The realm of the subjective is the last castle one may defend against the coercion of society and government and the vicissitudes of life and death. It is too precious to become the plaything of bureaucrats, however worthy their intentions to make a better life for us all. The U.S. Bureau of the Census has no business asking me or you "Are you very happy, pretty happy, or not too happy?"

To be sure, apart from the legally mandatory decennial census, government surveys do not compel respondent cooperation. But people do not necessarily know this and even if so informed they well may wonder if nonparticipation can have adverse consequences. It is known that the Current Population Survey of the Bureau of the Census enjoys a much lower refusal rate than nonfederal surveys. In parallel surveys by the Census Bureau and Survey Research Center (University of Michigan) the (unweighted) refusal rates were 5.8 percent and 13.0 percent, respectively (National Research Council, 1979:42 [Table 1]). It may be that at least 7 percent of the American people believe that participation in a federal survey is required by law or is otherwise an obligation of the citizen. That is too many people undergoing the admittedly "subjective" experience of being coerced into revealing what is on their minds.

In another publication (Duncan, 1972), I have argued persuasively that there are indeed things a government ought to know about how people feel, but that these are among the very things that government agents should be forbidden to ask. Whether people trust the government is one. Whether the public thinks there should be religious observance in the schools may be another. But, as I noted in 1972, there are reputable survey and polling organizations that can collect relevant statistics. Statistics pertinent to wise and benign governance need not be produced by government statisticians.

My recommendation is not controversial. It is, in fact, no more than a restatement of the policy of the Bureau of the Census, at any rate the one that prevailed a few years ago when, according to a high official (Levine, 1968), it "had a strong predilection for factual as opposed to attitudinal or motivational questions." Mr. Levine did go on to note that "it has been necessary to consider expanding the scope of the Bureau's work in recent years." I urge that the Bureau and sister federal statistical agencies continue to be strongly skeptical of all allegations of "necessity."

Statement, Baruch Fischhoff

One issue on which our panel declined to take a position was the appropriate level of federal government involvement in conducting surveys of subjective phenomena. I would like to note some arguments that motivated this decision.

Adopting a policy on any issue obviously requires a consideration of the costs and benefits associated with that policy and the consequences associated with the alternatives to that policy. From either perspective, a blanket recommendation regarding government involvement seems ill-advised for two reasons.

First, one legitimate goal of government is to meet the needs of the populace as felt by the populace. Two potential benefits of surveys in this regard are: helping civil servants to understand those needs and providing them with feedback on the adequacy of their performance. The extent of such benefits arising from a particular survey depend upon how well it is done, how sensitively it is used, and how great the need for it was. It may be particularly useful when civil servants' intuitive "reading" of the public will is erroneous. In any case, the value of surveys must be assessed on a case-by-case basis, rather than judged categorically. Assessing the costs of a survey also requires an individualized analysis. For example, one wants to know:

—Will the respondents' participation be solicited in as non-coercive a way as possible (so that any increase in response rate could be interpreted as reflecting a greater desire to participate in federal, rather than private, surveys)?

—Will its results be broadly disseminated rather than being put to private (and possibly manipulative) use?

—Are its questions designed to direct policy (or just to satisfy curiosity)?

Clearly, an assessment of the costs and benefits of a proposed survey should

serve not only an evaluative function, but also suggest ways to make it more attractive.

Second, if the government were to stop conducting surveys of subjective phenomena, one must ask what will come in its stead. Perhaps the most benign possibility is increased reliance on political punditry, with those individuals best able to gain a hearing for their views expounding on what the public thinks and wants. A more troublesome possibility is that surveys will continue to be conducted, but that they will be primarily proprietary in nature. That is, they will be conducted at the behest of those individuals or organizations with the resources to commission them. The results of such surveys may then be used to strategic advantage, perhaps being kept in secret, perhaps being released selectively. An individual who questioned the intentions of these institutions might argue that if "surveys are outlawed, only outlaws will have surveys." Government, on the other hand, is required to reveal the results of the studies it conducts or commissions. Moreover, it is entrusted with acting in the public's best interest and subject to the censure of that public should it fail to do so.

None of these comments, of course, would excuse the conduct of knowingly flawed or biased surveys, or the misrepresentation of their results for manipulative purposes. However, even such excesses reflect more poorly on those responsible for them than on the medium they use.

Dissent to Recommendation 6, Howard Schuman (Lester Frankel and Tom Smith join in this dissent.)

Independent peer review is desirable for all scientific research projects that have important goals or raise controversial issues. However, there is no compelling reason to single out all surveys, especially without qualification concerning their goals. The further stipulation regarding "paid reviews" raises troublesome problems regarding the independence of such reviews. In general, this recommendation needs to be carefully thought through before being endorsed.

B

Notes on the Size and Scope of the Survey Industry

If we had information precisely suited to our needs, we might begin to gauge the social and economic importance of the survey research enterprise by asking about:

1. The number of people employed in the industry;
2. Its economic value to the consumers who buy its products, as indicated by its gross revenues;
3. The total number of individual respondents (and the total burden per individual) required to produce the raw data that are the basic product of the industry;
4. The value placed on its future—as indicated by such things as capital investments in the industry.

Unfortunately, there is no comprehensive information on any of these questions. Consequently, we have tried to assemble available information on the major survey research firms that make subjective measurements.

This compilation is not representative of the survey industry in general: many small market research and political polling firms no doubt exist about which we have no knowledge, and there are firms that we know exist but about which we have no information, for example, on their size or revenues.

The Sectors of the Industry

ACADEMIC AND COMMERCIAL SURVEY ORGANIZATIONS

Honomichl (1977, 1978, 1980), using information from corporate annual reports and supplementary information from corporate officials, has published estimates of the research revenues and employment of the largest nonacademic survey firms. For privately held companies (which are not required to publish annual statements of revenues and earnings), Honomichl was sometimes able to obtain data from corporate officials or other informants.[1]

Table B.1 summarizes these data on the largest commercial survey research organizations that (mainly) do subjective surveys and also presents estimates for several academic and nonprofit survey organizations.[2] The survey organizations covered in Table B.1 had estimated total revenues of more than $200 million in 1979, and employed almost 6,000 people, at least half of whom were hourly or part-time workers. Since Table B.1 includes only the largest firms, this total is an underestimate of the revenues of all such survey firms. We should also note that this compilation does not reflect the internal costs for survey research done by corporations and government agencies that have their own research divisions. (Conjectural estimates of the total annual cost of survey research in the United States range from $1.2 billion [Honomichl, 1982] to $4 billion [Bailar, 1979].)

An independent compilation by Spaeth (1979) indicates—as does Honomichl's review—that the survey research industry includes several very large survey organizations, such as: Market Facts, which reported conducting more than 900 surveys in 1978; Chilton Research Services, 467 surveys; National Analysts (Booz, Allen, & Hamilton), more than 200 surveys; Market Opinion Research, 285 surveys; Opinion Research Corporation and N.E. Field Facts, 1,500 reported surveys; and an anonymous commercial organization that reported conducting 3,137 surveys in 1978. In addition to the giants, however, there seem to be a very large number of small survey operations; in Spaeth's sample, 20 of 41 organizations reported conducting fewer than 10 surveys in 1978.[3]

We know of no direct measurements of capital investments in the survey

research industry as a whole. However, some fragmentary and indirect information is available. For example, in 1975 it was reported that one medium-size firm, Louis Harris and Associates, was purchased at a price of approximately $4 million (*New York Times*, 1975). Indirect information is also available on capital outlays in one sphere: investment in computer-assisted telephone interviewing (CATI) equipment. The last column in Table B.1 lists known CATI installations among the firms shown. In 1980, the minimum cost of entry into the CATI field was estimated to be about $100,000 for equipment (for example, minicomputer configuration with 10 to 15 interviewing stations); large operations can cost several times that. And, as with many computer applications, the cost of software development may dwarf hardware costs, but estimates of software costs for CATI maintenance and development are unavailable.

GOVERNMENT SURVEYS

Of the large number of surveys conducted by government agencies, some are contracted out to commercial or academic survey organizations (and thus may be included in the revenue figures shown in Table B.1), but many are carried out by the government itself, in particular, by the Bureau of the Census.

In 1980, the Demographic Surveys Division of the Bureau of the Census conducted 27 projects.[4] Many of these projects included survey measurements of subjective phenomena, although such measurements did not constitute the majority of the questions in any single survey.[5]

As is discussed in detail by Bailar and Rothwell (Chapter 5, Volume 2), measurements of unemployment, although generally treated as objective, have significant subjective components. Measurements of employment and unemployment are made in the monthly Current Population Survey (CPS) program, which conducted approximately 740,000 interviews in 1980. In addition to the unemployment items, supplemental questions added to the monthly CPS sometimes involve explicitly subjective measurements. For example, the CPS supplemental survey of fertility, birth expectations, and marital history, asks questions concerning the number of children women expect to have in the future. Similarly, a supplemental survey of voting in the election of 1980 asked nonvoting respondents to state their reasons for not voting (or not registering to vote) in the election. The CPS program cost approximately $23 million in 1980 (exclusive of supplements). The annual fertility and voting supplements cost $200 thousand and $150 thousand, respectively. The Methods Development Program, a methodological component of CPS, conducts approximately 36 thousand interviews each year at an approximate annual cost of $845 thousand. Several other continuing Census Bureau surveys also incorporate measurements of subjective

TABLE B.1

Estimated Revenues and Employment of Some Major National Survey Organizations (excluding organizations primarily engaged in nonsubjective measurement)

Organization	Description[h]	Date Founded	Employees[d]	Revenues	
				Year	Dollars (millions)
I. *Nonprofit*[i]					
National Opinion Research Center (U. of Chicago)[a]	Nonprofit organization, part of U. of Chicago after 1947. Part of U. of Denver, 1941–47. Primarily government-funded survey research and analysis.	1941	400 (Total)	1979–80 1975–76 1970–71	16.8 8.6 3.1
Institute for Social Research (U. of Michigan)[a]	Nonprofit organization, part of U. of Michigan since 1946. CATI with 16 stations. Primarily government-funded survey research and analysis (72% of revenue from federal government in 1979–80).	1946	—[g]	1979–80 1975–76 1970–71	14.1 8.9 6.4
Institute for Survey Research (Temple U.)[a]	Nonprofit survey organization, part of Temple U. since 1967.	1967	60(FT) 750 (Field)	1979–80 1975–76 1970–71	5.0 3.8 1.5
Research Triangle Institute[a]	Nonprofit organization affiliated with Duke U., U. of N. Carolina, and N. Carolina State. Conducts surveys mainly for government and academic research projects. Has CATI facility.	1958	275	1981 1980 1979	19 17 12
Rand Corporation[a]	Nonprofit organization founded in 1948. Began survey operations in 1972. Conducts surveys on public policy issues.	1948 (1972)	60	1980 1976	8.5 7.5

SRI International[a]	Nonprofit organization originally affiliated with Stanford U., but independent since 1970. Does survey and other social science research for government and industry.	1946			
II. *For Profit*					
Burke Int'l. Research Corp.[b]	Privately held organization. Custom survey research in marketing and advertising (all done by telephone). CATI with 11 stations.	1947	250(FT)	1979 1977 1976	42.6 23.1 18.6
Market Facts[b]	Publicly held corporation. Field surveys, telephone and mail panel surveys, store audits.	1946	556(FT-S)	1979 1977 1976	19.3 17.1 14.4
Westat, Inc.[b]	Employee-owned corporation. Personal telephone and mail surveys mainly for civilian agencies of U.S. government.	1961	170(S-Prof) 150(Field) 259(Sup)	1979 1977 1976	14.4 8.4 6.5
Audits & Surveys[b]	Privately held corporation. Surveys and store audits. CATI with 125 stations.	1952	200(FT)	1979 1977 1976	14 12.2 9.4
Marketing and Research Counselors[b]	Privately owned subsidiary of Tracy Locke Co. Custom research including surveys and shopping mail studies. CATI with 170 stations.	1965	136(Prof) 1,500(H) 216(H-Prof)	1979 1977 1976	13.1 10.0 7.5
Chilton Research Service[b]	Wholly-owned subsidiary of American Broadcasting Co. Telephone and personal interviewing. CATI with 150 stations.	1957	170(FT) 700(PT)	1979 1977 1976	12 5.1 4.3

TABLE B.1 *(continued)*

Organization	Description[h]	Date Founded	Employees[d]	Revenues	
				Year	Dollars (millions)
Yankelovich, Skelly & White[b]	Wholly-owned subsidiary of Reliance Group. Surveys for marketing and social research.	1958	206(FT)	1979 1977 1976	11.8 6.9 6.0
National Family Opinion[b,c]	Privately held corporation. Mail and telephone surveys, including mail panel of 150,000 households.	1946	90(FT)[c] 260(PT)[c]	1979 1977	10[c] 7.9[c]
Louis Harris & Associates[b]	Wholly-owned subsidiary of Gannet Co. (newspaper chain owner). Syndicated and custom surveys of public opinion, and consumer preferences.	1956	135(FT)	1979 1977 1976	9.3 5.8 4.9
National Analysts[b,e]	Subsidiary of Booz, Allen & Hamilton. Surveys for marketing and social science research.	1943	170(FT)	1979 1977 1976	8.7 7[e] [e]
Opinion Research Corporation[b,f]	Subsidiary of Arthur D. Little Survey for corporate, employee relations, marketing, political, policy and government research.	1938	146(FT)	1979	8.2
Elrick & Lavidge[b,f]	Wholly-owned subsidiary of Equifax. Studies of shopping malls and telephone surveys for marketing research.	1951	99(FT)	1979	7.1

Gallup Organization[g]	Privately held corporation. Private and syndicated surveys for marketing and public opinion research.	1935	—[g]	1979	—[g]

[a]Information supplied by organization (i.e., Boseman, 1980; McDonald, 1980; LoSeinto, personal communication, May 20, 1981; Larabee, personal communication, June 2, 1982; Russell, personal communication, June 21, 1982; Hensler, personal communication, July 30, 1982).

[b]Information from J. Honomichl, "Top 23 research leaders up 18%," *Advertising Age*, May 19, 1980; "Research Top 20", *Advertising Age*, April 11, 1977; "Research Top 20", *Advertising Age*, April 24, 1978. Revenues are for research only.

[c]Organization did not cooperate with *Advertising Age* industry roundup (see b), hence the estimates made by *Advertising Age* may be unreliable.

[d]Employment reports are not comparably classified in source documents. Where broken down, we have indicated types of employees: FT (full time), PT (part time), S (salaried), H (hourly), Prof (professional), Sup (support), Field (field staff). Number of employees is for most recent year.

[e]Revenues for 1976 (11.9m) and 1977 (17.5m) were consolidated with other Booz, Allen & Hamilton marketing and development companies.

[f]No published estimates available for years prior to 1979.

[g]No published estimates are available.

[h]Firms *known* to have computer assisted telephone interviewing facilities (CATI) are noted together with the maximum number of CATI interviewing stations available, if known. This information was compiled by William Nicholls, U.S. Bureau of the Census (1980), Mary Spaeth (1979), and Nicholls (personal communication, May 25, 1981), and for RTI, Larabee (personal communication, June 2, 1982).

[i]List of nonprofit firms includes only those known to routinely conduct national surveys, hence regional survey organizations, such as the Survey Research Center (University of California at Berkeley), the Survey Research Laboratory (University of Illinois), and the Center for Survey Research (University of Massachusetts and Joint Center for Urban Studies, MIT-Harvard University) are not included.

[j]SRI employed 159 people in its social science section in 1981 and had revenues for all social science research of $11 million (total revenues for all SRI sections was $164 million). However, only a part of SRI's social science activity involves survey research.

phenomena. Subjective measurements of satisfaction with various aspects of neighborhoods are incorporated in the Annual Housing Survey program which conducts 177 thousand individual interviews annually at a cost of $11 million. Work aspirations, psychological dispositions, and similar topics are covered in the National Longitudinal Surveys in which approximately 20,000 respondents are interviewed either 8 or 9 times over a 10-year period; the cost in 1980 of this program was $1.2 million. Attitudes toward work and experiences in the CETA program (Public Law 95-524, title III) are measured in the Longitudinal Manpower Survey which includes 8,000 respondents interviewed at 1-, 2-, or 3-year intervals; the total cost of this program in 1980 was $6 million. Estimates of the effects of health disabilities are included in the Health Interview Survey program which conducts 40,000 interviews annually at a total cost in $4 million.

There are also quasi-subjective aspects to the questions asked in the National Crime Survey (144 thousand interviews were conducted annually at a cost of $5.6 million in 1980), and, in the Survey of Travel to Work[6] (conducted as a supplement to the Annual Housing Survey at a cost of approximately $5.2 million). The planned Survey of Fishing, Hunting and Wildlife Recreation also may include subjective measurements.[7]

Total costs for these Census Bureau surveys totaled more than $50 million and involved more than 1 million individual interviews in 1980.

IN-HOUSE SURVEYS

In addition to the commercial, academic, and government surveys just discussed, there are an unknown number of surveys conducted in-house by the research departments of large corporations. For example, the research division of General Foods is believed to conduct a large number of surveys for marketing purposes. AT&T has a large-scale survey operation that (primarily) gathers data on customer satisfaction. This survey operation, known as Telsam (Telephone Service Attitude Measurement program), is equipped with the most extensive CATI interviewing system in the nation. 460 computer terminals are linked to 10 computers that automatically display the appropriate sequence of questions and store the answers, which interviewers type directly into their terminals.[8] By the estimate of the manager of this operation, AT&T's Telsam program conducts more than 5 million survey interviews each year. These interviews provide aggregate statistics on customer perceptions of the quality of service across the telephone network, and it can be used to identify problems of particular classes of customers for remedial action.

During the 1970s, television, newspaper, and radio news organizations also began to establish in-house survey operations. The Columbia Broadcasting System (CBS) and the *New York Times (NYT)* established a joint

survey operation in 1976, and they have been conducting regular national surveys that are used in their news reporting. Also in 1976, the National Broadcasting Company (NBC) and the Associated Press (AP) wire service began a similar joint venture,[9] and the *Washington Post* began conducting its own polls.[10] The *Los Angeles Times* entered the field in 1979–80. In the CBS/*NYT* and NBC/AP operations, the survey design, interviewing, and analysis are done in-house. The *Washington Post* and *Los Angeles Times* have editors in charge of their polling operations, but the interviewing is contracted out to service bureaus. Questionnaire design, data analysis, and interpretation are done in-house, however, and news stories based on the poll results are ordinarily authored by the editors. (These results are also used in stories written by other reporters.)

These relatively well-known polls begun during the 1970s by the national news media are the visible manifestation of a much larger trend. According to data reported by Mitofsky (cited in Dionne, 1980c), there are at least 147 national or state polls sponsored by local newspapers or broadcast outlets around the country; one-third of them were established between 1978 and 1980.

The Number of People Surveyed Annually

Although accurate estimates of the number of survey organizations are not available, some data are available on the frequency with which people report being contacted by them. In a survey conducted jointly by the Survey Research Center (University of Michigan) and the Bureau of the Census, respondents were asked whether they had been contacted by a survey organization at any time "in the last four or five years." The respondents were a probability sample of noninstitutionalized people aged 18 and over residing in the United States (excluding Alaska and Hawaii).[11] Forty-six percent reported that they had been contacted; the average number of reported contacts was 3.26. Under the assumption that respondents reported contacts made with any member of their household, these data would suggest that approximately 32 million households were contacted and over 104 million interviews conducted during the years 1971–76.

This does seem to be an extraordinarily large number of survey contacts, and we might question the accuracy of respondents' recall as well as their ability to distinguish between bona fide surveys and salespersons who begin a sales pitch by posing as survey takers. However, if we keep in mind that a single survey, the Census' Current Population Survey, conducted 2.7 million interviews during that 5-year period, the figure begins to appear

more plausible. If we add in other large-scale operations such as AT&T's Telsam surveys and the work of the other firms we have noted previously, the figure becomes more plausible than it may seem at first sight.

Respondents also reported that the majority of these surveys were on topics that were clearly subjective. The following list shows the distribution of survey topics reported by the respondents:[12]

Ideas about a product	32%
Political views	21%
Other opinions	14%
Personal or family data	21%
How spent time or money	9%
Other	4%

It seems that whatever the magnitude of the survey enterprise, the subjective side of survey research accounts for a very substantial proportion of survey interviews—or at least of those that respondents recall.

We do not have any information on revenues or investments in survey research in Great Britain, but there is one source of information on the number of people surveyed. On the basis of a survey conducted by the Market Research Society, it has been estimated that 39 percent of British adults have taken part in market research studies (Bowen, 1979:78). We do not know what proportion of this total would be excluded from the domain of our study because they were strictly factual, such as surveys restricted solely to reports of past purchases; however, we suspect that this proportion would not be very large.

Notes

1. In the case of one privately held firm for which we have access to a reliable informant, we have been informed that Honomichl's (1980) published estimates are "roughly correct."

2. In making this compilation, we began with Honomichl's (1980) list of the 23 largest research firms, and then we excluded several firms whose business is concentrated in survey measurements of nonsubjective phenomena, for example, audits of store sales and inventories. The largest of these companies is A.C. Nielsen, the national TV rating service. This service, which relies on electronic monitoring of TV use in a sample of households, is clearly not subjective. Similarly, the self-report measurements of past television watching, which Nielsen gathers for local areas, involve a nonsubjective phenomenon: past behavior. Although A.C. Nielsen does some "custom survey research" which involves the measurement of subjective phenomena, we did not include this organization in our compilation because the bulk of its

$321 million in revenues appears to be derived from research that is outside the domain of our study.

Other firms not included in our compilation are IMS International ($88.8 million), which specializes in measurement of sales and inventories of pharmaceuticals; SAMI ($54.4 million), whose basic activity is the production of aggregate data on the movement of brand name items from wholesale warehouses to retail outlets; Arbitron ($47.1 million), whose business is mainly in radio and television ratings of audience size; ASI Market Research ($14.4 million), which specializes in controlled audience pretesting of commercials, pilot films and phonograph records; Ehrhardt-Babic Associates ($10.4 million), which concentrates in store audits; NPD Research ($9.7 million), which specializes in consumer diary studies apparently of purchases and consumption; and Walker Research ($7 million), Starch, INRA, Hooper ($5.5 million), Decisions Center ($5.1 million) and Data Development Corporation ($9.6 million), firms for which Honomichl (1980) did not provide sufficient information for us to assess whether their main activities involved measurements of subjective phenomena.

After this compilation was completed, Starch, INRA, Hooper acquired the Roper Organization, Inc. for $17,000 in cash and 100,000 shares of common stock (*source*: Starch, INRA, Hooper. Inc 1981 Annual Report). At the time of acquisition Roper's assets were listed at $249,909 and their liabilities were $553,200. During the five months subsequent to their acquisition: "Roper experienced a small loss on $666,197 of revenue . . . ." Roper's business is largely in the measurement of subjective phenomena.

3. Unfortunately, the "number of surveys conducted" is a vague concept that hides differences between organizations in such factors as average sample size and length of interview. If we relied only on this statistic, an organization which did a single survey involving one hour interviews with one million households might appear "smaller" than one doing five surveys of 10 minutes each with samples of 100 households. For this reason, Spaeth's data should be interpreted in conjunction with Honomichl's data on revenues.

4. Estimates of expenditures for the Demographic Surveys Division were obtained from a memorandum prepared by the division (Bureau of the Census, 1980c).

5. The Youth Employment and Training Programs and Youth Community Conservation and Improvement Projects (YETP/YCCIP) attitude supplement to the Longitudinal Manpower Survey is an exception to this rule.

6. Bureau of the Census (1980c) indicates that data collection for all surveys in this program was completed by June 1980; it is unclear whether the $5.2 million annual cost was incurred in 1980.

7. The Retirement History Survey, which reinterviewed a sample of persons born between 1906 and 1911 (original sample of 12,800 drawn in 1969) is not included in this discussion, nor is the 1978 Survey of Veterans since the field work ended prior to 1980 (although data processing and analysis expenses appear in the 1980 budget).

8. The computers (in 1981, 10 PDP Model 11/70s) are located at nine sites around the country. The facilities are owned by AT&T, but interviewing staffs are provided under contract by Chilton Research Services and several other contractors.

9. NBC originally began working with a consortium of newspapers.

10. During the course of the panel's work, the *Washington Post* made new arrangements for a joint polling operation with the American Broadcasting Company (ABC) television network.

11. This personal interview survey was conducted between September and December, 1976 for a study done by the Committee on National Statistics. The overall noninterview rate is reported (National Research Council, 1979:23) as 22 percent for the SRC sample and 15 percent for the Census sample. Respondents were asked three basic questions about survey contacts:

> We would like to start by asking you to recall any time in the last four or five years when you were asked to respond to a survey. This card lists some of the things surveys ask about. Before we go on please look it over to help you remember all the times you were asked to be part of a survey.
>
> A1. The first type is a mail survey. Has anyone taking a survey sent you a questionnaire in the mail in the last four or five years?
>
> A2. During the last four or five years has anyone taking a survey called to ask you questions on the telephone?
>
> A3. Besides me, has anyone taking a survey come to your door to interview you in the last four or five years?

For each type of survey (that is, mail, telephone, personal visit), respondents were also asked the frequency with which they had been surveyed, for example: "About how many different times have interviewers come to your home for a survey," and the topics covered in the most recent survey, for example, "What kinds of things did the (most recent) interviewer who came here want to know about? You can give me the letters on this card that describe the questions." The card listed 6 topic areas:

(a) Your ideas about a product or advertising;
(b) Your political views, ideas about issues or candidates;
(c) Your opinions about other things;
(d) Descriptive information about yourself—or your family—age, education, work, income;
(e) Other things you have done or that have happened to you—how you spend your time or money;
(f) Other: Please describe

We would note that the distribution of survey contacts is unusual: more than half of the respondents reported zero contacts, while the remainder reported an average of more than three contacts. This suggests that the various surveys that are being reported are unlikely to have been independent random samples of the national population. (We know that this must be true to some extent given the existence of local-area polls.) Unfortunately, although the survey researchers had hopes of doing more extensive studies of the error structure of these data (see National Research Council, 1979: Appendix A), there is little information on the error structure of these measurements. (In this regard, one notes that a similar commentary has been made about Walker Research's estimates of the frequency of telephone surveys; see Opinion Outlook, 1981).

12. Respondents' descriptions of individual surveys were coded to include all topics mentioned. Percentages were calculated using the total number of topics mentioned as denominator.

C

Details of Newspaper Study

This appendix documents procedural and methodological aspects of the clipping study, results of which are reported in Chapters 2 and 3.[1] As mentioned in these chapters, our attempt to systematically collect information about media reporting of polls and surveys should be regarded as exploratory. At the time we began the project, we knew of no similar investigations,[2] and we wished to conduct as comprehensive a study as possible, within the constraints imposed by the time and resources available to us. However, as detailed below, we found that our study could not be as comprehensive as one might have hoped for.

We contracted with two commercial clipping agencies (Press Intelligence in the United States and Apcut in Great Britain) to supply copies of all stories that were published in any of several hundred major newspapers (and, in Great Britain, magazines) during the period studied.[3] Although the clipping was supposed to be done over a simultaneous one-month period in the two countries, administrative difficulties resulted in the British clippings being obtained from July 31 through September 11, 1980, and American clippings being obtained from July 1 through July 31, 1980.

We received a total of 1,038 American and 168 British news stories, clipped from 207 American newspapers and 66 British newspapers and magazines. Since the numbers seemed low to us, we conducted an independent check of a sample of the newspapers. The check revealed that, indeed, the clipping agencies missed the overwhelming majority of refer-

ences to polls and surveys. An independent reading of one British newspaper, the *Daily Telegraph*, showed that only about one sixth of the British stories (6 of 37) were clipped. Similar checks on four American papers, the *New York Times*, the *Washington Post*, the *Boston Globe*, and the *Harrisburg Patriot*, revealed that only about one-tenth of the American stories (25 out of 252) were clipped. These are minimum figures since our checkers might also have missed some relevant articles. Although this massive undercoverage is distressing, we could not discover any systematic pattern to the omissions. Articles on the first and second pages were missed as well as articles on the back pages of the papers; surveys cited in headlines were missed as well as those which were not.

We decided to proceed with the analysis in an exploratory spirit recognizing that our datafile probably represents about 10 to 15 percent of the material about surveys to which the reading public was exposed during July and August, 1980. It is, in any case, the best approximation to a sample of poll reports that we could generate given our limited time and resources. We regard the whole inquiry as a pilot study for the investigation called for in Recommendation 18 in Chapter 10 (that is, to study of the role of polls and surveys in modern society).

Two separate data files were compiled from the clippings. One contains information about the clipping itself—where the poll or survey result appeared in the article, whether the results were described as a "mandate," and so on—and about the level of reporting of sampling and other details of the polls and surveys whose results were reported. A copy of the codes that were used in this content analysis exercise is included at the end of this appendix.[4]

Because of the large number of election polls reported in the American clippings, they were not all coded. A complete sample was coded for 13 days, and a random sample was coded for the remaining 18 days. The characteristics of the *sampled* election polls were weighted in proportion to the number of election polls published divided by the number included in the sample. The number of American clippings actually coded is 730; their weighted sample size is 1,038. We coded the details of two poll reports for those clippings in which more than one poll was mentioned. In our samples of 1,038 American and 168 British clippings, we obtained a total of 894 and 195 poll reports, respectively. (Note that the same poll may be reported in more than one clipping and that a single clipping may report more than one poll.)

The second data file contains information about the poll being reported as far as it could be determined from the various reports. If the clipping file notes whether or not the sample size, method of interview, and so forth, were reported in the article, the poll file notes the actual sample size, method of interview, subject matter, and so on. Because the same poll was

reported in numerous articles, the number of polls is much smaller than the number of clippings. The second data file contains information on 253 American and 71 British polls. Care must be taken in drawing inferences from this last sample; it is, of course, a sample of the kind of surveys that find their way into the papers (with probability of selection being a function of how widely the survey was reported, given the incompleteness of our clippings), and not a cross-section of survey activity itself.

The effect on the public of the way poll results are reported is amplified by the newspaper's circulation or the number of potential readers of the poll results. Our analysis takes this into account by weighting the data by the circulation of the newspapers in which the survey or poll was reported.

Because of the inherent deficiencies of our sample, the restricted period sampled, the problems of measurement inherent in defining the concepts and categories of our semantic analysis, and the restricted time and resources available to us, the results serve mainly to delineate problems for a future and more comprehensive study. Several suggestions are offered to those interested in expanding on this work.

First, and most obvious, future studies should obtain more complete coverage of the clippings for whatever period is sampled. The automated methods being developed by Erbring (1980) hold particular promise of solving the sampling problem.

Second, an investigator should choose either one very long time period or several shorter ones spaced over a long period of time in order to reduce the influence of one or two specific events. Our sample, for example, was heavily weighted with election polls because of the time at which the clipping was done. (Within the relatively short time frame of our project, however, any clipping period would have faced the same problem.)

Third, future studies will need a clearer definition of the purpose for which a poll result is reported in an article. We at times found it difficult to distinguish between poll results as the focus of an article and those cited to support or oppose an article's point of view. This was more of a problem for some types of articles than for others; in commentaries and news analyses, a more rigorous criterion is necessary, perhaps with more detailed categories than the ones used here.

Fourth, our initial desire to document the proclivity of newspapers to make judgments about the past or future effects of poll results had to be abandoned due to a lack of conceptual clarity and the consequent lack of coder reliability. Several dimensions need to be considered in this regard. Although we can enumerate the dimensions necessary to quantify these judgments, we have not been able to formulate rules to identify the presence or absence of a media judgment in the first place.

Fifth, although our coding of individual base *N*s for percentages was not problematic, it could be improved by distinguishing between reports in

which base *Ns* are given in addition to percentages and those in which only *Ns* are given, leaving the reader to calculate the percentages.

Lastly, we suggest that further studies of this kind make a more concerted effort to code the number of column inches in the poll report. We never did arrive at a satisfactory solution to such mundane problems of measurement as different size type, different width columns, and photographic reduction of the newspaper articles to fit into 8½ in. × 11 in. binders prior to coding.

I. Codes Used in Newspaper Clipping Data File[5]

Columns	Specification
Card 1	
1–4	I.D. Number of clipping: 0001–4999 (from U.K. clips) 5001–9999 (from U.S.A. clips)
5–10	Date of clipping in format: MMDDYY, where M = month, D = day, Y = year (if clipping is from a monthly publication, leave DD blank)
11	Type of article 1. News report, general 2. News report on polling or the polling industry in general (news articles reporting results of specific polls should be coded as "1") 3. Miscellany column (i.e., collection of three or more items which are physically separated from one another, but which appear under a single headline) 4. Editorial 5. Letter to the editor 6. Signed piece of commentary (e.g., op-ed or column explicitly identified as "news analysis" piece) 7. Syndicated poll column 8. "Poll" column (used only in U.S.A.) 9. Magazine article (used only in U.K.)
12	Where in the clipping does a survey result or a word containing "poll" or "survey" as a root first appear? 1. In the headline 2. In the first paragraph 3. Later in the story

Columns	Specification
13	Does clipping report at least one actual (not hypothetical) percentage or other statistic (mean, etc.,) derived from an actual survey? 1. No 2. Yes, numeric and exact 3. Yes, but non-numeric form, e.g., "a majority," "over 60 percent," "few," "most people in a recent survey," etc.
14*	Are poll results the focus of the article, or are they included to support the author's point of view about something? *(See Note 4.)* 1. Focus of the article 2. Cited to support or oppose article's point of view 3. Incidental mention of poll 4. Poll results not reported *Examples*: 1. Article reports the results of a recent poll taken for the *Boston Globe* on racial attitudes in metropolitan Boston. 2. Headline: "Polls Don't Support Anxiety over Draft"; Gallup and Harris poll results mentioned or referred to only in the second and third paragraphs of lengthy article; the rest of the article defuses arguments against the draft. 3. In a lengthy article outlining Ronald Reagan's plan for the economy, the last two paragraphs report results of the most recent Gallup candidate preference poll.
	Columns 15–16 refer to polling in general rather than to specific polls.
15	Does the clipping mention general or specific criticisms of polling practices except wording of questions? 0. No 1. Yes
16	Does the clipping mention general negative features of question wording? 0. No 1. Yes
17	Does the clipping criticize the accuracy, validity, or reliability of a particular poll result? 0. No 1. Yes

Column	Specification
18*	In reporting the results of a poll, is a political context within which they have particular meaning cited? *(See Note 4.)* 0. No 1. Yes *Examples of Code 1*: Survey results showing that most Americans favor school prayers mentioned in connection with upcoming House subcommittee hearings on prayer in public schools; survey results showing that Vietnam veterans disagree with U.S. involvement there but are glad they served mentioned in connection with Carter's authorizing a site for a national Vietnam memorial.
19	Is the poll cited as evidence that an organization or person *should* take an action or act differently? 0. No 1. Yes 9. Poll results not reported
20–26	IF YES: Who should take action or act differently? (0 = No, 1 = Yes, 9 = Poll results not reported)
20	Political party
21	Political leader
22	Government or government agency
23	Business or business organization(s)
24	Labor or labor organization(s)
25	Other (specify)
26	Not explicitly stated
27–28	What action should they take? (List on separate sheet)
29	Is the poll cited as evidence that the past action(s) of an organization or person is supported or opposed (includes past legislation)? 0. No 1. Yes
30–36	IF YES: Is it an action of: (0 = No, 1 = Yes, 9 = Poll results not reported)
30	Political party
31	Political leader
32	Government or government agency
33	Business or business organization(s)
34	Labor or labor organization(s)
35	Other (specify)
36	Not explicitly stated

Columns	Specification
37*	Does the clipping argue that someone should be encouraged or discouraged by the poll results? *(See Note 4.)* 0. No 1. Yes, it is explicitly stated
38–43	IF YES: Who should be encouraged or discouraged? (0 = No, 1 = Yes, 9 = Poll results not reported)
38	Political party
39	Political leader
40	Government or government agency
41	Business or business organization(s)
42	Labor or labor organization(s)
43	Other (specify)
44–49	Does the clipping use any of the following terms or their synonyms? 0. No (includes all clippings referring to results of electoral polls on candidate preferences) 1. Yes, used by paper 2. Yes, attributed to a source
44	–"vote," referring to the people voting for a policy, etc.
45	–"mandate," i.e., that the public has given someone a mandate
46	–"views of constituents," i.e., that the poll expresses the views of constituents on an issue (or "electorate," "voters")
47	–"verdict of public opinion," e.g., that the public has rendered a verdict on a policy
48	–"popular" or "public support" (or "approval") for a program
49	–"public taking the side of" or "backing" some person or organization
50	Does the clipping report individual base *N*s for percentages? 0. Never 1. Sometimes (total sample size is not sufficient) 2. Yes, always 9. Poll results not reported
51–58	Are any results broken down by the following characteristics? (0 = No, 1 = Yes, 9 = Poll results not reported)
51	Age
52	Race/ethnicity
53	Gender
54	Geographic region

Column	Specification
55	Social class (i.e., SES or any of the following: education, income, occupational level)
56	Trade union membership
57	Political party membership
58	Other (specify)
59	Does the clipping report a poll showing consumer satisfaction or dissatisfaction with a commercial product (e.g., 56 percent of users of X toothpaste report they have fewer cavities; 86 percent of beneficiaries of cancer insurance report they are happy they bought the insurance). 0. No 1. Yes
60*	How many different polls are mentioned in the clipping? *(See Note 4.)* 0. None—skip to column 71 1. One—skip to card 2 and complete a card for the poll 2. Two 3-8. Three to eight 9. Vague reference to recent polls
61	Does the clipping compare percentages from similar measures obtained at two or more points in time? 0. No 1. Yes
62	Does the clipping mention general patterns of agreement or disagreement between findings of two (or more) polls, except for trends across time? 0. No, not applicable 1. No, presents results of another poll but does not explicitly conclude that they agree or disagree 2. Yes, mentions agreement or disagreement with other poll results but no specific details (e.g., percentage distribution, etc.) of other poll 3. Yes, mentions agreement or disagreement with other poll and presents details (e.g., percentage distributions, etc.) of other poll (*If two polls are cited*, go to card 2 and complete a card for each. If more that two polls are cited, go to card 2 and complete a card for the two most detailed.)

Details of Newspaper Study

Columns	Specification
71–73	I.D. Number for newspaper in which clipping appeared
74	Type of newspaper: 1. U.K.—quality daily newspaper 2. U.K.—tabloid newspaper 3. U.K.—local newspaper 4. U.K.—Sunday newspaper 5. U.S.A.—national newspaper (daily or Sunday) 6. U.S.A.—local newspaper (daily or Sunday) 7. Magazine 8. Unknown
75–78	Circulation of newspaper in thousands (use *either* the daily *or* Sunday figure depending on the day of circulation of the clipping)
79	Record number 1. Card 1
80	"C" is coded in column 80 of all clipping records
Card 2	
1–4	I.D. Number of clipping
5–8	I.D. Number of poll: 0001–4999 (from U.K. clips) 5001–9999 (from U.S.A. clips)
9	Country in which poll was done: 1. U.S.A. 2. U.K. 3. Other Europe 4. Canada 5. Other (Specify in notes)
10*	Who is the sponsor? *(See Note 4.)* 1. Newspaper from which the clipping was made 2. News media other than the paper from which the clipping was made 3. Gallup poll with no explicit sponsor 4. Harris survey with no explicit sponsor 5. Other sponsor 6. Sponsor not mentioned

Column	Specification
11	Is the agency or survey organization conducting the poll mentioned? 0. No 1. Yes
12*	What restrictions of the target population are specified in the clipping? *(See Note 4.)* 1. Age restriction (e.g., adults only) 2. Geographic restrictions (e.g., national sample) 3. Age and geographic restrictions 4. Other types of restrictions (i.e., special sample, e.g., Vietnam veterans) 5. No restrictions specified 6. Voters
13	Is there any mention of the sample type, i.e., quota, full probability, etc.? (Mention of interview technique, e.g., telephone survey, is not sufficient.) 0. No 1. Yes
14*	Is the sample claimed to be "representative" or "scientific"? *(See Note 4.)* 0. No 1. "Representative" 2. "Scientific"
15	Is the sample size mentioned? 0. No 1. Yes, exact 2. Yes, inexact
16	Are the dates of interview mentioned? 0. No mention 1. Exact dates given 2. Approximate dates given (Include cases which mention early, middle, or late part of month.) 3. Month of survey given
17	Is the rate of nonresponse mentioned (for full probability samples, mention of some measure of nonresponse [i.e., noninterview of designated respondents], for quota samples, mention of refusal rate)? 0. No 1. Yes

Details of Newspaper Study

Columns	Specification
18	Is the sampling error provided for *any* estimate? 0. No 1. Yes 2. Vague phrase with no numbers
19	Is the possibility of nonsampling errors ever mentioned, including doubts about the adequacy of specific question wordings? 0. No 1. Yes
20	Is the method of interview mentioned, e.g., face-to-face, telephone, mail, etc.? 0. No 1. Yes (*Note.* Mention of "interviews" alone is not sufficient.)
21	Are the question wordings reported? 0. No question wordings are reported 1. Some question wordings are reported 2. All question wordings for which results are given are reported
22	Are the wordings of code categories (i.e., response alternatives) reported? 0. No response categories reported 1. Some response categories reported 2. All response categories reported (must include mention of DK category)
79	Record Number 2. If Card 2 3. If Card 3
80	"C" is coded in column 80 of all clipping records

II. Codes Used in Newspaper Poll Data File

Columns	Specification
1	Country in which poll was done: 1. U.S.A. 2. U.K. 3. Other Europe 4. Canada 5. Other

Column	Specification
2	Country in which clipping was made: 1. U.S.A. 2. U.K.
3–6	I.D. Number of poll: 0001–4999 (polls from U.K. clippings) 5001–9999 (from U.S.A. clippings)
8–11	I.D. Number of clipping giving fullest report of the details of the poll
13–16	I.D. Number of clipping containing example of bad reporting of poll details
18–19	Type of poll sponsor: 1. Newspaper, magazine 2. Polling organization itself (e.g., the Gallup Poll or Harris Survey) 3. Political party or politician 4. Lobbying group 5. Business or business organization 6. National government 7. State or local government 8. Radio station 9. TV 10. Wire service 11. Combination of 1, 9 or 10 12. Academic research, i.e., basic social science research done under auspices of academic institution 13. Other 00. Not known
21–23	Agency conducting poll or survey: (unique I.D. for all poll and survey organizations)
24	Was the sample: 0. Totally unknown: not described in any report 1. A volunteer sample 2. A street sample, e.g., roving reporter 3. Described as: "representative," "scientific," "cross-section(al)," "random" 4. Other, e.g., full probability, quota sample, etc. 5. Census

Columns	Specification
25	Does the sampling frame consist of households or other units? 0. Unknown 1. Households 2. Other units 3. Electoral register
26	How was the selection of units accomplished? 0. Totally unknown: not described 1. Probability sample at least to cluster level 2. Other procedure 3. Not described, but survey described as "quota sample," "probability sample," or "scientific sample," "random sample," "cross-sectional" 4. Survey described as telephone survey with random-digit dialing 5. Survey described as telephone survey with other sample method or no method described 6. Volunteer sample 7. Haphazard method (e.g., roving reporter)
27	Is within household selection procedure known? 0. No, not described in any report 1. Yes, quotas used 2. Yes, full probability 3. Yes, other respondent selection strategy 9. Not applicable: no within household sampling required
28	Is there any mention of geographical area covered in the sample? 0. No mention whatsoever, universe cannot be determined 1. Yes, it is based on a national sample (48 or 50 states in U.S.A.) 2. Yes, it is restricted to a region (more than one state) 3. Yes, it is restricted to a state (not U.K.) 4. Yes, it is restricted to a single metropolitan area, county, or city. (*Note.* Metropolitan areas that overlap 2 or more states such as Washington, Kansas City SMSAs, etc. should be coded as 4 not 2.) 5. Yes, other (e.g., international)
29	Is target population restricted in other ways? 0. Includes at least all adults (ages 18+ or 21+) 1. Restricted to voting population or registered voters 2. Restricted to other special population (e.g., non-adults, elite, businesspersons)

Column	Specification
	7. Magazine or newspaper readers: survey conducted by including questionnaire in publication 8. Other haphazard sample of ill-defined population (e.g., roving reporter "survey") 9. Cannot be determined (use only if you have reason to believe code 0 does not apply)
30	Sample size: 0. Unknown 1. 0–100 2. 101–500 3. 501–1000 4. 1001–1500 5. 1501–2000 6. 2001 or more
31	How long was the field period? 0. Not known 1. 1 day 2. 2–3 days 3. 4–7 days 4. 8–14 days 5. 15–30 days 6. Between 1 and 2 months 7. More than 2 months
32	Is method of interview: 0. Not known 1. Face-to-face 2. Telephone 3. Mail or other self-completion
33–57	Subject matter of polls (0 = No; 1 = Yes) (*Code as many as apply*)
33	Presidential or prime ministerial popularity
34	Popularity of other political leaders and elected officials
35	Election polls and candidate preference polls
36	Evaluations of problems in the country
37	Party support
38	"Government" popularity
39	Attitudes toward current legislation or proposed legislation
40	Attitudes toward labor unions
41	Attitudes toward pay increases
42	Attitudes toward government monetary or fiscal policy

Column	Specification
43	Attitudes toward international affairs, war, ECM, etc.
44	Attitudes toward religion
45	Attitudes toward sports and leisure activities
46	Attitudes toward commercial products or services
47	Attitudes toward crime
48	Attitudes toward drugs
49	Attitudes toward race relations
50	Attitudes toward women and issues relating directly to women
51	Attitudes toward unemployment
52	Attitudes toward inflation
53	Attitudes toward defense or defense spending
54	Attitude toward other social issues (specify)
55	Measurements of consumer confidence
56	Measurements of business confidence
57	Other (specify)
79	Is this poll record derived from a single clipping in which the poll was not the focus of the story? (Not coded in Britain) 0. No 1. Yes
80	"P" is coded in column 80 of all records in poll file.

Notes

1. We have deposited a copy of the data from this study with the Roper Public Opinion Research Center (P.O. Box 1732, Yale Station, New Haven, CT 06520).

2. Recently, Paletz and others (1980) have reported the results of a similar content analysis exercise, using the *New York Times* Information Bank for three years (1973, 1975, 1977). While many of their conclusions about poor reporting of poll results are similar to ours (see discussion under Recommendation 3, Chapter 10), we found that the use of the *New York Times* Information Bank for such studies introduces some difficult problems (see note 3).

3. We considered using the *New York Times* Index (that is, the computerized *New York Times* Information Bank), but we became disillusioned with it because of the underreporting we discovered. For example, since the *New York Times* index does not contain a keyword for "public opinion polls," the names of the 19 survey research organizations that are indexed had to be used as surrogates. Use of such a surrogate yields an erroneous estimate of the frequency of use of poll data because polls or surveys conducted by organizations other than the 19 included in the index (or stories that do not mention the organization collecting the data) are not counted. Moreover, we discovered a substantial undercoverage of published reports even

when the report mentioned one of the 19 indexed organizations; such undercoverage apparently occurs because reprinted stories (for example, a *Washington Post* reprinting of a *New York Times* poll story) are not always included in the index. Consequently, although our choice of a data source is imperfect, we came to feel that it might have more potential than the *New York Times* index.

4. Fifty-two clippings were independently recoded by a second coder. (Only clipping variables were recoded since poll variables required one to first identify the best report of a particular poll.) Most clipping variables showed relatively high levels of intercoder agreement (median intercoder agreement exceeded 90 percent). Much of this agreement, however, appeared to reflect consensus about when an element (e.g., reporting of question wording in story) was entirely absent from a clipping.

Variables for which intercoder agreement was less than 75 percent are makred with an asterisk in the preceding code descriptions.

5. For many variables, coders were given option of coding a 9 if variable was "not applicable," e.g., because clipping did not cite a poll result and variable seeks to code how poll results were used in news article.

D

On the Review of Federal Surveys

In addition to the surveys conducted by the Census Bureau (described in appendix B), the government sponsors a wide range of surveys that are carried out under contract with private surveying organizations. We attempted to discover the full extent of government-sponsored survey activities by examining the docket for clearance of government sponsored surveys. Almost all government forms requesting information from the public, including those for government-sponsored surveys, must be submitted for approval to the forms clearance offices of the Office of Management and Budget (OMB) or the General Accounting Office (GAO) (Federal Reports Act of 1942 as amended, 44 USC 3501-3512).[1] In order to assess the government's use of surveys, we searched the docket of forms approved by the GAO (as of September 30, 1980) and the records of the currently approved (as of November 17, 1980) forms cleared by the OMB. (For a description of the procedures for approval of forms, see Executive Office of the President, [1976].)

Since the request-for-approval forms do not specifically identify surveys, the boundaries of our search included all forms for which the purpose of the information collected was listed as "program evaluation," "general purpose statistics," and "research", and the type of respondents was listed as indi-

viduals or households. We thereby eliminated forms whose purposes were described only as "applications for benefits", "regulatory or compliance", and "program planning or management." Forms for which multiple purposes were listed were kept if at least one purpose was within the bounds of our search. We also excluded forms if the respondents were limited to state or local governments, farms, businesses, or other institutions.

Within such a large net, a number of nonsurvey forms were caught. The definition of a survey that we used to narrow the scope was: reports of information about individuals or households that are initially collected for the purpose of generating aggregate statistics (rather than for obtaining information on a particular individual), but excluding forms that requested only physiological or meteorological reports. Use of this definition excluded such forms as those used for record-keeping, applications for benefits or employment, incident reports (for example, fish tagging, boating accidents), weather reports, and medical forms reporting the diagnosis and treatment of physical illnesses. Of the 339 OMB forms discovered in the initial search, 228 fit our definition of surveys; the GAO docket yielded 6 forms, of which only one was a survey.[2] Except for the coding of surveys as factual, quasi-factual, or subjective (see Chapter 2), all data were derived from the records of the surveys as recorded in the forms clearance dockets (that is, on form SF-83).

Notes

1. The major exemptions to the clearance process mandated by the act are forms used by the Internal Revenue Service, and forms completed only by government employees. Since the act covers only the executive branch, surveys sponsored by the judiciary or by congressional committees are not subject to the clearance process.

2. In reviewing the GAO docket (as of September 30, 1980) in the categories of program evaluation, statistical surveys or reports, and "other," in which respondents were listed as individuals or households, we found six eligible documents. Of these, only one was relevant—a screening survey to identify recent buyers and sellers of homes. The number of interviews involved was 4,200 and the total estimated burden was 420 hours. We have subsequently learned that the dearth of surveys submitted to GAO for approval results from the way in which GAO interprets its charge. They only require approval for forms that are *completed* (that is, actually filled out) by the public. This effectively eliminates all but mail surveys, since interviewers actually record the answers for personal and telephone surveys. Thus, the one survey we found is not necessarily indicative of a limited use of surveys by regulatory agencies. Indeed, some agencies, such as the Federal Trade Commission have active survey programs (see Chapter 2).

E

Details of Survey Results Cited in Union Carbide Example

All survey results cited in the Union Carbide example may be found in Union Carbide (1980a:Chapter 7). In the following pages, we reproduce from that report all parts of each question cited in Chapter 3.

The samples on which these percentages are based are described by Union Carbide (1980a:i), as follows:

> Two simultaneous surveys were conducted between September 26 and October 28, 1979, among matched national probability samples of Americans 18 years and older. One thousand interviews were completed in each of the two surveys. Interviews were conducted from the Seasonwein Associates headquarters in about 1,000 communities in 300 county assignments representative of the contiguous United States. A 71% cooperation rate among households in the probability sample was obtained.

Question Wording and Marginals

I am going to read you some statements. Are you inclined to agree or disagree with the statement I just read? [If agree] Are you inclined to feel the condition I just described slows down economic growth in this country, or don't you feel this is the case?

	Agree	Dis-agree	Don't know	Slows growth	Does not	Don't know	Not asked
The government spends more than it takes in and borrows too much.	84%	11%	5%	65%	6%	5%	16%
Inflation and taxes are greatly reducing saving by individuals.	82	15	3	70	8	3	18
Too many Americans just don't like to work very hard anymore.	71	26	3	59	9	3	29
American businesses can't make a profit competing with the cheaper foreign imports that are flooding the country.	67	29	4	59	6	3	33
Businessmen are [more] afraid to take risks than they used to be.	66	29	5	56	8	2	34
A few big companies divide up business between them, and there is just not much competition anymore.	60	34	6	46	10	4	40
High taxes leave businessmen with too little money to invest in new equipment and new products.	53	40	7	44	6	3	47

Details of Survey Results Cited in Union Carbide Example

	Agree	Dis-agree	Don't know	Slows growth	Does not	Don't know	Not asked
Government regulations make it necessary for businesses to devote so much money to things like cleaning up the environment and improving product safety that there is not enough money left for new equipment.	47	46	8	38	6	2	53

I am going to read some statements. Are you inclined to agree or disagree?

	Agree	Disagree	Don't know
When the economy is booming, most of the increase is kept by business in the form of higher profits, higher executive salaries, and fat expense accounts. The average person doesn't benefit very much.	67%	28%	5%
When the economy goes down, business keeps its profits high and cuts jobs and wages.	67	29	4
All companies need to make some profits to stay in business, but the profits that businesses are making today are much higher than they need to be.	67	28	6
Without high profits, business cannot invest in the new plants and equipment needed to keep our economy going.	66	29	5
A person does better working for a company making high profits than working for a company making low profits.	63	28	9

Is it your feeling that these types of regulations should be made more strict, made less strict, or kept about the way they are now?

	More strict	About the same	Less strict	Don't know
To protect employees from work conditions that cause health risks like cancer	70%	20%	6%	5%
To protect consumers from products that cause health risks like cancer	65	23	8	5
To protect consumers from products that cause accidents	60	28	8	4
Water pollution	60	28	8	4
To protect employees from work conditions that cause accidents	56	30	9	6
Regulating the introduction of new prescription drugs	53	28	14	5
Air pollution	52	29	15	5
Regulating banks and interest rates	43	30	20	8
Government antitrust laws—laws limiting the size of corporations	42	33	15	10
Laws and regulations that reduce the number of firms selling a product or providing a service	35	32	24	10

Details of Survey Results Cited in Union Carbide Example

Is it your feeling that the federal government is doing more than it should, less than it should, or about the right amount to regulate major corporations in areas like product safety and other things that have to do with protecting the public?

Doing more than it should	22%
Doing the right amount	27
Doing less than it should	43
Don't know	7

F

Intersurvey Comparisons: Early Work

This appendix provides details of the intersurvey comparisons from early studies (Cantril, 1944) which were summarized in Section 5.2. The summary measure of discrepancy between survey estimates used in these comparisons (index of dissimilarity) indicates the percentage of respondents who would have to be changed from one category of response to another in order to bring the two response distributions into complete agreement.

For all comparisons in this appendix, "don't know" and "no opinion" answers were excluded from the analysis. The chi-square values shown in the following pages are likelihood ratio chi-square statistics (L^2) computed for the illustrated table (excluding "don't know" and "no answer" categories) using Cantril's statement that samples were approximately of sizes 5,000 (*Fortune*), 3,500 (AIPO), 2,500 (NORC), and 1,200 (OPOR). (In listing, AIPO refers to Gallup's American Institute of Public Opinion; OPOR refers to Office of Public Opinion Research [Princeton]; *Fortune* refers to polls conducted by *Fortune* magazine; and NORC is the National Opinion Research Center.)

1. In general, do you think the government has gone too far or not far enough in asking people to make sacrifices for the war?

	AIPO (April 25, 1944)	*OPOR (April 20, 1944)*
Too far	12%	8%
About right	42	37
Not far enough	37	49
No opinion	9	6
	100	100

Index of Dissimilarity = 4.4; L^2 = 49.8, d.f. = 2

2. Which of these two things do you think the United States should try to do when the war is over: Stay out of world affairs as much as we can, or take an active part in world affairs?

	AIPO (April 25, 1944)	*OPOR (April 5, 1944)*
Stay out	18%	28%
Take active part	74	66
No opinion	8	6
	100	100

Index of Dissimilarity = 3.9; L^2 = 48.2, d.f. = 1

3. Which of these four statements comes *closest* to expressing your own opinion of Mr. Roosevelt?

	Fortune (*November 1940*)	*AIPO* (*October 24, 1940*)	
		K Form	*T Form*
A. In times like these, it is absolutely essential to have a man like Roosevelt for President.	25.5%	32%	27%
B. There may be some reasons against having Roosevelt for another four years, but on the whole it is the best thing to do.	27.8	19	22
C. While Roosevelt has done some good things, the country would be better off under Willkie for the next four years.	21.2	21	19
D. The re-election of Mr. Roosevelt for another four years would be a very bad thing for the country.	19.0	20	23
Don't know	6.5	8	9
	100.0	100	100

Index of Dissimilarity = 3.9; L^2 = 64.2, d.f. = 3 (*Note.* Results for AIPO's K and T forms averaged to make comparison.)

4. Which of the following statements comes *closest* to expressing your own opinion of Mr. Roosevelt?

	Fortune (October 1940)	*AIPO* (October 19, 1940) K Form	*AIPO* (October 19, 1940) T Form
A. In times like these, it is absolutely essential to have a man like Roosevelt for President.	26.7%	29%	27%
B. There may be some reasons against having Roosevelt for another four years, but on the whole it is the best thing to do.	26.5	22	23
C. While Roosevelt has done some good things, the country would be better off under Willkie for the next four years.	22.3	19	19
D. The re-election of Mr. Roosevelt for another four years would be a very bad thing for the country.	16.4	21	22
Don't know	8.1	9	9
	100.0	100	100

Index of Dissimilarity = 3.6; L^2 = 54.0, d.f. = 3

5. Which of these four statements comes *closest* to expressing your opinion of Wendell Willkie?

	Fortune (October 1940)	*AIPO* (October 19, 1940) K Form	*AIPO* (October 19, 1940) T Form
A. Willkie is just the man the country needs for President during the next four years.	11.5%	16%	14%
B. Even though Willkie hasn't as much political and international experience as he needs, he still would make a better President than Roosevelt.	24.1	25	26
C. Willkie is probably an honest and capable businessman but he hasn't the right experience to be President in times like these.	36.1	33	34
D. The election of a man like Mr. Willkie would be a very bad thing for the country.	9.4	15	14
Don't know	18.9	11	12
	100.0	100	100

Index of Dissimilarity = 3.2; L^2 = 48.5, d.f. = 3

6. Which of these four statements comes *closest* to expressing your opinion of Wendell Willkie?

	Fortune (November 1940)	*AIPO* (October 24, 1940) K Form	*AIPO* (October 24, 1940) T Form
A. Willkie is just the man the country needs for President during the next four years.	11.1%	14%	16%
B. Even though Willkie hasn't as much political and international experience as he needs, he still would make a better President than Roosevelt.	26.8	28	27
C. Willkie is probably an honest and capable businessman but he hasn't the right experience to be President in times like these.	36.5	32	31
D. The election of a man like Mr. Willkie would be a very bad thing for the country.	13.8	16	17
Don't know	11.8	10	9
	100.0	100	100

Index of Dissimilarity = 3.7; $L^2 = 57.3$, d.f. = 3

7. If the war is STILL GOING ON and President Roosevelt runs for the Democrats against Governor Dewey for the Republicans, how do you think you will vote?

	AIPO (March 29, 1944)	*OPOR (March 1, 1944)*
Roosevelt	51%	51%
Dewey	41	32
No opinion	8	17
	100	100

Index of Dissimilarity = 2.2; L^2 = 11.3, d.f. = 1

8. Taking everything into consideration, which side do you think will win the war?

	Fortune (December 1941)	*NORC (November 1941)*
Allies	72.5%	69.0%
Axis	7.0	4.2
Neither	4.3	3.5
Don't know	16.2	23.3
	100.0	100.0

Index of Dissimilarity = 1.5; L^2 = 18.5, d.f. = 2

9. After this war is over, do you think every young man should be required to serve one year in the Army or Navy?

	AIPO (December 1942)	*OPOR (April 20, 1944)*
Yes	67.4%	70%
No	26.6	22
No opinion	6.0	8
	100.0	100

Index of Dissimilarity = 1.7; L^2 = 8.3, d.f. = 1

10. Do you think Russia can be trusted to cooperate with us when the war is over?

	AIPO (April 6, 1943)	*OPOR (March 30, 1943)*
Yes	45%	42%
No	34	39
No opinion	21	19
	100	100

Index of Dissimilarity = 2.0; L^2 = 7.6, d.f. = 1

11. Which country is the greatest military threat to the United States—Germany or Japan?

	AIPO (March 24, 1943)	*OPOR (March 16, 1943)*
Germany	25%	29%
Japan	53	53
Both	18	14
No opinion	4	4
	100	100

Index of Dissimilarity = 1.6; L^2 = 14.2, d.f. = 2

12. With which of these statements do you come closest to agreeing?

	AIPO (September 15, 1944)	*OPOR (October 5, 1944)*
A. Dewey has an excellent record and it would be the best thing for the country to elect him President for the next four years.	25.1%	24.9%
B. While Dewey may not be ideal for the job, it certainly would be better to have him than Roosevelt for the next four years.	19.3	16.7
C. Although Dewey may have some very good qualities, he still could not do the job as well as Roosevelt during the next four years.	34.3	29.1
D. It would be a very bad thing for this country to elect Dewey as President for the next four years.	15.0	19.1
Don't know	6.3	10.2
	100.0	100.0

Index of Dissimilarity = 1.9; L^2 = 20.0, d.f. = 3

13. Which party do you want to see win the Presidential election next year—the Democrats or the Republicans?

	AIPO (March 24, 1943)	*OPOR (March 30, 1943)*
Democrats	41%	43%
Republicans	30	29
Other	1	2
No opinion	22	23
Qualified answer	6	3
	100	100

Index of Dissimilarity = 1.0; L^2 = 6.7, d.f. = 2

14. Do you think Russia can be trusted to cooperate with us when the war is over?

	AIPO (January 18, 1944)	*OPOR (January 24, 1944)*
Yes	40%	42%
No	36	33
No opinion	24	25
	100	100

Index of Dissimilarity = 1.3; L^2 = 3.1, d.f. = 1

15. If the war is over and Roosevelt runs for a fourth term next year, do you think you will vote for him or against him?

	AIPO (*February 23, 1943*)	*OPOR* (*March 30, 1943*)
For	37%	40%
Against	50	49
No opinion	13	11
	100	100

Index of Dissimilarity = 0.9; L^2 = 1.9, d.f. = 1

16. If the war is OVER and Roosevelt runs for the Democrats against Dewey, how do you think you will vote?

	AIPO (*March 29, 1944*)	*OPOR* (*March 1, 1944*)
Roosevelt	35%	30%
Dewey	55	51
No opinion	10	19
	100	100

Index of Dissimilarity = 0.7; L^2 = 1.1, d.f. = 1

17. With which of these statements do you come closest to agreeing?

	Fortune (*September 7, 1944*)	*OPOR* (*October 5, 1944*)
A. Roosevelt has done an excellent job and it's very important that he should be President during the next four years.	26.3%	25.5%
B. While he has made mistakes and he's been in office a long time, it is still better to elect Roosevelt President again for the next four years.	28.3	26.0
C. Although Roosevelt has done some good things, he has been President long enough and the country would be better off to elect Dewey for the next four years.	27.3	28.7
D. It would be a very bad thing for this country to re-elect Roosevelt for another four years.	13.6	12.5
Don't know	4.5	7.3
	100.0	100.0

Index of Dissimilarity = 0.7; L^2 = 2.9, d.f. = 3

18. If our military leaders believe it will be necessary to bomb historic religious buildings and shrines in Europe, would you approve or disappove of their decisions?

	AIPO (March 29, 1944)	*OPOR (March 1, 1944)*
Approve	77%	79%
Disapprove	17	17
No opinion	6	4
	100	100

Index of Dissimilarity = 0.2; L^2 = 0.1, d.f. = 1

19. Do you think the United States is doing all it possibly can to win the war?

	NORC (April 8, 1944)	*OPOR (April 20, 1944)*
Yes	71%	74%
No	21	23
Don't know	8	3
	100	100

Index of Dissimilarity = 0.4; L^2 = 0.4, d.f. = 1

20. Which party to you want to see win the presidential election next year—the Democrats or the Republicans?

	AIPO (March 10, 1943)	*OPOR (March 30, 1943)*
Democrats	61%	62%
Republicans	39	38
	100	100

Index of Dissimilarity = 0.4; L^2 = 0.4, d.f. = 1

G

Intersurvey Comparisons: Recent Work

Details are provided in tables G.1 through G.4 of the intersurvey comparisons reviewed in recent studies (Smith, 1978, 1979, 1981, 1982; Turner, 1981a) and summarized in Section 5.2.

For dichotomous (or dichotomized variables), discrepancies are the absolute value of the percentage-point difference between estimates for one response category (for example, if the variable was gender and the estimates of the male-female ratio were 48–52 and 50–50, the discrepancy would be 2 percentage points). For variables with more than two response categories, a comparable value is provided by the index of dissimilarity (*D*); it indicates the percentage of respondents who would have to be changed from one category of response to another in order to bring two response distributions into complete agreement. The notes to each table indicate whether the "don't know/no opinion" category of responses was kept in the analysis. Table G.5 presents a detailed tabulation of the distribution of discrepancies found in these comparisons and those shown in Appendix F.

TABLE G.1
Intersurvey Comparisons from Smith (1978)

Question (Mnemonic): Response Category	Time	House	Proportion	N	Difference (T_1–T_2)
Capital punishment (CAPPUN): Execute	2/72	AIPO	.510	1,507	−.020
	3/72	GSS	.530	1,609	
	11/72	AIPO	.570	NA	−.032
	3/73	GSS	.602	1,492	
	3/76	GSS	.655	1,496	−.010
	4/76	AIPO	.666	1,540	
Ideal number of children (CHLDIDEL): 4 or more	2/74	AIPO	.177	1,562	−.104
	3/74	GSS	.191	1,484	
Judicial punishment (COURTS): Harsher	12/72	AIPO	.744	1,504	.013
	3/73	GSS	.731	1,494	
Women for president (FEPRES): Vote for	3/75	GSS	.778	1,489	−.043
	8/75	AIPO	.735	1,515	
Suitability for politics (FEPOLY): Men	3/74	GSS	.326	730	−.061
	9/74	SRC	.387	488	
Marijuana law (GRASS): Legalize	1/73	AIPO	.157	1,508	−.026
	3/73	GSS	.183	1,501	
Gun permit (GUNLAW): Opposes	3/72	GSS	.702	1,610	−.014
	5/72	AIPO	.716	1,540	

	2/75	SRC	.706	452	−.021
	3/75	GSS	.737	1,488	
	2/76	SRC	.726	638	.011
	3/76	GSS	.715	1,493	
Evaluation of people (MISANTHROPY):					
FAIR: Fair	11/72	SRC	.589	2,179	.016
	3/73	GSS	.573	1,503	
	11/74	SRC	.576	1,543	−.040
	3/75	GSS	.616	1,488	
HELPFUL: Helpful	11/72	SRC	.469	2,174	.001
	3/73	GSS	.468	1,501	
	11/74	SRC	.507	1,528	−.055
	3/75	GSS	.562	1,488	
TRUST: Trusts people	11/72	SRC	.458	2,179	−.001
	3/73	GSS	.459	1,502	
	11/74	SRC	.466	1,551	−.073
	3/75	GSS	.393	1,485	
Spending for foreign aid (NATAID): Too little	12/73	Roper	.021	1,766	−.009
	3/74	GSS	.030	1,481	
Spending for military (NATARMS): Too little	12/73	Roper	.150	1,764	−.019
	3/74	GSS	.169	1,479	
Spending for cities (NATCITY): Too little	12/73	Roper	.432	1,757	−.067
	3/74	GSS	.499	1,474	
Spending for crime prevention (NATCRIME): Too little	12/73	Roper	.640	1,757	−.026
	3/74	GSS	.666	1,481	
Spending for drug prevention (NATDRUG): Too little	12/73	Roper	.579	1,759	−.021
	3/74	GSS	.600	1,478	

TABLE G.1 *(continued)*

Question (Mnemonic): Response Category	Time	House	Proportion	N	Difference (T_1–T_2)
Spending for education (NATEDUC): Too little	12/73	Roper	.458	1,752	−.049
	3/74	GSS	.507	1,474	
Spending for environment (NATENVIR): Too little	12/73	Roper	.459	1,766	−.131
	3/74	GSS	.590	1,476	
Spending for welfare (NATFARE): Too little	12/73	Roper	.170	1,762	−.051
	3/74	GSS	.221	1,481	
Spending for medical care (NATHEAL): Too little	12/73	Roper	.599	1,761	−.040
	3/74	GSS	.639	1,477	
Spending for space exploration (NATSPAC): Too little	12/73	Roper	.035	1,768	−.042
	3/74	GSS	.077	1,480	
Political identification (PARTYID): Democratic	11/72	SRC	.403	2,702	−.018
	3/73	GSS	.411	1,493	
	11/74	SRC	.385	1,570	−.019
	3/75	GSS	.404	1,486	
	11/76	SRC	.396	2,244	−.049
	3/77	GSS	.444	1,520	
Work if rich (RICHWORK): Continue working	Winter 1972/73	SRC	.658	2,148	−.023
	3/73	GSS	.681	831	

NOTE: "Don't know/No opinion" responses were maintained in the analysis. Mnemonic Labels in column one may be ignored in reading this table; they are shown to provide a cross-reference to the complete question wordings which are listed at the end of this appendix.

TABLE G.2

Alternative Estimates of the Proportion of People Having a "Great Deal of Confidence" in the People Running Various Institutions (from Smith 1981; see also Turner and Krauss 1978)

	Survey											
Institution[a]	**Harris 2319**	**GSS 1973**	**Harris 2354**	**NORC 4179**	**Harris 7482**	**GSS 1974**	**Harris 2515**	**GSS 1975**	**Harris 7581**	**Harris 2521**	**Harris 7681**	**GSS 1976**
Business	.338	.325	.276	.218	.241	.314	.181	.193	.197	.163	.215	.220
Religion	.332	.331	.289	.321	—	.443	—	.244	.322	.237	—	.307
Government	—	—	.134	.142	.117	.136	—	.133	.131	.108	.165	.135
Labor	.229	.142	.162	.187	—	.182	—	.101	—	.099	—	.116
Press	—	—	.278	.251	—	.259	—	.239	.259	.201	.213	.285
Medicine	.629	.553	—	—	—	—	—	.505	.428	.420	—	.541
Supreme Court	—	—	—	.341	—	.332	—	.308	.287	.219	.316	.354
Science	—	—	—	—	—	—	—	.377	.479	—	—	—
Congress	—	—	.171	.227	—	.171	.124	.133	.136	.088	.179	.137
Military	—	—	—	—	—	—	.267	.352	.245	.225	.362	.392
Banks, etc.	—	—	—	—	—	—	—	.319	.415	.335	—	.395
Date of Survey	3/73	3/73	12/73	1/74	2/74	3/74	3/75	3/75	4/75	2/76	3/76	3/76

SOURCE: Smith (1981, Table 6); results for Labor in Harris 2515 and 7581 have been deleted since the label used (labor unions) was different from that used by GSS 1975 (organized labor). Otherwise, labels are all the same or very similar (see Smith, 1981, p. 137). Samples were of adult population (18+) of the continental United States (noninstitutionalized), except that Harris 2319 used electoral participants (GSS 1973 was adapted to approximate this universe). See Smith (1981) for further details on surveys.

NOTES: Summary tabulation (Chapter 5, Table 5.8) includes all across house comparisons that were possible in the following pairings: Harris 2319 vs. GSS 1973; Harris 2354 vs. NORC 4179; Harris 7482 vs. NORC 4179; Harris 7482 vs. GSS 1974; Harris 2515 vs. GSS 1975; Harris 7581 vs. GSS 1975; Harris 2521 vs. GSS 1976; Harris 7681 vs. GSS 1976; Except for Harris survey 2319 (N = 2,993), all surveys had sample sizes of approximately 1,500. "Don't know/no opinion" responses were maintained in the analysis.

[a]Institutions were described as listed in stub of table (sometimes with a "the," except as follows: "major companies" (Business), "organized religion" (Religion), "executive branch of the federal government" (Government), "the U.S. Supreme Court" (Supreme Court), "scientific community" (Science), "banks and financial institutions" (Banks, etc.), "organized labor" (Labor).

TABLE G.3
Intersurvey Comparisons from Smith (1982)

A1. RESPONDENT'S POSITION ON DEFENSE SPENDING

Survey Organization	N	1	2	3	4	5	6	7
SRC	1,658	3.1%	2.2%	5.0%	16.6%	26.7%	23.4%	23.0%
NORC	1,303	3.6%	3.8%	7.4%	23.6%	25.7%	14.5%	21.3%

NOTE: 1: "Greatly decrease spending;" 7: "Greatly increase spending."
Index of dissimilarity = 5.7

A2. FEDERAL GOVERNMENT'S POSITION ON DEFENSE SPENDING

Survey Organization	N	1	2	3	4	5	6	7
SRC	1,521	3.3%	5.9%	18.5%	27.9%	26.0%	12.1%	6.4%
NORC	1,207	4.9%	10.4%	25.6%	26.5%	16.2%	9.5%	6.8%

NOTE: 1: "Greatly decrease spending;" 7: "Greatly increase spending."
Index of dissimilarity = 6.8

B1. RESPONDENT'S POSITION ON ASSISTANCE FOR MINORITIES

Survey Organization	N	1	2	3	4	5	6	7
SRC	855	6.2%	4.8%	8.9%	24.4%	19.6%	17.9%	18.1%
NORC	1,361	5.0%	5.1%	10.2%	24.9%	17.8%	15.2%	21.7%

NOTE: 1: "Greatly decrease spending;" 7: "Greatly increase spending."
Index of dissimilarity = 2.7

B2. FEDERAL GOVERNMENT'S POSITION ON ASSISTANCE FOR MINORITIES

Survey Organization	N	1	2	3	4	5	6	7
SRC	750	15.3%	23.5%	22.7%	16.3%	12.5%	6.3%	3.3%
NORC	1,250	19.4%	22.0%	19.8%	16.5%	10.5%	7.8%	4.1%

NOTE: 1: "Greatly decrease spending;" 7: "Greatly increase spending."
Index of dissimilarity = 3.1

C1. RESPONDENT'S POSITION ON SOCIAL SPENDING

Survey Organization	N	1	2	3	4	5	6	7
SRC	1,625	7.2%	9.8%	13.6%	18.3%	16.8%	15.0%	19.4%
NORC	1,308	9.3%	9.4%	13.7%	20.7%	14.9%	12.1%	20.0%

NOTE: 1: "Greatly decrease spending;" 7: "Greatly increase spending."
Index of dissimilarity = 2.6

TABLE G.3 *(continued)*

C2. FEDERAL GOVERNMENT'S POSITION ON SOCIAL SPENDING

Survey Organization	N	1	2	3	4	5	6	7
SRC	1,479	2.5%	2.9%	9.7%	21.0%	25.0%	22.8%	16.2%
NORC	1,224	2.7%	5.1%	11.2%	21.6%	21.5%	19.7%	18.3%

NOTE: 1: "Greatly decrease spending;" 7: "Greatly increase spending."
Index of dissimilarity = 3.3

D. SPENDING FOR SPACE EXPLORATION

Survey Organization	N	Too Much	Too Little	About Right
SRC	717	47.7%	15.8%	36.5%
NORC	1,344	42.7%	19.8%	37.6%

Index of dissimilarity = 2.3

E. SPENDING ON ENVIRONMENT

Survey Organization	N	Too Much	Too Little	About Right
SRC	750	14.7%	42.8%	42.5%
NORC	1,382	16.4%	50.7%	32.9%

Index of dissimilarity = 4.4

F. SPENDING ON HEALTH

Survey Organization	N	Too Much	Too Little	About Right
SRC	768	9.1%	51.6%	39.3%
NORC	1,407	8.0%	57.0%	35.0%

Index of dissimilarity = 2.5

G. SPENDING ON CITIES

Survey Organization	N	Too Much	Too Little	About Right
SRC	677	22.5%	45.3%	32.2%
NORC	1,278	24.4%	45.8%	29.8%

Index of dissimilarity = 1.1

TABLE G.3 *(continued)*

H. SPENDING TO HALT CRIME

Survey Organization	N	Too Much	Too Little	About Right
SRC	749	5.3%	66.7%	28.0%
NORC	1,400	6.3%	72.0%	21.7%

Index of dissimilarity = 2.9

I. SPENDING ON DEALING WITH DRUG ADDICTION

Survey Organization	N	Too Much	Too Little	About Right
SRC	739	8.0%	58.7%	33.3%
NORC	1,353	8.4%	64.5%	27.1%

Index of dissimilarity = 2.8

J. SPENDING ON EDUCATION

Survey Organization	N	Too Much	Too Little	About Right
SRC	771	10.9%	54.6%	34.5%
NORC	1,404	10.6%	54.9%	34.5%

Index of dissimilarity = 0.2

K. SPENDING ON IMPROVING THE CONDITION OF BLACKS

Survey Organization	N	Too Much	Too Little	About Right
SRC	727	27.4%	26.0%	46.6%
NORC	1,347	25.9%	26.2%	47.9%

Index of dissimilarity = 0.7

L. SPENDING ON THE MILITARY

Survey Organization	N	Too Much	Too Little	About Right
SRC	767	11.7%	65.5%	22.8%
NORC	1,370	12.3%	60.2%	27.5%

Index of dissimilarity = 2.4

TABLE G.3 *(continued)*

M. SPENDING ON FOREIGN AID

Survey Organization	N	Too Much	Too Little	About Right
SRC	781	79.0%	3.5%	17.5%
NORC	1,389	73.7%	5.4%	20.9%

Index of dissimilarity = 2.4

N. SPENDING ON WELFARE

Survey Organization	N	Too Much	Too Little	About Right
SRC	775	59.2%	15.5%	25.3%
NORC	1,401	59.0%	14.0%	27.0%

Index of dissimilarity = 0.8

NOTE: "Don't know/No opinion" responses were excluded from all analyses.

TABLE G.4
Intersurvey Comparisons from Turner (1981a) and Smith (1979)

A. FERTILITY

1. Percentage of married women aged 18–39 expecting no more children in next five years

	1972	1974	1975	1977
Census-CPS[a]	67%	68%	68%	68%
NORC[b]	69	69	61	62

2. Percentage of married women aged 18–39 expecting no more children in all future years

	1972	1974	1975	1977
Census-CPS[a]	64%	67%	66%	67%
NORC[b]	67	66	59	58

B. HAPPINESS: PERCENTAGE REPORTING SELVES TO BE "VERY HAPPY."

	1973	1974
NORC-CNS[c]	33.1%	28.0%
NORC-GSS[d]	35.9	37.9

[a]Census-CPS surveys conducted in month of June; sample sizes were approximately 4,000 in each year. Estimates are projections to national population.
[b]NORC surveys conducted in month of March; sample sizes ranged from 201–246. To provide projections to national population, NORC data were weighted in proportion to the number of (eligible) adults in household.
[c]Dates of surveys: April–May 1973, and March 1973; sample sizes: 723 and 610.
[d]GSS surveys conducted in month of March; sample sizes: 1,496 and 1,476.

TABLE G.5

Summary Tabulation of Discrepancies

	Source of Comparison					
Discrepancy	**Appendix F**	**Table G.1**	**Table G.2**	**Table G.3**	**Table G.4**	**Total**
0.0–0.9	6	3	11	3	—	23
1.0–1.9	6	9	4	1	—	20
2.0–2.9	2	6	7	8	2	25
3.0–3.9	5	1	4	2	1	13
4.0–4.9	1	6	2	1	—	10
5.0–5.9	—	2	4	1	1	8
6.0–6.9	—	2	1	1	1	5
7.0–7.9	—	1	6	—	1	8
8.0–8.9	—	—	3	—	—	3
9.0–9.9	—	—	2	—	2	4
10.0–10.9	—	—	2	—	—	2
11.0–11.9	—	—	—	—	—	—
12.0–12.9	—	—	2	—	—	2
13.0–13.9	—	1	1	—	—	2
14.0–14.9	—	—	—	—	—	—
15.0–15.9	—	—	—	—	—	—
16.0–16.9	—	—	1	—	—	1
Totals	20	31	50	17	8	126

I. Wording of Questions Cited in Table G.1 (Listed by Acronym)

CAPPUN

(a) AIPO and NORC-GSS 1972–73: Are you in favor of the death penalty for persons convicted of murder?

(b) NORC-GSS 1974–76: Do you favor or oppose the death penalty for persons convicted of murder?

CHLDIDEL

(a) AIPO and NORC-GSS: What do you think is the ideal number of children for a family to have?

(b) AIPO has response "No opinion" while NORC-GSS has responses "As many as you want" and "Don't know."

COURTS

(a) AIPO and NORC-GSS: In general, do you think the courts in this area deal too harshly or not harshly enough with criminals?

FEPRES

(a) AIPO omits "were" from NORC-GSS question below.

(b) NORC-GSS: If your party nominated a woman for President, would you vote for her if she were qualified for the job?

FEPOLY

(a) SRC and NORC-GSS: Would you say that most men are better suited for politics than are most women, that men and women are equally suited, or that women are better suited than men in this area?

GRASS

(a) AIPO and NORC-GSS: Do you think the use of marijuana should be made legal, or not?

GUNLAW

(a) AIPO and NORC-GSS: Would you favor or oppose a law which would require a person to obtain a police permit before he or she could buy a gun?

(b) SRC omits "or she."

MISANTHROPY

(a) SRC and NORC-GSS:

(1) FAIR: Do you think most people would try to take advantage of you if they got a chance, or would they try to be fair?

(2) HELPFUL: Would you say that most of the time people try to be helpful, or that they are mostly just looking out for themselves?

(3) TRUST: Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?

(b) NORC-GSS accepts responses of "depends" and "don't know." SRC uses "don't know" only.

NATIONAL PROBLEMS (NATAID to NATSPACE)

(a) NORC-GSS and Roper: We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount. First (READ ITEM A) . . . are we spending too much, too little, or about the right amount on (ITEM)?

A. Space exploration program

B. Improving and protecting the environment

C. Improving and protecting the nation's health

D. Solving the problems of the big cities

E. Halting the rising crime rate
F. Dealing with drug addiction
G. Improving the nation's education system
H. Improving the conditions of Blacks (Roper omits "H. Improving the conditions of Blacks.")
I. The military, armaments and defense
J. Foreign aid
K. Welfare

PARTYID

(a) SRC and NORC-GSS: Generally speaking, do you usually think of yourself as a Republican, [a] Democrat, [an] Independent, or what?
(b) NORC-GSS has a precoded response of "other" each year and added the response "no preference" in 1975–1977. SRC includes both categories each year.
(c) SRC includes the bracketed articles.

RICHWORK

(a) NORC-GSS: If you were to get enough money to live as comfortably as you would like for the rest of your life, would you continue to work or would you stop working?
(b) SRC uses: "you'd" instead of "you would" and omits "or would you stop working."

II. Wording of Questions Cited in Table G.2

As far as the people running (institution) are concerned, would you say you have a great deal of confidence, only some confidence, or hardly any confidence at all in them?

Note. Footnote to table G.2 provides the labels used to describe the various instititutions. Minor variations have appeared in the basic question (see Turner and Krauss 1978; Smith 1981) and in the wording of the labels. The data presented have been limited to instances in which the labels were identical or extremely similar; thus, for example, measurements of confidence in "labor unions" were *not* compared with measurements of confidence in "organized labor."

III. Wording of Questions Cited in Table G.3

Seven-point scales, NORC-GSS

A. Some people believe that we should spend much less money for defense. Suppose these people are at one end of the scale at point num-

ber 1. Others feel that defense spending should be greatly increased. Suppose these people are at the other end, at point 7. And, of course, some other people have opinions somewhere in between at points 2, 3, 4, 5, or 6. [HAND CARD]

DEFENSE SPENDING

01	02	03	04	05	06	07
Greatly Decrease Defense Spending						Greatly Increase Defense Spending

A–1 Where would you place yourself on this scale, or haven't you thought much about this?

Haven't thought much 08
Don't know 98

A–2 Where would you place what the federal government is doing at the present time?

Haven't thought much 08
Don't know 98

B. Some people feel that the government in Washington should make every possible effort to improve the social and economic position of blacks and other minority groups, even if it means giving them preferential treatment. (Suppose these people are at one end of the scale at point number 1.) Others feel that the government should not make any special effort to help minorities because they should help themselves. (Suppose these people are at the other end, at point 7. And of course, some other people have opinions somewhere in between at points 2, 3, 4, 5, or 6.) [Format for reply same as before.]

C. Some people think the government should provide fewer services, even in areas such as health and education, in order to reduce spending. Other people feel it is important for the government to continue the services it now provides even if it means no reduction in spending. [Format for reply same as before.]

Seven-point scales, SRC

SRC questions are the same except for the addition of intervening items; for example, while NORC asks only about where respondent and federal government stand on issues, SRC asks for position of: Respondent, Jimmy Carter, Ronald Reagan, Ted Kennedy, John Connally, Jerry Brown, Howard Baker, George Bush, and then for the position of federal government.

Three-point scales: NORC-GSS and SRC

First I would like to talk with you about some things people think about today. We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount. First (ITEM) . . . are we spending too much, too little, or about the right amount on (ITEM)? READ EACH ITEM.

D. Space exploration program
E. Improving and protecting the environment
F. Improving and protecting the nation's health
G. Solving the problems of the big cities
H. Halting the rising crime rate
I. Dealing with drug addiction
J. Improving the nation's education system
K. Improving the conditions of Blacks
L. The military, armaments and defense
M. Foreign aid
N. Welfare

III. Wording of Questions Cited in Table G.3

A. *Fertility Expectations*

Census Bureau, 1972–74

Do you expect to have any (more) children?
How many (more) do you expect to have?
How many (more) do you expect to have in the next five years?

Census Bureau, 1975–77

Looking ahead, do you expect to have any (more) children?
How many (more) do you expect to have?
How many (more) do you expect to have in the next five years?

NORC-GSS, 1972–77

Do you expect to have any (more) children?
How many (more)?
How many (more) in the next five years?

B. *Happiness, NORC-GSS and NORC-CNS*

Taken all together, how would you say things are these days—would you say that you are very happy, pretty happy, or not too happy?

H

Scheme for Classifying Survey Questions According to Their Subjective Properties

Many survey researchers treat survey questions about matters of fact as generally more reliable than questions about respondents' feelings and attitudes. One reason for doing so is that it is easier to agree on a common interpretation for an objective question referring to external, observable events than a question referring to subjective states. Respondents who are asked, "Have you purchased a black and white television set in the past six months?" (for example) presumably have a clear and common interpretation of the event in question. In contrast, respondents who are asked, "Taken all together, how would you say things are these days—would you say that you are very happy, pretty happy, or not too happy?" may or may not share a common interpretation (with each other or with the investigator) of what it means to be happy. In addition, in the former case it would be possible (at least in principle) to compare a respondent's report with an external criterion.

This appendix takes as its starting point the intuitive notion that answers to different types of survey questions are not equally valid and that the validity and reliability of a question depends somehow on the extent, and in what way, it is subjective. We evaluate three senses in which a survey question can be subjective: the referent, the nature of the question, and the object.

First, a question can focus on respondents' internal subjective states rather than on external, observable events. Second, different questions pose different mental tasks, and the kind of mental problem a respondent is trying to solve—recall, generalization, introspection, prediction, evaluation, and so on—influences the sorts of errors and biases that are likely to characterize his response. (In this sense, all survey questions are subjective in that answers to them are products of a thought process.) Third, a survey question can be subjective in the sense that respondents are asked to take themselves as the object of report and provide information about their own behaviors, possessions, attitudes, and so on. Each separate meaning of subjective is taken as a dimension of the scheme displayed in Figure H.1. The scheme can be used to classify survey questions, yielding a three-way cross-classification of survey questions according to their subjective properties as presented in Table H.1.[1]

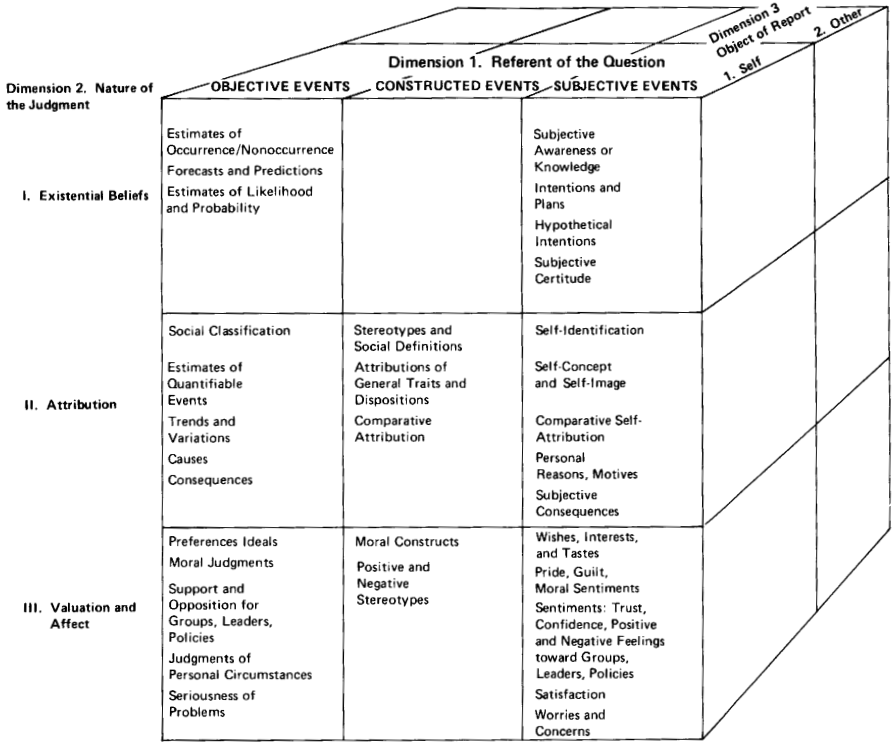
At the outset, it should be noted that the classification scheme diagrammed in Figure H.1 is provisional, simplistic, sometimes arbitrary and incomplete, and intended mainly for heuristic purposes. Sometimes, when the task requirements seem ambiguous, a question is classified in more than one category. The careful reader will note that some of the categories built into the scheme prove on examination to be based on distinctions that are less clearcut than they seem to be at first. The sophisticated reader who is well-versed in philosophical and linguistic analyses of the meaning of "subjective" and "objective" will note that these concepts are used here in a naive and intuitive fashion. We hope that the careful reader will be stimulated to think up a better scheme for analyzing meaningful distinctions among survey questions, and that the sophisticated reader will be moved to apply his or her knowledge of the conceptual issues to clarify the measurement problems that confront survey researchers.

Dimension 1. The Referent of a Question: Objective Versus Subjective Events

We classify as objective those questions referring to external events that could in principle be observed and reported by parties in addition to the

FIGURE H.1

Dimensions of subjectivity in survey questions.



respondent (see Table H.1 for examples). We classify as subjective those questions that refer to a respondent's internal states of awareness, knowledge, intention, belief, sentiment, and so forth, which are in principle directly knowable only by her.

We also distinguish a third category of questions, constructed events, which combine subjective and objective elements. This category includes questions that give respondents a construct or category and ask them to supply events or attributes to fit it or that refer to a general ambiguous event for which respondents must supply meaning. For example, an event may be real, but defined in terms of the respondent's own feelings; the event is constructed by the respondent based on his own idiosyncratic subjective reactions: "Do you sometimes drink more than you think you should?" This category also includes questions about phenomena that are labeled in vague, quasi-evaluative, loaded terms: for example, "Do you feel the negotiations in Paris will end in an honorable settlement or not?" or

"Do you have anyone close to you who is or has been mentally disturbed?" Such questions refer to events that cannot be easily reduced to a set of observable occurrences, nor are they wholly subjective; rather, they are cultural constructs and labels used to interpret and make sense of the social world.[2] The third kind of questions in this category refers to general motives and dispositions attributed to others: for example, "Do you agree or disagree that the Russian leaders are basically trying to get along with America?" This is a constructed event, since it does not correspond to a set of observable occurrences nor is it wholly subjective.

Because objective referents are tied to concrete events, we expect to find greater consensus and stability among respondents about their meaning than about subjective or constructed events. Therefore, in order to ensure that survey respondents interpret questions in a comparable fashion, surveys often try to refer to events in concrete, behaviorally specific, and neutral terms, so that respondents do not have to supply their own interpretations of the referent of a question. For example, in the National Crime Survey (NCS), the interviewer says, "Now I'd like to ask you some questions about crime," and then proceeds to question the respondent about a variety of concrete incidents that may have taken place. Crime is defined for respondents in terms of a fairly specific set of events; for example, for assaultive violence: "Did anyone beat you up, attack you or hit you with something, such as a rock or bottle?" "Were you knifed, shot at, or attacked with some other weapon . . . ?" "Did anyone *threaten* to beat you up or *threaten* you with a knife, gun or some other weapon . . . ?" "Did anyone *try* to attack you in some other way?"

Yet a fundamental problem with such questions is that the meanings of basic concepts—such as crime, threat, attack, and weapon—are still undefined and depend on respondents' interpretations. Such interpretations normally take into account the inferred intent behind an offender's action, the circumstances in which an incident occurs, and a variety of other situational cues.

Experimental evidence demonstrates the contextual variability of people's interpretations of incidents of violence. For example, Duncan (1976) recruited white male college students to observe and classify a videotaped interaction between two male confederates; the races of the confederates were the experimental variables. The behavior of the confederates was programmed by a script and included an ambiguous incident in which one confederate behaved aggressively toward the other. Subjects' interpretations of the incident depended on the race of the aggressor: 70 percent of the subjects classified a black aggressor's action as "violent behavior" (rather than "aggressive behavior," "dramatizing," or "playing around"), while only 12 percent classified a white aggressor's action as "violent." Interestingly,

subjects' classifications were not influenced by the race of the "victim." This experiment was replicated by Sagar and Schofield (1980) using pictorial stimuli (instead of a videotape) with children as subjects. They found that black as well as white subjects tended to perceive black males as more threatening than white males.

Interpreting and labeling an action as an attack or a threat depends not only on contextual cues, but also on the meaning a person attaches to the label. The specific content of a general interpretive category such as "threat" can vary according to the context in which the label has been learned and used. Such variability can be cultural, or individual, or both. Rothwell (1971) finds considerable individual variability in the concept of crime among a small number of respondents in a single neighborhood. Respondents who were presented with a list of 12 actions identified between 3 and 12 of them as crimes. Some people interpreted crime as sin, some as harm toward others, and some as illegal behavior.

Variability in how respondents and interviewers define incidents of interpersonal violence has been adduced as one reason for biased measurements. A report by the National Research Council (Penick and Owens, 1976:135) noted:

> At the present time it is unclear how far the underreporting of certain incidents is due to memory failure or to other forms of response bias in the narrow sense of that term and how far it is due to differences between respondents in the definitions of particular types of incidents (for example a subcultural tendency in certain groups to regard physical violence as permissible or, at any rate, nondeviant rather than as an "assault").

Findings from an interviewer variance study "suggest that there may have been differences in the method by which NCS interviewers applied the concepts or definitions directly related to a determination of the incidence of assaultive violence without theft," and the investigators speculate that "Perhaps the variability among interviewers reflects the manner in which they view assaultive violence involving acquaintances or friends" (Bailey, Moore, and Bailar, 1978:21; see also Chapter 4, Section 4.2).

The purpose of our discussion of concepts related to crime is to suggest that even when survey questions refer to real, external events, defined in concrete and specific terms, the problem of subjectivity is by no means eliminated. This conclusion is not limited to measurements of victimization by crime, but applies to virtually all measurements of social behavior. Whether a person is classified as unemployed or out of the labor force, for example, depends crucially on whether she claims to have looked for work in the prior two weeks, which is again a matter of intention and interpreta-

tion. (See Schuman et al. Chapter 3, Part II "Quasi-Facts" in Volume 2 for other examples of seemingly objective measurements that prove, on closer inspection, to be partly subjective.)

When respondents are asked to report on purely subjective states, the problems of definition are even more severe. First, we do not know and often cannot assess the degree to which respondents use comparable definitions of the subjective states or events about which they are asked to report, nor do we know whether respondents' definitions correspond to what the investigator intended to measure. For some types of measurement, it is possible to assess the validity and comparability of respondent reports. For example, if a respondent claims knowledge of the French language, it would be possible to devise a test to validate the claim. But when respondents are asked to report on their happiness, it is difficult to know if they share a common concept of happiness among themselves and with the investigator.

Part of the problem may be that the English language is relatively impoverished of labels for subjective states of feeling, attention, need, and so on. William James (1890:195) said:

> This absence of a special vocabulary for subjective facts hinders the study of all but the very coarsest of them. The lack of a word . . . [leads us] . . . to suppose that no entity can be there; and so we come to overlook phenomena whose existence would be patent to us all, had we only grown up to hear it familiarly recognized in speech. It is hard to focus our attention on the nameless. . . .

In addition, we cannot be sure that survey questions intended to measure purely subjective states actually do so. We sometimes assume that when respondents are questioned about their attitudes or motives, they introspect about their feelings and thoughts and report the results of their self-examination to the interviewer. Some experimental psychologists suggest that people themselves may not know what they really feel, and that when asked questions about their subjective states they base their answers on external cues and on implicit ideas of what they ought to feel under the circumstances (see, for example, Nisbett and Wilson, 1977). By their account, we are led to the ironical conclusion that when one attempts to measure subjective states one inadvertently taps objective aspects of the situation. Not only are objective survey questions less objective than they first appear, but subjective questions may be less subjective than they seem as well. Not all psychologists agree with the conclusion that people cannot report accurately their own attitudes and motives. Ericsson and Simon (1980) suggest that accurate reports of some internal states can be obtained if questions focus upon a specific, real, current event, and do not require

respondents to reconstruct past feelings or to generalize about feelings or thoughts over a series of experiences. Importantly, they also suggest that people do not provide accurate reports about feelings to which they do not normally attend. Other results suggest that factors such as the interview setting and methods of questioning can be manipulated to focus respondents' attention upon themselves rather than the environment, resulting in more valid reports of subjective states (see Chapter 9, Section 9.4).

The question of whether a referent is defined in meaningful and comparable terms by respondents is not really resolved by a distinction between referents that are objective or subjective. People do not necessarily agree on the meaning or interpretation even of objective events. This implies that framing a question in concrete and behaviorally specific terms will not necessarily eliminate subjectivity. This issue of the meaningfulness and comparability of a referent perhaps resolves itself instead into a series of questions about respondents' state of knowledge about an event, and the source of that knowledge.

First, is a category or event known to respondents; do they attach meaning to it at all?

When the event in question has no familiar meaning, a respondent faces the task of constructing one. Respondents are adept at constructing a frame of reference even for remote topics, as suggested by the example of a survey question which, by design, referred to an extremely remote event. Respondents were asked: "Congress is considering the Agricultural Trade Act of 1978. Do you favor or oppose passage of this act?" One purpose of the question was to test respondents' willingness to give their opinions on something about which they presumably knew nothing (see Schuman and Presser, 1981). Although most respondents did not venture an opinion about the Agricultural Trade Act of 1978, those who did offer opinions "do not simply guess mindlessly, but are able to construct some meaning from the question . . . some people favor or oppose Congressional passage of the Agricultural Trade Act on the basis of their general confidence in government, regardless of the specific content of the Act, and . . . some people reach an opinion based on what the words 'Agricultural Trade' suggest to them" (Schuman and Presser, 1981:153-154). Schuman and Presser's findings suggest that respondents' attempts to construct meanings, even for distant and ambiguous topics, are by no means passive or mindless, but are systematically related to their prior attitudes and to information provided by the question itself. As the authors note, "the distinction between meaningful and meaningless responses to attitude survey questions is not as simple as is sometimes assumed" (1981:154). A respondent actively attempts to make sense of the question, and in doing so, uses whatever information is available. If the event in question is remote, vaguely specified, or meaningless, then a respondent (if he ventures an opinion at all) will attempt to

construct a definition using information derived from the question or its context, or from his knowledge about related subjects.

The example of the Agricultural Trade Act suggests the importance of a second set of questions. On what basis do respondents construct meaning for a category? What are their sources of knowledge about an event or category? Is the category or event used to organize the respondent's own experience, and for what purpose is it used?

As is true of opinions about the Agricultural Trade Act, many attitudes are hastily constructed on the basis of very vague and fragmentary information provided by the question itself. Converse (1970) has long argued that studies of attitudes should be grounded in careful assessments of how much respondents know about an attitude object. The meaningfulness and comparability of the referents of survey questions depend not only on how much respondents know about them, but also on the sources of that knowledge. The interpretive framework in terms of which an event (for example, crime) is understood may be derived from informal, direct experience; from the mass media; from formal schooling; formal codes and laws; and so on. The content of a category as well as the extent to which its meaning is shared will depend on sources of knowledge about it. Violence among friends and family members is perhaps a category that has been understood in terms of informal and heterogeneous frames of reference that vary from one cultural setting to another. We might expect this to change as more people come to depend upon mass media portrayals of crime as their source of knowledge about crime events and concepts. In contrast, voting in elections is widely understood in terms of a common framework that is national in scope. The meanings (and meaningfulness) of categories will vary according to how (and if) respondents use the category or event in question to organize their experience. For example, we expect that the (underlined) referent of the question, "During the last 12 months, did anyone *break into or somehow illegally get into* your (apartment/home), garage, or another building on your property?" might have a more precise and different meaning for policemen and lawyers than it does for most people. Police officers and lawyers have different sources of knowledge about that kind of event, a more precise frame of reference (the criminal code), and they use the category differently and more actively, as part of their professional work. (See Chapter 7, Section 7.1 for a discussion of sources of interpretive frameworks.)

Dimension 2. The Nature of the Judgment

We distinguish three major categories of judgments: existential beliefs, attributions, and valuations. These categories involve different intellectual

tasks, most of which have been the subject of psychological investigation and resulted in literature that is informative about the biases characterizing human judgment. Although the literature will not be reviewed here, we try to supply some useful general references which can provide information about the nature of the subjective phenomena we try to measure in surveys. (Each of the major categories is subdivided further in Table H.1, although not all of the subcategories will be discussed.)

I. EXISTENTIAL BELIEFS

A great many survey questions ask respondents for simple judgments about the possible or actual occurrence of events or states. Questions classified in this category primarily involve the mental tasks of recall or memory for past events and forming expectations or intentions about future ones.

There is of course a large literature on the psychology of memory, and survey researchers have also explored the problems of asking retrospective questions (see Chapter 9, Section 9.4). We note here that even incidents that one would intuitively expect to be memorable (for example, victimization by crime) are forgotten fairly rapidly, and that recall of past events is generally characterized by a strong recency bias. Evidence suggests that expectations and predictions about future events are biased by wishful thinking. Experimental studies (McGuire, 1960) indicate that subjects overestimate the probability of desired outcomes. (Although beliefs and attitudes are classified here as distinct mental judgments, they are fundamentally confounded: the literature on selective perception shows that feelings about a stimulus influence how a person perceives and describes it (see, for example, Erdelyi, 1974).

II. ATTRIBUTIONS

Survey questions that require respondents to attribute general characteristics and dispositions to social aggregates or actors are classified in this category. Questions in this category require the additional tasks of *generalization* and *inference*. The respondent must estimate the frequency, rate, or level of events (attribution/objective referent); attribute general traits and dispositions to social objects or actors (attribution/constructed referent); or describe his own traits (attribution/subjective referent). These questions are often more ambiguous and difficult than those in category I because the referents are typically general (for example, "most people," "Russian leaders") rather than specific. (They are often implicitly evaluative as well.) Since the referents of these questions are less constrained, respondents face the task of deciding who and what the question refers to. In addition, instead of being asked whether an event did or did not occur, a respondent must decide how often it occurred, whether it generally holds true, and so forth. Many questions in this category require even more complex infer-

ences and comparisons: respondents may be asked to compare levels or rates at different times or for different groups, or they may be asked to infer the causes or consequences of events. Questions about causal relationships often involve hypothetical situations: for example, "What government programs do you feel would suffer if federal taxes were reduced by 20 percent?" Such questions probably impose an additional burden on a respondent's inferential capacity.

We raise two separate questions here. First, can respondents even approximate the inferences that are called for by the questions? Common sense suggests that answers to such complex questions should be viewed skeptically, and this speculation is confirmed by psychological evidence. Second, if respondents cannot really answer a question as posed, how do they go about answering survey questions, and how should their responses be interpreted?

Psychological evidence suggests that when confronted with complex inferential tasks, people are likely to adopt some shortcut or heuristic to solve the problem. Tversky and Kahneman (1973) suggest that people estimate frequency or probability by the ease with which instances or associations can be brought to mind; for example, people may assess the divorce rate in a given community by recalling divorces among their acquaintances. (For discussions of this and other problem-solving heuristics, see Tversky and Kahneman [1974].) As noted, the ease of retrieval may be influenced by such factors as the recency and saliency of past occurrences as well as their actual frequency. In addition, evidence suggests that people draw the same conclusions from a small number of observations as from a large sample and that the salience of information, not its reliability, influences subjective inferences. Together, these findings imply that subjective inferences may be quite sensitive to the availability of relevant instances. Thus, for example, Combs and Slovic (1979) suggest that subjects overestimate the probability of well-publicized causes of death.

Certainly, it seems advisable not to take answers to complex inferential questions at face value, but it does not follow that responses to such questions are meaningless. Evidence and common sense suggest that the more complex the question, the more likely respondents are to simplify or reinterpret it so that they can answer it. If so, then survey researchers need to understand how respondents go about reinterpreting complex questions so that they can (a) provide reasonable interpretations of their answers, and (b) ask simpler questions that can be answered. It should be apparent that not all respondents will reinterpret complex questions in the same way and that their varying interpretations are a source of response variability that makes valid interpretations of the results difficult.

A second characteristic of attributive questions (and evaluative ones; see category III) is their extreme generality. Researchers frequently assume,

without verification, that respondents can meaningfully report their attitudes and actions abstracted from any situational context. Survey questions frequently ask respondents how they feel or act "in general," or "taking all things together," or "all in all," without knowing in what context (if any) the reports have meaning (for example, "In general, would you say you have become more favorable toward racial integration in recent years, or are you less favorable?"). Even if respondents are willing to answer such questions, their answers may not be meaningful if the attitudes and dispositions referred to are not coherent or general across situations. At least, survey questions ought to allow for the possibility that attributions and attitudes may be situation-specific. Bem and Allen (1974) provide some clues as to how one might go about this in the survey context. They asked subjects to rate their own attributes (for example, how "conscientious" and "friendly and outgoing" they were). These self-ratings were compared with ratings by peers and parents and by direct observation of behaviors that, in the view of the investigators, corresponded to the traits in question. Two results are relevant for survey measurements of general traits. First, subjects were asked how much they varied from one situation to the next in the rated attributes, and this measure proved very predictive of consistency between rated and observed traits. Or, to put it another way, subjects' self-descriptions were more valid predictors of their behavior when they rated themselves as consistent. Second, the investigators' concepts of the trait in question did not always correspond to subjects' concepts of the same trait. (For example, "neatness" was regarded by the investigators as one manifestation of "conscientiousness," but was not so defined by subjects or raters.) These results suggest that survey investigators can obtain better predictive measures by including direct questions about consistency across situations. In addition, investigators should not assume that their own definitions of attributes or traits correspond to those used by respondents, or that respondents have coherent and consistent trait concepts. The problem of whether two questions measure the same or different conceptual dimensions can be addressed using techniques discussed in Chapter 6. It seems advisable to include more survey questions of the type labeled "social definitions" in category II (for example, "What is the first thing that comes to your mind when you think of someone who is a liberal or conservative?"). More careful exploration of respondents' interpretations of the attributes and concepts asked about in survey questions could help investigators interpret the meaning of responses to fixed-format survey questions (see also Chapter 7, Section 7.3).

III. VALUATION

Questions in category III require respondents to orient themselves toward some social object or condition in terms of its "goodness" or "badness,"

"importance" or "nonimportance." (There are many different formats for evaluative tasks; psychometric literature has much information on the relationship of form of task—for example, rating or ranking—on reliability.)

We note that dimension 1 (objective vs. constructed vs. subjective events) assumes a somewhat different meaning for evaluative questions. We classify as objective those questions that focus on external events and ask a respondent to take a stand for or against, in favor or opposed. Evaluative questions that are objective in this sense (for example, "Do you think the use of marijuana should be made legal, or not?") typically have the quality of final judgments, preferences, or decisions; more than any others, questions in this category include what might be defined as "public opinion." For the most part, such questions have implications for public or private action: they ask respondents to approve or reject policies, laws, rights, and practices.

Other evaluative questions focus more on respondents' feelings and sentiments about public and personal matters. The respondent is asked to look inward and evaluate his own state of mind with respect to some object—his personal tastes, feelings of satisfaction or dissatisfaction with personal or public affairs, or wishes and concerns. Typically, questions in this category have no clearcut implications for collective (or other) action.

We distinguish five types of evaluation:

1. *Ideals and preferences* are what the respondent would like to have or see happen. These include questions about wishes, preferences, and values.
2. *Approval ratings* include explicit normative judgments of practices, roles, or rules as right or wrong, which should or should not occur or of which the respondent approves or disapproves. There is a clear moral tone to these questions.
3. *Attitudes* towards groups, public figures and organizations, and policies. Attitude questions classified as objective measure respondents' support or opposition. Attitude questions classified as subjective measure sentiments of trust, confidence, and so on. Attitude questions are not cast in moral terms, although moral judgments may or may not lie behind respondents' evaluations.
4. *Satisfaction* and "quality of life" evaluations.
5. *Problems and worries*.

The basis for a respondent's evaluation is an important determinant of his response. Thus, for example, it makes a difference if the respondent is asked to voice support or opposition for a law prohibiting abortion or is asked his or her private sentiments about abortion as a practice. Investigators should not necessarily expect these different attitude measurements to

be consistent. Moreover, sometimes it is not clear what sort of evaluation is required by a question. The "happiness" question is perhaps a case in point. One might reasonably interpret it as a question about self-concept—do you think of yourself as a happy person?—or as a question about one's mood or feeling over some undefined period of time. (For this reason the question is classified in two categories in Table H.1.) The type of evaluation elicited by a survey question is probably very sensitive to the context and phrasing of the question. There may be particularly subtle distinctions in the realm of moral evaluation. For example, subtle changes in wording may change a question from one which asks whether a practice is good or acceptable into a question which asks whether a practice represents a right or a freedom. A possible example is the effect of varying the words "forbid" versus "allow" in the question "Do you think the United States should allow (forbid) public speeches against democracy?" (see Chapter 5, Section 5.1). A question that asks whether a practice that is not legally restricted "should be allowed" may be interpreted as asking whether the practice should be acknowledged or recognized as acceptable. (The question has a different meaning if it asks about practices that are prohibited by law: then to say "It should be allowed" implies that the legal restriction should be removed.) A question asking whether the same practice "should be forbidden," on the other hand, asks whether people should have the legal right to do it. There is a subtle distinction between what is acceptable and desirable, what is undesirable but should not be forbidden by law, and what should be forbidden by law. It may be argued that the "allow" version of the speeches-against-democracy question is ambiguous with respect to the middle category. This illustration suggests the possibility that we may alter an evaluative response by inducing a change in the basis for the respondent's evaluation. (See Chapter 6 for ways of assessing whether alternative wordings of a question measure the same or different attitudes.)

Dimension 3. The Object of the Report: Self Versus Other

Survey questions often ask the respondent to take herself as an object to report on, describe, or evaluate. (Self-report questions are marked with an asterisk in Table H.1.) Social psychological evidence suggests that there are systematic differences in attributions of qualities or motives to oneself and to others. In part these differences appear motivated by the desire to think of or present oneself in a favorable light. For example, experimental findings suggest that subjects are prone to attribute success, as in a basketball game, to their own ability and effort and to attribute failure to external

factors, such as luck. Although the self-serving tendency to take responsibility for good outcomes and to deny responsibility for bad outcomes appears general (see Bradley, 1978), there are circumstances in which it does not occur. If subjects believe that their performance will be evaluated independently, for example, their self-appraisals are more modest, perhaps because an obviously inflated opinion of oneself can be a source of embarrassment.

Similarly, respondents in surveys are likely to answer questions so as to present themselves favorably. (See the review of the literature on social desirability by DeMaio, Chapter 9, Volume 2.) The experimental literature suggests that such tendencies are likely to be less if a respondent believes that an exaggerated claim can be falsified by other evidence.

A second difference between attributions to oneself and to others appears due to the different perspectives of actor and observer. Psychological evidence suggests that people tend to see their own actions as situationally caused and the behavior of others as resulting from stable personality tendencies. This conclusion follows both from experimental evidence (see, for example, Jones and Nisbett, 1972) and from paper-and-pencil tests in which subjects are asked to describe the traits of themselves and others. Using virtually all trait-descriptive adjectives in the English language, Goldberg (1978) finds that the response "depends on the situation" is used most often when describing oneself, less often when describing a neutral other, and least often when describing either a well-liked or disliked other. The effect is small, but remarkably consistent: more than 92 percent of the 2,800 trait words showed the general effect. (He notes, however, that the situational response is also the "middle, neutral, average, uncertain, and ambiguous response" [p. 1028] so his findings may be consistent with other interpretations.)

Finally, we note that attributions to self and to others are linked in a fundamental way. The behavior and attitudes of other people are not evaluated independently, but are perceived and evaluated in the context of one's own behavior and attitudes: what is congruent is seen as normal, and what is discrepant is viewed as deviant and in need of explanation (see, for example, Ross, 1977). Experimental evidence suggests that there is a general tendency to project one's own traits and behaviors to others. (See Holmes [1968] and Ross [1977] for reviews of evidence and a discussion of the bases for the phenomenon.) For example, Bender and Hastorf (1950) found that subjects' estimates of the attitudes of their best friends were highly correlated with their own attitudes ($r = .71$), but not at all with their friends' actual attitudes, independently measured ($r = .07$; see also Fields and Schuman, 1976). Clearly, egocentric attribution may be the basis of a false consensus, with each person viewing his own attitudes and actions as normal and appropriate.

Holmes (1968) finds evidence of a second type of projection, labeled "complementary projection." People tend to attribute traits and attitudes to others that complement and justify their own feelings. For example, when subjects are made fearful, they are more likely to see others as threatening or aggressive (Holmes, 1968:250). Complementary projection may underlie some anomalous survey results, such as the finding reported by Schanck (1932; also see Chapter 8) that churchgoers falsely perceived a consensus opposed to cardplaying. The false attribution may have served to justify their own fear of public censure; respondents' fears of expressing their more liberal private views may have led them to exaggerate the conservatism and intensity of their neighbors' feelings. Ironically, this process of complementary projection would make them even more inhibited about expressing what they presumed were contrary views, thus reinforcing the false consensus. It is also possible that Schanck's respondents were truly opposed to cardplaying, but these sentiments were either concealed from interviewers or denied at the unconscious level and defensively attributed to others. Concealment is possible (although Schanck discounts it), but defensive projection appears unlikely in view of evidence reviewed by Holmes (1968). Schanck finds no evidence suggesting that people project a trait onto others as a defense against recognizing it in themselves. Finally, Schanck's findings also may be explained in terms of the common tendency to rate oneself as a little smarter, more enlightened, and above average (see Myers [1980] for evidence on this point).

The fact that other people are perceived in the context of one's own behavior and attitudes has very important implications for survey attempts to measure perceptions of the social world. Obviously, attributions to others should not be taken at face value, as if they were arrived at independently and in an unbiased fashion. Clearly, a respondent's own attitudes and beliefs color (or even dominate) perceptions of others, and must be taken into account when analyzing such data.

Notes

1. The classification scheme in Figure H.1 and Table H.1 was developed inductively by analyzing and grouping a large and haphazard batch of survey questions from a variety of sources (including AIPO, Harris, SRC, NCS, CPS, NORC's-General Social Survey, and the Detroit Area Study). Questions were classified on an ad hoc and intuitive basis. We might however test more rigorously whether two different questions measure the same dimension, using the techniques explicated in Chapter 6 in this Volume and in Duncan, Chapter 12 in Volume 2.

2. The definition of what constitutes an event is itself far more a matter of interpretation than is implied by our discussion here.

TABLE H.1

Classification of Survey Questions According to Three Dimensions of Subjectivity

	Dimension 1. Referent of the Question		
Dimension 2. Nature of the Judgment	**Objective Events[a]**	**Constructed Events[b]**	**Subjective Events**
I. Existential beliefs about the occurrence or nonoccurrence of events or states.			
A. Past events	[During the last 6 months] Did anyone *threaten* to beat you up or *threaten* you with a knife, gun, or some other weapon, NOT including telephone threats? *Have you ever attended grand opera?	*What have you been doing in the last 4 weeks to find work? *Did you give up or do anything differently for Lent?	*Awareness or Knowledge:* Was the person someone you knew or a stranger you had never seen before? Have you ever heard or read about flying saucers?
B. Present events	*Do you smoke? Do you have any children under 6 years of age?	*Is there any area right around here—that is, within a mile—where you would be afraid to walk at night?	Can you tell me the name of your congressman? *Can you swim?
C. Expectations about future events	*Forecasts:* If you had to guess, who do you think will win for President this November—Nixon or McGovern? Do you expect the United States to fight in another world war within the next ten years?	Do you feel the negotiations in Paris will end in an honorable settlement or not?	*Intentions and Plans:* *Are you planning to go to work? *Do you expect to have any (more) children? *Do you, yourself, plan to vote in the election for President in November 1968?

D. Hypothetical events	Do you think Richard Nixon could return to a role in public life?	Do you feel that some kind of human life exists on other planets in the total universe? Do you believe in life after death?	*Hypothetical Intentions:* *If you discovered that a son or daughter of yours had a supply of marijuana in his or her room at home, would you turn him or her into the legal authorities or not?
E. Likelihood and certainty	If he were President, who do you think would be more likely to end U.S. involvement in Vietnam sooner—Nixon or McGovern?		*Certitude:* *Many people, including saints and great religious leaders, have had doubts from time to time throughout their lives about the existence of God. Which of these statements about your belief in God's existence best fits the way you feel? Not sure; . . . Not too sure; . . . Rather sure; . . . Very sure.
II. Attribution			
A. Categorizing people	*Social Classification:* Thinking in terms of *your* closest friends, what proportion are Jewish?	*Stereotypes and Social Definitions:* Which words seem to you to describe the German people best? What is the first thing that comes to mind when you think of someone who is a liberal or conservative?	*Self-Identification:* *In politics as of today, do you consider yourself a Republican, Democrat, or Independent?

NOTE: *Asterisks denote questions that ask respondent to take himself or herself as the object of report—that is, distinguishing **Dimension 3.**
[a]Objective events are events which are in principle observable by people other than the respondent.
[b]Constructed events are external events defined in quasi-evaluative terms, or by R in subjective terms.

TABLE H.1 *(continued)*

Dimension 2. Nature of the Judgment	Dimension 1. Referent of the Question: Objective Events[a]	Constructed Events[b]	Subjective Events
(II. Attribution cont.)			
B. Generalities and estimates (frequency, rate, or level)	*Estimates of Quantifiable Events:* *At what length do you wear your hair? *We would like to know how you and your (husband/wife) divide up some of the family jobs . . . . Now who does the grocery shopping? Husband always, husband more than wife; husband and wife exactly the same; . . . Do you feel that a great many young couples have premarital sex before getting married these days, some but not a lot, or almost none have premarital sex?	*Attributions of General Traits and Dispositions:* Do you agree with this statement? —The Russian leaders are basically trying to get along with America. How much of the population of S. Vietnam do you feel is under communist influence—none, under 10%, 10–20%, . . . ? Do you feel that wiretapping for political purposes is a common occurrence, or not?	*Self-Concept and Self-Image:* *Some people feel they can run their own lives pretty much the way they want to, others feel the problems of life are too big for them. Which one are you most like? *Taken all together, how would you say things are these days—would you say that you are very happy, pretty happy, or not too happy? *How much thought have you given to the coming November elections for Congress—quite a lot, some, or only a little?
C. Comparisons and trends	Do you think the country has had more strikes and labor disputes recently than a few years ago, fewer strikes, or about the same number? Do you feel wages paid school teachers are average, above average, or below average?	Do you feel that patriotism has declined in the past few years in this country, increased, or stayed about the same? Tell me if you agree or disagree with this statement: Women will always be more emotional and less logical than men.	*In general, would you say you have become more favorable toward racial integration in recent years, or are you less favorable?

	*Compared with American families in general, would you say your family income is far below average, below average, average, . . . ?	Which city has the gayest night life?	
Causal attributions:			
D. Causes	Do you feel the Kennedy assassination was the work of one man or was it part of a broader plot? What is the chief reason for auto accidents in your opinion?	If you had to choose, do you feel that most student unrest is the result more of deeply felt injustices in society, or more the work of radical troublemakers? What do you think usually makes a person take to heavy drinking?	*Personal Reasons, Motives, and Causes:* *Women have different reasons for working. What would be your main reason for working? *How important is a candidate's position on penalties for marijuana use when you decide how to vote in an election—is it one of the most important factors you would consider, a very important factor, somewhat important, or not too important?
E. Consequences	Do you think birth control pills can be used safely—that is, without danger to a person's health?	Do you agree or disagree that union membership discourages productivity?	*Has the television coverage of this war made you feel more opposed to the war or not? *Could you see yourself being happy in a marriage without sex?
F. Hypothetical consequences	What government programs do you feel would suffer if federal taxes were reduced by 20 percent?		*Do you think you would be happier if you had gone farther in school?

TABLE H.1 *(continued)*

Dimension 2. Nature of the Judgment	Dimension 1. Referent of the Question: Objective Events[a]	Constructed Events[b]	Subjective Events
III. Valuation and Affect			
A. Ideals, Tastes, Preferences	*Ideals and Preferences:* What do you think is the ideal number of children for a family to have?	If you could look into tomorrow's paper and see any headline, what would you most like to see?	*Wishes, Interests, Tastes:* *Do you wish you had a part-time or full-time job outside the home, or not?
	Do you agree or disagree?—Next to health, money is the most important thing in life.	*Do you have a favorite author?	*Would you like to go along on the first rocket trip to the moon?
	I want to read you a list of things describing life here in America. For each item please tell me how important it is to you personally—very important, somewhat important, only slightly important, or not important at all—living in a democratic society; having a high standard of living; . . .		*How interested would you be in making a trip to the following countries?
	*Would you please look at this card and tell me which one thing on this list you would most prefer in a job?		

	Moral Judgments:	*Moral Constructs:*	*Pride, Guilt, Moral Sentiments:*
B. Approval Ratings 1. Approved identities and roles	Now let me ask you which, in your judgment, is worse, if you could only choose *ONE*—an adulterer, OR a politician who takes bribes?	Are there some kinds of work you feel women should not have?	*Would you say you take pride in being an American or not?
2. Approval of individual behavior and practices	If a man and a women have sex relations before marriage, do you think it is always wrong, almost always wrong, wrong only sometimes, or not wrong at all? Do you feel it is justified or not justified to take the life of another person in self-defense?	*What was the most difficult moral decision you have had to make in your life? Are there any situations that you can imagine in which you would approve of a man punching an adult male stranger?	*Have you often, sometimes, or never felt bad because you didn't go to church enough? Would you respect a man more, less, or about the same if he decided to stay at home and take care of the children while his wife worked?
3. Approval of institutional policies, practices, actions	Do you think the present draft system is fair, or not? When college students occupy a building or an office, do you think the authorities are right or wrong to call in the local police or the federal guard? Do you think the United States should allow public speeches against democracy?	Now, on the issue of war, do you believe in war, or not?	Do you find yourself getting annoyed a lot, getting somewhat annoyed, or hardly getting annoyed at all by the Miss America Beauty Pageant and how it patronizes women?

TABLE H.1 *(continued)*

Dimension 2. Nature of the Judgment	Dimension 1. Referent of the Question: Objective Events[a]	Constructed Events[b]	Subjective Events
(III. Valuation and Affect cont.)			
4. Approval (acknowledgement) of rights and freedoms	Do you think some of the books you have seen or heard about should be censored by government or do you think people have a right to them if they want them? Do you think a 12-year-old child should be allowed to decide for himself whether he will go to church or Sunday school, or should his parents decide this? How do you feel about the right to have a union—do you favor or oppose workers having that right?		*Do you feel you have the right to question what your church teaches, or not? *Do you feel threatened in any way by having information about yourself in some files, or don't you feel threatened by that?
5. Approval of formal rules or laws	Do you think there should be laws against marriages between (Negroes/blacks) and whites? Do you think the use of marijuana should be made legal, or not? Do you think the United States should forbid public speeches against democracy?		

6. Sanctions	Are you in favor of the death penalty for persons convicted of murder? Do you favor or oppose taking away a student's scholarship or loan if he participates in an unlawful protest?		
C. Attitudes 1. Groups	*Support and Opposition:* Do you feel Negroes have tried to move too fast, too slow, or about the right pace?	*Positive and Negative Stereotypes:* Tell me if you agree or disagree with this statement: "Most men are better suited emotionally for politics than women." Do you agree or disagree?—Too many people on welfare cheat by getting money they're not entitled to.	*Sentiments (trust, confidence, positive and negative feelings):* Do you personally trust most white people? How would it make you feel if a close relative of yours were planning to marry a [(Negro/black)/white]? Would you be very uneasy, somewhat uneasy, or not uneasy at all.
2. Public figures and organizations	How would you rate the job President Nixon is doing as president—excellent, pretty good, only fair, or poor? Are you in favor of labor unions?	What 2 or 3 things do you especially like about President Nixon personally?	For each of the following countries, would you say you generally have good feelings toward the country or not?—Russia As far as the people in charge of running the military are concerned, would you say you have a great deal of confidence, only some confidence, or hardly any confidence at all in them?

TABLE H.1 *(continued)*

	Dimension 1. Referent of the Question		
Dimension 2. Nature of the Judgment	**Objective Events[a]**	**Constructed Events[b]**	**Subjective Events**
(III. Valuation and Affect cont.)			
3. Policies	In view of the developments since we entered the fighting, do you think the United States made a mistake in sending troops to fight in Vietnam? Do you favor or oppose the U.S. selling wheat and other grains to the Soviet Union? Do you favor or oppose busing children to schools outside their neighborhood to achieve racial desegregation?	In general, do you think the courts in this area deal too harshly, or not harshly enough with criminals?	
D. Satisfaction and "Quality of Life" Ratings	*Judgments:* *Do you feel you work too many hours a week, too few, or about the right number of hours? Do you feel that the taxes you pay are too high, too low, or are reasonable?	*What do you personally find to be the two or three most enjoyable things about being a woman today? In general, over the past 10 years, do you think that America has become a better place to live, a worse place to live, or is it just about the way it was 10 years ago?	*Sentiments (satisfaction; positive or negative feelings):* *Do you sometimes feel down, lonely, or depressed, or don't you ever feel that way? *Taken all together, how would you say things are these days—would you say you are very happy, pretty happy, or not too happy?

	Compared to 10 years ago, do you feel the quality of most things you buy has become better, worse, or not changed much?	Do you agree or disagree?—Children born today have a wonderful future to look forward to.	Are you well satisfied, more or less satisfied, or not at all satisfied with city garbage collection here? *Do you feel a sense of accomplishment in your job, or not?
E. Problems and Worries	All in all, do you feel the student protests and demonstrations are a very serious national problem, somewhat serious, or hardly serious at all?	In your opinion, what is the most important problem facing the country today?	*Worries and Concerns:* How do you feel about the war personally? Do you feel grim about it, determined about it, sad about it, nervous about it—or how? As you know, the value of the dollar has gone down sharply in recent months. Does this decline in the worth of the dollar abroad worry you a great deal, some but not a lot, or not at all?

References

Abelson, R. P. (1954) A technique and a model for multi-dimensional attitude scaling. *Public Opinion Quarterly* 18:405-418.

Abelson, R. P. (1972) Are attitudes necessary? In B. T. King and E. McGinnies, eds., *Attitudes, Conflict, and Social Change.* New York: Academic Press.

Abelson, R. P. (1979) Social clusters and opinion clusters. In P. Holland and S. Leinhardt, eds., *Perspectives on Social Network Research.* New York: Academic Press.

Abelson, R. P. (1982) Three modes of attitude-behavior consistency. In M. P. Zanna, E. T. Higgins, and C. P. Herman, eds., *Consistency in Social Behavior: The Ontario Symposium,* Vol. 2. Hillsdale, N.J.: Erlbaum.

Abelson, R. P., Aronson, E., McGuire, W. J., Newcomb, T. M., Rosenberg, M. J., and Tannenbaum, P. H., eds. (1968) *Theories of Cognitive Consistency.* Chicago: Rand McNally.

Abrams, M. (1973) Subjective social indicators. In U.K. Central Statistical Office, *Social Trends: 1973.* London: Her Majesty's Stationery Office.

Abrams, M. (1980) Social research and market research. *Sociology* 14:113-115.

Acland, H. (1979) Are randomized experiments the Cadillacs of design? *Policy Analysis* 5:223-241.

Adams, E. W. (1965) Elements of a theory of inexact measurement. *Philosophy of Science* 32:205-228.

Adams, E. W. (1966) On the nature and purpose of measurement. *Synthese* 16: 125-169.

Adler, K., and Wertman, D. (1981) Is NATO in trouble? A survey of European attitudes. *Public Opinion* 4(Aug.-Sept.).

Adorno, T., Frenkel-Brunswik, E., Levinson, D., and Sanford, N. (1950) *The Authoritarian Personality.* New York: Harper and Brothers.

Advertising Age (1975) Senate unit investigates consumer agency poll. April 7.

Ajzen, I., and Fishbein, M. (1977) Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin* 84:888-918.

Alexander, H. (1976) *Financing the 1972 Election.* Lexington, Mass.: Lexington Books.

Alwin, D., and Stephens, S. (1979) Use of sample surveys in the social sciences. Paper presented at the 145th national meeting of the American Association for the Advancement of Science, Houston, Texas, January 3-8, 1979.

American Anthropological Association (1971) *Annual Report 1970 and Directory.* Washington D.C.: American Anthropological Association.

American Association for Public Opinion Research (1976) Report of the Ad Hoc Evaluation Committee appointed to investigate the survey conducted by Opinion Research Corporation on the issue of consumer advocacy. January 16. New York: American Association for Public Opinion Research.

American Association for Public Opinion Research (1979-80) Code of professional ethics and practices. In *AAPOR Directory of Members.* New York: American Association for Public Opinion Research.

American Association for Public Opinion Research (1980) Final Report of the Standards Committee, 1979-1980, by D. Hughes. *AAPOR Newsletter* July:13.

American Bar Association (Project on Standards for Criminal Justice) (1968) *Standards Relating to Fair Trial and Free Press.* Chicago: American Bar Association.

American Statistical Association (1974) Report of the ASA conference on surveys of human populations. *American Statistician* 28:30-34.

Anastasi, A. (1976) *Psychological Testing.* 4th ed. New York: Macmillan.

Andersen, R., Kasper, J., Frankel, M. R., and Associates (1979) *Total Survey Error.* San Francisco: Jossey-Bass.

Anderson, N. H. (1970) Functional measurement and psychophysical judgment. *Psychological Review* 77:153-170.

Anderson, T. W. (1954) Probability models for analyzing time changes in attitudes. In P. F. Lazarsfeld, ed., *Mathematical Thinking in the Social Sciences.* Glencoe, Ill.: Free Press.

Andrews, F., and Withey, S. (1976) *Social Indicators of Well-Being.* New York: Plenum Press.

Aronson, E., and Mills, J. (1959) The effect of severity of initiation on liking for a group. *Journal of Abnormal and Social Psychology* 59:177-181.

Ashby, W. R. (1960) The relativity of "meaning." *Nature* 187:532.

Babbie, E. (1973) *Survey Research Methods.* Belmont, Calif.: Wadsworth Publishing.

Back, K. W. and Cross, T. S. (1982) Response effects of role restricted respondent characteristics. In W. Dijkstra and J. van der Zouwen, eds., *Response Behaviour in the Survey-interview.* London: Academic Press.

References

Baddeley, A. D. (1979) The limitations of human memory: Implications for the design of retrospective surveys. In L. Moss and H. Goldstein (eds.), *The Recall Method in Social Surveys*. London: University of London Institute of Education.

Baddeley, A. D. and Hitch, G. J. (1977) Recency reexamined. In S. Dornic, ed., *Attention and Performance VI*. Hillsdale, N.J.: Erlbaum.

Baddeley, A. D., Lewis, V. and Nimmo-Smith, M. I. (1978) When did you last . . . ? In M. M. Gruneberg, P. E. Morris, and R. N. Sykes, eds., *Practical Aspects of Memory*. New York: Academic Press.

Bailar, B. (1975) The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association* 70:23-30.

Bailar, B. (1979) Evaluation of sample survey data. Paper presented at the 145th national meeting of the American Association for the Advancement of Science, Houston, Texas, January 3-8, 1979.

Bailar, B., Bailey, L., and Stevens, J. (1977) Measures of interviewer bias and variance. *Journal of Marketing Research* 14:337-343.

Bailar, B., and Lanphier, M. (1978) *Development of Survey Methods to Assess Survey Practices*. Washington, D.C.: American Statistical Association.

Bailey, L., Moore, T., and Bailar, B. (1978) An interviewer variance study for the eight impact cities of the National Crime Survey cities sample. *Journal of the American Statistical Association* 73:16-23.

Barath, A., and Cannell, C. F. (1976) Effect of interviewer's voice intonation. *Public Opinion Quarterly* 40:370-373.

Baxter, R. (1964) An inquiry into the misuse of the survey technique by sales solicitors. *Public Opinion Quarterly* 28:124-134.

Belson, W. A. (1968) Respondent understanding of survey questions. *Polls* 3:1-13.

Bem, D., and Allen, A. (1974) On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review* 81:506-520.

Bender, I., and Hastorf, A. (1950) The perceptions of persons: Forecasting another person's responses on three personality scales. *Journal of Abnormal and Social Psychology* 45:556-561.

Beniger, J. R. (1976) Winning the presidential nomination: National polls and state primary elections, 1936-1972. *Public Opinion Quarterly* 40:22-38.

Beniger, J. R., Watkins, S., and Ruz, J. E. (1978) Trends in the abortion issue as measured by events, media coverage and public opinion indicators. *Proceedings of the American Statistical Association (Social Statistics Section)* 1978:118-123.

Benny, M., Riesman, D., and Star, S. (1956) Age and sex in the interview. *American Journal of Sociology* 62:143-152.

Berk, R., Hennessy, M., and Swan, J. (1977) The vagaries and vulgarities of "scientific" jury selection: A methodological evaluation. *Evaluation Quarterly* 1:143 158.

Biderman, A. D. (1975) The survey method as an institution and the survey institution as a method. In H. W. Sinaiko and L. A. Broedling, eds., *Perspectives on Attitude Assessment*. Washington D.C.: Smithsonian.

Biderman, A. D. and Lynch, J. P. (1981) Recency bias in data on self-reported victimization. *Proceedings of the American Statistical Association (Social Statistics Section)* 1981: 31-40.

Biderman, A. D. and Moore, J. C. (1980) Report of the workshop on applying cognitive psychology to recall problems of the National Crime Survey. Unpublished ms., Washington D.C.: Bureau of Social Science Research.

Biderman, A. D., and Sharp, L. M. (1973) *The Competitive Evaluation Research Industry*. Washington, D.C.: Bureau of Social Science Research.

Bishop, G. F., Tuchfarber, A. J., and Oldendick, R. W. (1978) Change in the structure of American political attitudes: The nagging question of question wording. *American Journal of Political Science* 22:250-269.

Bishop, G. F., Tuchfarber, A. J., Oldendick, R. W., and Bennett, S. E. (1979) Questions about question wording: A rejoinder to revisiting mass belief systems revisited. *American Journal of Political Science* 23:187-192.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975) *Discrete Multivariate Analysis*. Cambridge, Mass.: MIT Press.

Black, M., and Metzger, D. (1965) Ethnographic description and the study of law. *American Anthropologist* 67(6:part 2, special publication):141-165.

Blair, E. (1977-78) More on the effects of interviewer's voice intonation. *Public Opinion Quarterly* 41:544-548.

Blumer, H. (1948) Public opinion and public opinion polling. *American Sociological Review* 13:542-552.

Blumler, J. (1977) The intervention of television in British Politics. In Appendix E of Committee on the Future of Broadcasting (chair Lord Annan), *Report of the Committee on the Future of Broadcasting*. (CMND Nos. 6753, 6753-I). London: Her Majesty's Stationery Office.

Boffey, P. M. (1975) *The Brain Bank of America*. New York: McGraw.

Bogardus, E. S. (1968) Comparing racial distance in Ethiopia, South Africa, and the United States. *Sociology and Social Research* 52:149-156.

Bolinger, D. (1974) On the passive in English. In A. Makkai and V. B. Makkai, eds., *The First Lacus Forum*. Columbia, S.C.: Hornbeam Press.

Bolinger, D. (1977) *Meaning and Form*. London: Longman.

Bonafede, D. (1981) As pollster to the president, Wirthlin is where the action is. *National Journal* December 12:2184-2187.

Bonora, B., and Krauss, E., eds. (1979/1983) *Jurywork: Systematic Techniques*. Berkeley, Calif.: National Jury Project. (2nd ed., New York: Boardman.)

Boseman, A. T. (1980) ISR Operations for Fiscal Year 1979-1980. Institute for Social Research, University of Michigan, Ann Arbor. October memorandum.

References

Bowen, J. (1979) A survey of the general public's attitudes to market research. *Journal of the Market Research Society* 21:75-102.

Bower, G. (1981) Mood and memory. *American Psychologist* 36:129-148.

Bradburn, N. M. (1969) *The Structure of Psychological Well-Being*. Chicago: Aldine.

Bradburn, N. M. (1980) Verified Statement of Dr. Norman M. Bradburn, entered in United States District Court, Southern District of New York, *Carey et al.* vs. *Klutznick et al.*, 80 Civ 4550, Nov. 10.

Bradburn, N. M., and Miles, C. (1979) Vague quantifiers. *Public Opinion Quarterly* 43:92-101.

Bradburn, N. M., Sudman, S., and Associates (1979) *Improving Interview Method and Questionnaire Design*. San Francisco: Jossey-Bass.

Bradley, G. W. (1978) Self-serving biases in the attribution process: A reexamination of the fact or fiction question. *Journal of Personality and Social Psychology* 36:56-71.

Brody, R. A., and Page, B. I. (1975) The impact of events on presidential popularity: The Johnson and Nixon administrations. In A. Wildavsky, ed., *Perspectives on the Presidency*. Boston: Little Brown.

Brooks, C. A., and Bailar, B. (1978) *An Error Profile: Employment as Measured by the Current Population Survey*. Statistical Policy Working Paper 3. Washington, D.C.: U.S. Department of Commerce.

Brown, M. P., Cardozo, R. N., Cunningham, S. M., Salmon, W. J., and Sultan, R. G. M. (1968) *Problems in Marketing*. 4th ed. New York: McGraw-Hill.

Bryce, J. (1899) *The American Commonwealth*. 2 vols. New York: Macmillan.

Bryson, M. C. (1976) The Literary Digest poll: Making of a statistical myth. *American Statistician* 30:184-185.

Bureau of the Census (1968) 1960 Census of Population and Housing. Evaluation and Research Program. Series ER60, No. 7. *Effects of Interviewers and Crew Leaders*. Washington, D.C.: U.S. Department of Commerce.

Bureau of the Census (1970) Victim recall pretest (Washington, D.C.) Household survey of victims of crime. June 10 memorandum no. 70-237.

Bureau of the Census (1973) 1970 Census of Population and Housing. Evaluation and Research Program. PHC(E)-4. *Estimates of Coverage of Population by Sex, Race, and Age: Demographic Analysis*, by J. S. Siegel. Washington, D.C.: U.S. Government Printing Office.

Bureau of the Census (1974) 1970 Census of Population and Housing. Evaluation and Research Program. PHC(E)-9. *Accuracy of Data for Selected Population Characteristics as Measured by Reinterviews*. Washington, D.C.: U.S. Government Printing Office.

Bureau of the Census (1975) *Current Population Reports.* Series P-20, No. 277. Fertility expectations of American women: June, 1974. Washington, D.C.: U.S. Department of Commerce.

Bureau of the Census (1977) 1970 Census of Population and Housing. Evaluation and Research Program. PHC(E)-12. *The Employer Record Check.* Washington, D.C.: U.S. Government Printing Office.

Bureau of the Census (1979) 1970 Census of Population and Housing. Evaluation and Research Program. PHC(E)-13. *Enumerator Variance in the 1970 Census.* Washington, D.C.: U.S. Government Printing Office.

Bureau of the Census (1980a) *Social Indicators III.* Washington D.C.: U.S. Government Printing Office.

Bureau of the Census (1980b) *Current Population Survey: Interviewers Reference Manual.* CPS-250. Washington D.C.: U.S. Department of Commerce.

Bureau of the Census (1980c) Project descriptions. Demographic Surveys Division, June memorandum.

Bureau of the Census (1980d) Statement prepared by the staff of the Center for Social Science Research for the director's news conference, October 31.

Bureau of the Census (1981) *Current Population Survey: Guide for Training New Interviewers.* CPS-270. Washington, D.C.: U.S. Department of Commerce.

Bushery, J. M., Cowan, C. D., and Murphy, L. R. (1978) Experiments in telephone-personal visit surveys. *Proceedings of the American Statistical Association (Survey Research Methods Section),* 1978:564-567.

Caddell, P. (1979) Crisis of confidence I: Trapped in a downward spiral. *Public Opinion* 2(Oct./Nov.).

Campbell, A. A. (1945) Two problems in the use of the open question. *Journal of Abnormal and Social Psychology* 40:340-343.

Campbell, A. (1976) Subjective measures of well-being. *American Psychologist* 31:117-124.

Campbell, A., and Converse, P. E., eds. (1972) *The Human Meaning of Social Change.* New York: Russell Sage.

Campbell, A., Converse, P. E., Miller, W. E., and Stokes, D. E. (1960) *The American Voter.* New York: Wiley.

Campbell, A., Converse, P. E., and Rodgers, W. L. (1976) *The Quality of American Life.* New York: Russell Sage.

Campbell, A., Heeringa, S., and Miller, A. (1979) Full text of an evaluation of the Minnesota Poll. *Minneapolis Tribune,* February 25.

Campbell, B. A. (1981) Race-of-interviewer effects among Southern adolescents. *Public Opinion Quarterly* 45:231-44.

Campbell, D. T. (1957) Factors relevant to the validity of experiments in social settings. *Psychological Bulletin* 54:297-312.

References

Campbell, D. T. (1963) Social attitudes and other acquired behavioral dispositions. In S. Koch, ed., *Psychology: A Study of a Science*, Vol. 6. New York: McGraw-Hill.

Campbell, D. T., and Fiske, D. W. (1959) Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* 56:81-105.

Campbell, N. (1920) *Physics: The Elements*. Cambridge, England: Cambridge University Press.

Cannell, C. F., Fisher, G., and Bakker, T. (1965) Reporting of hospitalization in the Health Interview Survey. *Vital and Health Statistics*, Series 2, No. 6. Washington, D.C.: U.S. Government Printing Office.

Cannell, C. F., Fowler, F. J., and Marquis, K. H. (1968) The influence of interviewer and respondent psychological and behavioral variables on the reporting in household interviews. *Vital and Health Statistics*, Series 2, No. 26. Washington, D.C.: U.S. Government Printing Office.

Cannell, C. F., Lawson, S. A., and Hausser, D. L. (1975) *A Technique for Evaluating Interviewer Performance*. Ann Arbor: Institute for Social Research, University of Michigan.

Cannell, C. F., Marquis, K. H., and Laurent, A. (1977) A summary of studies of interviewing methodology. *Vital and Health Statistics*, Series 2, No. 69. DHEW publication, no. (HRA)77-1343. Washington, D.C.: U.S. Government Printing Office.

Cannell, C. F., Miller, P. V., and Oksenberg, L. (1981) Research on interviewing techniques. In S. Leinhardt, ed., *Sociological Methodology 1981*. San Francisco: Jossey-Bass.

Cannell, C. F., Oksenberg, L., and Converse, J. M. (1977) Striving for response accuracy: Experiments in new interviewing techniques. *Journal of Marketing Research* 14:306-315.

Cannon, L. W. (1980) On the absolute or relative basis of perception: The case for middle class identification. *Social Indicators Research* 8:347-363.

Cantril, A. H. ed. (1980) *Polling on the Issues*. Cabin John, Md.: Seven Locks Press.

Cantril, H. (1940) Experiments in the wording of questions. *Public Opinion Quarterly* 4:330-332.

Cantril, H. and associates ([1944]1972) *Gauging Public Opinion*. Reprint. Port Washington, N.Y.: Kennikat Press.

Cantril, H. (1945) Do different polls get the same results? *Public Opinion Quarterly* 9:61-69.

Carey et al. v. *Klutznick et al.* (1980) Southern District of New York, 80 Civ 4550, Nov. 10.

Carlsson, G. (1970) Social trends and human decision: Discretionary change as a field of research. In P. Lindblom, ed., *Theory and Methods in Behavioural Science*. Stockholm: Scandinavian University Books.

Carmines, E. G., and Stimson, J. A. (1980) The two faces of issue voting. *American Political Science Review* 74:78-91.

Carpenter, N. (1934) Social surveys. In E. Seligman, ed., *Encyclopaedia of the Social Sciences*, Vol. 14. New York: Macmillan.

Carroll, J. D., and Arabie, P. (1980) Multidimensional scaling. *Annual Review of Psychology* 31:607-649.

Carroll, J. D., and Kruskal, J. B. (1978) Multidimensional scaling. In W. H. Kruskal and J. M. Tanur, eds., *International Encyclopedia of Statistics*, Vol. 2. New York: Free Press.

Center for Political Studies (1976) *A Continuity Guide to the American National Election Surveys of the Center for Political Studies: 1952-1974*. Ann Arbor, Mich.: Inter-university Consortium for Political and Social Research.

Changing Times (1980a) Opinion polls are [ ] accurate, [ ] slanted, [ ] don't know. March.

Changing Times (1980b) Your money and your opinion. March.

Cicourel, A. V. (1974) *Theory and Method in a Study of Argentine Fertility*. New York: Wiley.

Clarren, S., and Schwartz, A. (1976) Measuring a program's impact: A cautionary note. In W. Skogan, ed., *Sample Surveys of the Victims of Crime*. Cambridge, Mass.: Ballinger.

Clausen, A. (1968-69) Response validity: Vote report. *Public Opinion Quarterly* 32:588-606.

Clogg, C. C. (1979) Some latent structure models for the analysis of Likert-type data. *Social Science Research* 8:287-301.

Clogg, C. C. (1982a) Using association models in sociological research: Some examples. *American Journal of Sociology* 88: 114-134.

Clogg, C. C. (1982b) Some models for the analysis of association in multi-way cross-classifications having ordered categories. *Journal of the American Statistical Association* 77:803-815.

Clogg, C. C., and Sawyer, D. O. (1981) A comparison of alternative models for analyzing the scalability of response patterns. In S. Leinhardt, ed., *Sociological Methodology 1981*. San Francisco: Jossey-Bass.

Cochran, W. G. (1953) *Sampling Techniques*. New York: Wiley.

Cohen, S. (1980) Measuring the data collection organization effect on health care variables in the National Medical Care Expenditure Survey. Paper presented at the annual meeting of the American Public Health Association, Detroit, Mich., October.

Cohen, S., and Horvitz, D. (1979) Estimated data collection organization effect in the National Medical Care Expenditure Survey. Paper presented at the annual meeting of the American Public Health Association, New York, November.

References

Cole, E. (1979) The significance to the Society of the survey on public attitudes to market research. *Journal of the Market Research Society* 21:103-106.

Coleman, J. S. (1964) *Introduction to Mathematical Sociology.* New York: Free Press.

Collins, S. (1978) The use of social research in the courts. In L. Lynn, ed., *Knowledge and Policy.* Washington, D.C.: National Academy of Sciences.

Combs, B., and Slovic, P. (1979) Newspaper coverage of causes of death. *Journalism Quarterly* 56:837-843, 849.

Conklin, H. (1962) Lexicographical treatment of folk taxonomies. *International Journal of American Linguistics* 28(2):Part 2 Supplement: 119-141.

Conklin, H. (1964) Ethnogenealogical method. In W. H. Goodenough, ed., *Explorations in Cultural Anthropology.* New York: McGraw-Hill.

Connor, W. D., and Gitelman, Z. Y., with Huszczo, A., and Blumstock, R. (1977) *Public Opinion in European Socialist Systems.* New York: Praeger.

Converse, J. M. (1976-77) Predicting no opinion in the polls. *Public Opinion Quarterly* 40:515-530.

Converse, J. M. (1981) *Survey Research in the United States: Roots and Emergence, 1890–1960.* Institute for Social Research, University of Michigan. Manuscript.

Converse, J. M., and Schuman, H. (1974) *Conversations at Random: Survey Research as Interviewers See It.* New York: Wiley.

Converse, P. E. (1964) The nature of belief systems in mass publics. In D. E. Apter, ed., *Ideology and Discontent.* New York: Free Press.

Converse, P. E. (1970) Attitudes and non-attitudes: Continuation of a dialogue. In E. Tufte, ed., *The Quantitative Analysis of Social Problems.* Reading, Mass.: Addison-Wesley.

Converse, P. E. (1972) Change in the American electorate. In A. Campbell and P. E. Converse, eds., *The Human Meaning of Social Change.* New York: Russell Sage.

Converse, P. E. (1975) Public opinion and voting behavior. In F. I. Greenstein and N. W. Polsby, eds., *Nongovernmental Politics.* Vol. 4 of *Handbook of Political Science.* Reading, Mass.: Addison-Wesley.

Converse, P. E. (1979) The impact of polls on national leadership. Paper presented at Symposium on the Fiftieth Anniversary of the Social Sciences Research Building, University of Chicago, December 16-17.

Converse, P. E., Dotson, J., Hoag, W., and McGee, W. (1980) *American Social Attitudes Data Sourcebook 1947-1978.* Cambridge, Mass.: Harvard University Press.

Coombs, C. H. (1964) *A Theory of Data.* New York: Wiley.

Coombs, C. H., Dawes, R. M., and Tversky, A. (1970) *Mathematical Psychology.* Englewood Cliffs, N.J.: Prentice-Hall.

Council of American Survey Research Organizations (n.d.) CASRO: An organization of and for the survey research business person. (Brochure) Port Jefferson, N.Y.: Council of American Survey Research Organizations

Council of American Survey Research Organizations (n.d./1982) *Code of Standards for Survey Research.* Port Jefferson, N.Y.: Council of American Survey Research Organizations.

Cowan, C., Murphy, L. R., and Wiener, J. (1978) Effects of supplemental questions on victimization estimates from the National Crime Survey. *Proceedings of the American Statistical Association (Survey Research Methods Section)* 1978:277-282.

Crain, R. L., and York, R. L. (1976) Evaluating a successful program: Experimental method and academic bias. *School Review* 84:233-254.

Cronbach, L. J. (1975) Beyond the two disciplines of scientific psychology. *American Psychologist* 30:116-127.

Cronbach, L. J., and Gleser, G. C. (1965) *Psychological Tests and Personnel Decisions,* 2d ed. Urbana: University of Illinois Press.

Cronbach, L. J., and Suppes, P., eds. (1969) *Research for Tomorrow's Schools.* New York: Macmillan.

Davis, J. A. (1975a) Communism, conformity, cohorts and categories: American tolerance in 1954 and 1972-73. *American Journal of Sociology* 81:491-513.

Davis, J. A. (1975b) Does economic growth improve the human lot? Yes indeed about .0005 per year. Paper presented at the International Conference on Subjective Indicators of the Quality of Life, Cambridge, England, September.

Davis, J. A. (1976) *Studies of Social Change Since 1948.* Chicago: National Opinion Research Center.

Davis, J. A. (1978) *General Social Surveys, 1972-1978: Cumulative Codebook.* Chicago: National Opinion Research Center.

Deci, E. L. (1975) *Intrinsic Motivation.* New York: Plenum.

Deming, W. E. (1950) *Some Theory of Sampling.* New York: Wiley.

Dexter, L. A. (1970) *Elite and Specialized Interviewing.* Evanston, Ill.: Northwestern University Press.

Dionne, E., Jr. (1980a) Pollsters do their number on what black America thinks. *New York Times,* January 20.

Dionne, E., Jr. (1980b) Opinion polling by the Russians is oblique art. *New York Times,* February 3.

Dionne, E., Jr. (1980c) 1980 brings more pollsters than ever. *New York Times,* February 16.

Dionne, E., Jr. (1980d) Polls, once scorned, gain new esteem. *New York Times,* April 5.

References

Dionne, E., Jr. (1980e) Experts find polls influence activists. *New York Times*, May 4.

Dionne, E., Jr. (1980f) The business of the pollsters. *New York Times*, June 29.

Dionne, E., Jr. (1980g) Abortion poll: Not clear-cut; wording of a question makes difference. *New York Times*, August 18.

Dodd, S. C. (1947) Standards for surveying agencies. *Public Opinion Quarterly* 11:115-130.

Drew, E. (1979) A reporter at large: Phase—in search of a definition. *New Yorker*, August 27:45-73.

Duncan, B. and Duncan, O. D. (1978) *Sex Typing and Social Roles*. New York: Academic Press.

Duncan, B. L. (1976) Differential social perception and attribution of intergroup violence: Testing the lower limits of stereotyping of blacks. *Journal of Personality and Social Psychology* 34:590-598.

Duncan, O. D. (1961) A socioeconomic index for all occupations. In A. J. Reiss, Jr. et al., *Occupations and Social Status*. New York: Free Press.

Duncan, O. D. (1972) Federal statistics, non-federal statisticians. *Proceedings of the American Statistical Association (Social Statistics Section)* 1972:151-153.

Duncan, O. D. (1975) Does money buy satisfaction? *Social Indicators Research* 2:267-274.

Duncan, O. D. (1978a) Sociologists should reconsider nuclear energy. *Social Forces* 57:1-22.

Duncan, O. D. (1978b) Multiway contingency analysis. *Contemporary Sociology* 7:403-405.

Duncan, O. D. (1979a) How destination depends on origin in the occupational mobility table. *American Journal of Sociology* 84:793-803.

Duncan, O. D. (1979b) Indicators of sex typing: Traditional and egalitarian, situational and ideological responses. *American Journal of Sociology* 85:251-260.

Duncan, O. D. (1979c) Constrained parameters in a model for categorical data. *Sociological Methods and Research* 8:57-68.

Duncan, O. D. (1981) Two faces of panel analysis: Parallels with comparative cross-sectional analysis and time-lagged association. In S. Leinhardt, ed., *Sociological Methodology 1981*. San Francisco: Jossey-Bass.

Duncan, O. D. (1982) Review essay: Statistical methods for categorical data. *American Journal of Sociology* 87:957-964.

Duncan, O. D., and Brody, C. (1982) Analyzing *n* rankings of 3 items. In R. M. Hauser, D. Mechanic and A. O. Haller, eds., *Social Structure and Behavior: Essays in Honor of William Hamilton Sewell*. New York: Academic Press.

Duncan, O. D., and Lieberson, S. (1959) Ethnic segregation and assimilation. *American Journal of Sociology* 64:364-374.

Duncan, O. D., and Schuman, H. (1980) Effects of question wording and context: An experiment with religious indicators. *Journal of the American Statistical Association* 75:269-275.

Duncan, O. D., Schuman, H., and Duncan, B. (1973) *Social Change in a Metropolitan Community.* New York: Russell Sage.

Duncan, O. D., Sloane, D. M., and Brody, C. (1982) Latent classes inferred from response-consistency effects. In K. J. Jöreskog, ed., *Systems under Indirect Observation.* Amsterdam: North Holland.

Durkheim, E. ([1895]1938) *The Rules of Sociological Method.* 8th ed. Trans. S. A. Solovay and J. H. Mueller. Ed. G.E.G. Catlin. Chicago: University of Chicago Press.

Dutka, S., and Frankel, L. R. (1976) *Let's Not Forget About Response Error.* New York: Audits and Surveys, Inc.

Dutka, S., and Frankel, L. R. (1980) Sequential survey design through the use of computer assisted telephone interviewing. *Proceedings of the American Statistical Association (Business and Economic Statistics Section)* 1980:73-76.

Easterlin, R. (1973) Does money buy happiness? *Public Interest* No. 30:3-10.

Easterlin, R. (1974) Does economic growth improve the human lot? Some empirical evidence. In P. Davis and M. Reder, eds., *Nations and Households in Economic Growth.* New York: Academic Press.

Eisenberg, D. (1981) A scientific gold rush. *Science* 213:1104-1105.

Eisenhart, C. (1963) Realistic evaluation of the precision and accuracy of instrument calibration systems. *Journal of Research of the National Bureau of Standards (C. Engineering and Instrumentation)* 67C:161-187.

Ellsworth, P., Cowan, C., and Thompson, W. (1979) Juror attitudes and conviction proneness. Unpublished ms., Department of Psychology, Stanford University.

Ellsworth, P., and Fitzgerald, R. (1979) Due process vs. crime control: The impact of death qualification on jury attitudes. Unpublished ms., Department of Psychology, Stanford University.

Erbring, L. (1980) Media monitoring and models of public opinion dynamics. Paper presented at the 35th annual conference of the American Association for Public Opinion Research, Kings Island, Ohio, May-June.

Erdelyi, M. H. (1974) A new look at the new look: Perceptual defense and vigilance. *Psychological Review* 87:1-25.

Ericsson, K. A., and Simon, H. A. (1980) Verbal reports as data. *Psychological Review* 87:215-251.

Erikson, R. (1980) Descriptive or evaluative indicators in social reporting. (No. 1980-03-18), Stockholm: Swedish Institute for Social Research.

Executive Office of the President (1973) *Social Indicators, 1973.* Office of Management and Budget. Washington, D.C.: U.S. Government Printing Office.

References

Executive Office of the President (1976) *Federal Statistics: Coordination, Standards, Guidelines*. Office of Management and Budget. Washington, D.C.: U.S. Government Printing Office.

Fay, R. F. (1979) On adjusting the Pearson chi-square statistic for clustered sampling. *Proceedings of the American Statistical Association (Social Statistics Section)* 1979:402-406.

Fay, R. F. (1982) CPLX: Contingency table analysis for complex sample designs. Unpublished ms., Bureau of the Census, Statistical Methods Division.

Federal Trade Commission (FTC) (1980a) Initial decision in the matter of Litton Industries, Inc., a corporation, and Litton Systems, Inc., a corporation. Docket No. 9123, filed June 6.

Federal Trade Commission (FTC) (1980b) Impact evaluation: Projects in progress. Staff memorandum, March 1.

Fellegi, I. P. (1980) Approximate tests of independence and goodness of fit based on stratified multistage samples. *Journal of the American Statistical Association* 75:261-268.

Ferber, R. (1966) *The Reliability of Consumer Reports of Financial Assets and Debts*. Bureau of Economic and Business Research, University of Illinois, Urbana.

Ferber, R., Sheatsley, P., Turner, A., and Waksberg, J. (1980) *What is a Survey?* Washington, D.C.: American Statistical Association.

Ferber, R., and Sudman, S. (1974) Effects of compensation in consumer expenditure surveys. *Annals of Economic and Social Measurement* 3:319-331.

Fernandez, E. W., and McKenney, N. R. (1980) Identification of the Hispanic population: A review of Census Bureau experiences. Paper presented at the annual meeting of the American Statistical Association, Houston, Texas.

Ferreira, J., and Slesin, L. (1976) *Observations on the Social Impact of Large Accidents*. Technical Report #122. Cambridge, Mass.: Massachusetts Institute of Technology, Operations Research Center.

Fields, J., and Schuman, H. (1976) Public beliefs about beliefs of the public. *Public Opinion Quarterly* 40:427-448.

Fienberg, S. E. (1977/1980) *The Analysis of Cross-Classified Categorical Data*. 2d. ed. Cambridge, Mass.: MIT Press.

Fienberg, S. E. (1979) The use of chi-squared statistics for categorical data problems. *Journal of the Royal Statistical Society*, (Series B), 41:54-64.

Fienberg, S. E., and Goodman, L. A. (1974) Social Indicators, 1973: Statistical considerations. In R. Van Dusen, ed., *Social Indicators, 1973: A Review Symposium*. New York: Social Science Research Council.

Fischer, E. (1974) Change in anomie in Detroit from the 1950s to 1971. Ph.D. dissertation, University of Michigan, Ann Arbor.

Fischhoff, B., Lichtenstein, S., Slovic, P., Derby, S., and Keeney, R. (1981) *Acceptable Risk*. New York: Cambridge University Press.

Fischhoff, B., Slovic, P., and Lichtenstein, S. (1980) Knowing what you want: Measuring labile values. In T. S. Wallsten, ed., *Cognitive Processes in Choice and Decision Behavior*. Hillsdale, N.J.: Erlbaum.

Fischhoff, B., Slovic, P., and Lichtenstein, S. (1982) Lay foibles and expert fables in judgments about risk. *American Statistician* 36:240-255.

Frake, C. (1961) The diagnosis of disease among the Subanun of Mindanao. *American Anthropologist* 63:113-132.

Frake, C. (1964) Notes on queries in ethnography. *American Anthropologist* 66:-132-145.

Frankel, J., and Sharp, L. (1981) Measurement of respondent burden. *Statistical Reporter* 81:105-111.

Frankel, L. R. (1981) The future of survey research in China—an American viewpoint. *Proceedings of the American Statistical Association (Social Statistics Section)* 1981:119-122.

Franks, F. (1981) *Polywater*. Cambridge, Mass.: MIT Press.

Freeman, J. and Butler, E. W. (1976) Some sources of interviewer variance in surveys. *Public Opinion Quarterly* 40:80-91.

Funkhouser, G. R. (1973) The issues of the sixties: An exploratory study in the dynamics of public opinion. *Public Opinion Quarterly* 37:62-75.

Gallup, G. (1941) Question wording in public opinion polls: Comments on points raised by Mr. Stagner. *Sociometry* 4:259-268.

Gallup, G. (1972) Opinion polling in a democracy. In J. M. Tanur, ed., *Statistics: A Guide to the Unknown*. San Francisco: Holden-Day.

Gallup, G. (1978) *The Gallup Poll: Public Opinion, 1972-1977*. 2 vols. Wilmington, Del.: Scholarly Resources.

Gallup, G., and Rae, S. ([1940]1968) *The Pulse of Democracy*. Reprint. New York: Greenwood Press.

Galtung, J. (1967) *Theory and Methods of Social Research*. New York: Columbia University Press.

General Accounting Office (1978) *Better Guidance and Controls are Needed to Improve Federal Surveys of Attitudes and Opinions*. Comptroller General of the United States. Report GGD 78-24, September 15. Washington, D.C.: General Accounting Office.

Gergen, K. J. (1973) Social psychology as history. *Journal of Personality and Social Psychology* 26:309-320.

Gibson, C. O., Shapiro, G. M., Murphy, L. R., and Stanko, G. J. (1978) Interaction of survey questions as it relates to interviewer-respondent bias. *Proceedings of the American Statistical Association (Survey Research Methods Section)* 1978: 251-256.

References

Gilbert, J., Light, R., and Mosteller, F. (1975) Assessing social innovations: An empirical base for policy. In R. Zeckhauser et al., eds., *Benefit-Cost and Policy Analysis, 1974*. Chicago: Aldine.

Gimbel, F. M. (1979) Press statement. Citizens for Mayor Meier Press Conference, February 21.

Glaser, B., and Strauss, A. (1967) *The Discovery of Grounded Theory*. Chicago: Aldine.

Glasser, G., Metzger, G., and Miaoulis, G. (1970) Measurement and control of interviewer variability. *Proceedings of the American Statistical Association (Business and Economic Statistics Section)* 1970:314-319.

Glock, C., ed. (1967) *Survey Research in the Social Sciences*. New York: Russell Sage.

Goffman, E. (1974) *Frame Analysis*. Cambridge, Mass.: Harvard University Press.

Goldberg, D., Sharp, H., and Freedman, R. (1959) The stability and reliability of expected family size data. *Millbank Memorial Fund Quarterly* 37:369-385.

Goldberg, F. (1971) Toward expansion of Witherspoon. *Harvard Civil Rights and Civil Liberties Law Review* 5:53ff.

Goldberg, L. (1978) Differential attribution of trait descriptive terms to oneself as compared to well-liked, neutral, and disliked others: A psychometric analysis. *Journal of Personality and Social Psychology* 36:1012-1028.

Goldberger, A. S., and Duncan, O. D., eds. (1973) *Structural Equation Models in the Social Sciences*. New York: Seminar Press.

Goldfield, E., Steinberg, J., and Welch, E. (1948) The monthly report on the labor force. *Estadistica* 6:61-67.

Goldthorpe, J. H., and Hope, K. (1974) *The Social Grading of Occupations*. Oxford: Clarendon Press.

Gollin, A. (1980) Exploring the liaison between polling and the press. *Public Opinion Quarterly* 44:445-461.

Gonzalez, M., Ogus, J., Shapiro, G., and Tepping, B. (1975) Standards for discussion and presentation of errors in survey and census data. *Journal of the American Statistical Association* 70(no. 351, pt. 2).

Goodman, L. A. (1965) On the statistical analysis of mobility tables. *American Journal of Sociology* 70:564-585.

Goodman, L. A. (1968) The analysis of cross-classified data: Independence, quasi-independence, and interactions in contingency tables with or without missing entries. *Journal of the American Statistical Association* 63:1091-1131.

Goodman, L. A. (1969) On the measurement of social mobility: An index of status persistence. *American Sociological Review* 34:831-850.

Goodman, L. A. (1971) The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for building models for multiple classifications. *Technometrics* 13:33-61.

Goodman, L. A. (1972) Some multiplicative models for the analysis of cross-classified data. In L. LeCam, J. Neyman, and E. L. Scott, eds., *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. Berkeley: University of California Press.

Goodman, L. A. (1978) *Analyzing Qualitative/Categorical Data*. Ed. J. Magidson. Cambridge, Mass.: Abt Books.

Goodman, L. A. (1979a) Multiplicative models for the analysis of occupational mobility tables and other kinds of cross-classification tables. *American Journal of Sociology* 84:804-819.

Goodman, L. A. (1979b) Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association* 74:537-552.

Goodman, L. A., and Kruskal, W. H. (1979) *Measures of Association for Cross-Classifications*. New York: Springer-Verlag.

Gouldner, A. W. (1970) *The Coming Crisis of Western Sociology*. New York: Basic.

Greenberg, D. (1980a) Polls and other superstitious rituals. *Washington Post*, September 9.

Greenberg, D. (1980b) The plague of polling. *Washington Post*, September 16.

Groves, R., and Kahn, R. (1979) *Surveys by Telephone*. New York: Academic Press.

Gulliksen, H. (1950) *Theory of Mental Tests*. New York: Wiley.

Guttman, L. (1950) The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, and J. A. Clausen, eds., *Measurement and Prediction*. Princeton, N.J.: Princeton University Press.

Haberman, S. J. (1974) Log-linear models for frequency tables with ordered classifications. *Biometrics* 30:589-600.

Haberman, S. J. (1978-1979) *Analysis of Qualitative Data*. 2 vols. New York: Academic Press.

Haller, A. O., and Bills, D. B. (1979) Occupational prestige hierarchies: Theory and evidence. *Contemporary Sociology* 8:721-734.

Hambleton, R., and Cook, L. (1977) Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement* 14:75-96.

Hamblin, R. L. (1971) Mathematical experimentation and sociological theory: A critical analysis. *Sociometry* 34:423-452.

Handlin, O. (1979) *Truth in History*. Cambridge, Mass.: Belknap Press.

References

Hansen, M., Hurwitz, W., and Bershad, M. (1961) Measurement errors in censuses and surveys. *Bulletin of International Statistical Institute* 38:359-374.

Hansen, M., Hurwitz, W., and Madow, W. (1953) *Sample Survey Methods and Theory.* 2 vols. New York: Wiley.

Hansen, M., Hurwitz, W., and Pritzker, L. (1964) The estimation and interpretation of gross differences and the simple response variance. *Contributions to Statistics.* Calcutta, India: Pergamon.

Harman, H. (1976) *Modern Factor Analysis.* 3d ed. Chicago: University of Chicago Press.

Harrop, M. (1980) Social research and market research: A critique of a critique. *Sociology* 14:277-281.

Hartigan, J. A. (1975) *Clustering Algorithms.* New York: Wiley.

Hatchett, S., and Schuman, H. (1975-76) White respondents and race-of-interviewer effects. *Public Opinion Quarterly* 39:523-28.

Henson, R., Cannell, C. F., and Lawson, S. (1973) *Effects of Interviewer Style and Question Form on Reporting of Automobile Accidents.* Ann Arbor: Survey Research Center, University of Michigan.

Hershey, J. C., and Schoemaker, P. J. H. (1980) Risk taking and problem context in the domain of losses: An expected utility analysis. *Journal of Risk and Insurance* 47:111-132.

Ho, C. Y., Powell, R. W., and Liley, P. E. (1974) Thermal conductivity of the elements. *Journal of Physical and Chemical Reference Data* 3(supplement 1):1-244.

Hodge, R. W., Siegel, P. M., and Rossi, P. H. (1964) Occupational prestige in the United States, 1925-63. *American Journal of Sociology* 70:286-302.

Holmberg, S. (1980) The breakthrough of opinion polls on the editorial pages of Swedish newspapers. Paper presented to European Society for Opinion Surveys and Market Research seminar on opinion polls, Bonn, Federal Republic of Germany, January.

Holmes, D. S. (1968) Dimensions of projection. *Psychological Bulletin* 69: 248-268.

Honomichl, J. (1977) Research top 20. *Advertising Age* April 11.

Honomichl, J. (1978) Research top 20: Companies posted 16% gain last year. *Advertising Age* April 24.

Honomichl, J. (1980) Top 23 research leaders up 18%. *Advertising Age* May 19.

Honomichl, J. (1982) How much spent on research? Follow me. *Advertising Age* June 21.

Hope, K. (1982) A liberal theory of prestige. *American Journal of Sociology* 87: 1011-1031.

House J. S., and Wolf, S. (1978) Effects of urban residence on interpersonal trust and helping behavior. *Journal of Personality and Social Psychology* 36:1029-1043.

Hovey *v.* Superior Court of Alameda County. (1980) S. F. 24093. Decision of California Supreme Court, August 28, 1980.

Hovland, C. I., Harvey, O., and Sherif, M. (1957) Assimilation and contrast effects in reactions to communication and attitude change. *Journal of Abnormal and Social Psychology* 55:244-252.

Howell, N. (1979) *Demography of the Dobe !Kung*. New York: Academic Press.

Hunter, J. S. (1977) Quality assessment of measurement methods. In National Research Council, *Environmental Monitoring*, Vol. 4a. Washington, D.C.: National Academy of Sciences.

Hunter, J. S. (1980) The national system of scientific measurement. *Science* 210: 869-874.

Hyman, H. H. (1949) Inconsistencies as a problem in attitude measurement. *Journal of Social Issues* 5:38-42.

Hyman, H. H. (1972) *Secondary Analysis of Sample Surveys*. New York: Wiley.

Hyman H. H., with Cobb, W. J., Feldman, J. J., Hart, C. W., and Stember, C. H. (1954) *Interviewing in Social Research*. Chicago: University of Chicago Press.

Hyman, H. H., and Sheatsley, P. (1950) The current status of American public opinion. In J. Payne, ed., *The Teaching of Contemporary Affairs*. Washington, D.C.: National Council for Social Studies, National Education Association.

Inkeles, A., and Rossi, P. H. (1956) National comparisons of occupational prestige. *American Journal of Sociology* 61:329-339.

Inter-University Consortium for Political and Social Research (1981) *The American National Election Study, 1980: Traditional Time-Series, September-October (C-3) Survey, November-December (C3P0) Survey*. Ann Arbor, Mich.: Inter-University Consortium for Political and Social Research

Isen, A. M., Shalker, T., Clark, M., and Karp, L. (1978) Affect, accessibility of material in memory, and behavior: A cognitive loop? *Journal of Personality and Social Psychology* 36:1-12.

Issacs, S. (1971) Asia anthropology: Science or spying? *Washington Post*, November 22.

Jackman, M. (1973) Education and prejudice or education and response set? *American Sociological Review* 38:327-339.

Jackson, D., Lee, W., and Liu, C. (1980) Aseismic uplift in southern California: An alternative interpretation. *Science* 210:534-536.

James, W. (1890) *The Principles of Psychology*, Vol. 1. New York: H. Holt.

Jensen, R. (1977) Public opinion polls: Early problems of method and philosophy. Paper presented at Oxford Conference on the History and Theory of the Social Sciences, July 23.

Johnson, T. A. (1980) N.A.A.C.P. plans a polling system. *New York Times*, January 15.

References

Jones, E. E., and Nisbett, R. E. (1972) The actor and the observer: Divergent perceptions of the causes of behavior. In E. E. Jones, D. Kanouse, H. Kelley, R. Nisbett, S. Valins, and B. Weiner, eds., *Attribution: Perceiving the Causes of Behavior*. Morristown, N.J.: General Learning Press.

Jöreskog, K. G., and Sörbom, D. (1979) *Advances in Factor Analysis and Structural Equation Models*. Ed. J. Magidson. Cambridge, Mass.: Abt Books.

Jurow, G. (1971) New data on the effect of a death qualified jury on the guilt determination process. *Harvard Law Review* 84:567-611.

Kahn, R. L., and Cannell, C. F. (1957) *The Dynamics of Interviewing*. New York: Wiley.

Kahneman, D., Slovic, P., and Tversky, A. (1982) *Judgment Under Uncertainty*. New York: Cambridge University Press.

Kahneman, D., and Tversky, A. (1979) Prospect theory: An analysis of decision under risk. *Econometrica* 47:263-292.

Kalton, G., Collins, M., and Brook, L. (1978) Experiments in wording opinion questions. *Applied Statistics* 27:149-161.

Kalton, G., and Stowell, R. (1979) A study of coder variability. *Journal of the Royal Statistical Society*, Series C (Applied Statistics) 28:276-289.

Kaplan, A. (1964) *The Conduct of Inquiry*. San Francisco: Chandler Publishing.

Karatnycky, A. (1981) Poland diarist: growing pains. *New Republic*, June 27:43.

Katz, D. (1960) The functional approach to the study of attitudes. *Public Opinion Quarterly* 24:163-204.

Katz, E., and Lazarsfeld, P. F. (1955) *Personal Influence*. Glencoe, Ill.: Free Press.

Kelly, G. A. (1955) *The Psychology of Personal Constructs*. New York: Norton.

Kelman, H. C. (1961) Processes of opinion change. *Public Opinion Quarterly* 25:57-78.

Kempf, W. F., Hampapa, P., and Mach, G. (1975) Conditional maximum likelihood estimation for a dynamic test model. IPN-Arbeitsberichte 13. Kiel: Institut für die Pädagogik an der Naturwissenschaften an der Christian-Albrechts-Universität.

Kempf, W. F., and Repp, B. H., eds. (1977) *Mathematical Models for Social Psychology*. New York: Wiley.

Kiesler, C. A. (1971) *The Psychology of Commitment*. New York: Academic Press.

Kiesler S. B., and Turner, C. F., eds. (1977) *Fundamental Research and the Process of Education*. Washington, D.C.: National Academy of Sciences.

King, L., and Klockars, A. (1982) Imposed differences in comparative ratings. *Journal of Personality and Social Psychology* 42:631-636.

Kinsley, M. (1981) The art of polling: how you and Union Carbide came to agree. *New Republic*, June 20:16-19.

Kish, L. (1962) Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association* 57:92-115.

Klapper, J. T. (1960) *The Effects of Mass Communication*. Glencoe, Ill.: Free Press.

Koch, G. G. (1973) An alternative approach to multivariate response error models for sample survey data with applications to estimators involving subclass means. *Journal of the American Statistical Association* 68:906-913.

Koch, G. G., Freeman, J. L., and Lehnen, R. (1975) A contingency table approach to the analysis of ranked policy preference data. *Proceedings of the American Statistical Association (Social Statistics Section)* 1975:507-512.

Koch, G. G., Freeman, J. L., and Lehnen, R. (1976) A general methodology for the analysis of ranked policy preference data. *International Statistical Review* 44:1-28.

Kohn, M. L. (1969) *Class and Conformity*. Homewood, Ill.: Dorsey.

Kornhauser, A. (1946-47) Are opinion polls fair to organized labor? *Public Opinion Quarterly* 10:484-500.

Kramer, J. (1981) Letter from Europe. *New Yorker*, December 21.

Krantz, D. H., Luce, R. D., Suppes, P., and Tversky, A. (1971) *Foundations of Measurement*, Vol. 1. New York: Academic Press.

Krantz, D. H., and Tversky, A. (1971) Conjoint-measurement analysis of composition rules in psychology. *Psychological Review* 78:151-169.

Kraus, S., and Davis, D. (1976) *The Effects of Mass Communication on Political Behavior*. University Park: Pennsylvania State University Press.

Kraut, R. E., and McConahay, J. B. (1973) How being interviewed affects voting: An experiment. *Public Opinion Quarterly* 37:398-406.

Kruskal, J. B. (1978) Factor analysis and principal components, I. Bilinear methods. In W. H. Kruskal and J. M. Tanur, eds., *International Encyclopedia of Statistics*. Vol. 1. New York: Free Press.

Kruskal, J. B., and Wish, M. (1978) *Multidimensional Scaling*. Beverly Hills, Calif.: Sage.

Kruskal, W. H., and Mosteller, F. (1980) Representative sampling, IV: The history of the concept in statistics, 1895-1939. *International Statistical Review* 48: 169-195.

Kunreuther, H., Ginsberg, R., Miller, L., Sagi, P., Slovic, P., Borkan, B., and Katz, N. (1978) *Disaster Insurance Protection*. New York: Wiley.

Ladd, E. C., Jr., and Lipset, S. M. (1977) The faculty mood: Pessimism is predominant. *Chronicle of Higher Education*, Oct. 3.

Ladd, E. C., Jr., and Lipset, S. M. (1978) Professors found to be liberal but not radical. *Chronicle of Higher Education*, Jan. 16.

Lane, R. E. (1962) *Political Ideology*. New York: Free Press.

References

Lang, S. (1978) The professors: A survey of a survey. *New York Review of Books*, May 18.

Lang, S. (1981) *The File*. New York: Springer-Verlag.

Lanouette, W. (1980) Polls and pols—with a grain of SALT. In A. Cantril, ed., *Polling on the Issues*. Cabin John, Md: Seven Locks Press.

Lansing, J. B., Ginsberg, G. P., and Braaten, K. (1961) *An Investigation of Response Error*. Studies in Consumer Finance, No. 2. Urbana, Ill.: Bureau of Economic and Business Research, University of Illinois.

La Porte, T., with Chisolm, D. (1980) Indicators of public attitudes toward science and technology. *Scientometrics* 2:439-468.

Lasswell, H. D. (1930) *Psychopathology and Politics*. Chicago: University of Chicago Press.

Laumann, E. O. (1973) *Bonds of Pluralism*. New York: Wiley.

Lazarsfeld, P. F. (1937) Some remarks on the typological procedure in social research. *Zeitschrift für Sozialforschung* 6:119-139.

Lazarsfeld, P. F. (1944) The controversy over detailed interviews—an offer for negotiation. *Public Opinion Quarterly* 8:38-60.

Lazarsfeld, P. F., Berelson, B., and Gaudet, H. (1944) *The People's Choice*. New York: Columbia University Press.

Lazarsfeld, P. F., and Henry, N. W. (1968) *Latent Structure Analysis*. New York: Houghton Mifflin.

Lazarsfeld, P. F., and Thielens, W., Jr. (1958) *The Academic Mind*. Glencoe, Ill.: Free Press.

League of Women Voters (1980a) News release: 1980 Presidential debates, September 9, Washington, D.C.

League of Women Voters (1980b) News release: 1980 Presidential debates, October 17, Washington, D.C.

Lenski, G., and Leggett, J. (1960) Caste, class, and deference in the research interview. *American Journal of Sociology* 65:463-467.

Lepkowski, J. (1981) Survey Research Center sampling paper. Unpublished ms., Survey Research Center, University of Michigan, Ann Arbor.

Levine, D. B. (1968) The use of "sensitive" questions in Census Bureau surveys. Paper presented at the Social Science Research Council meeting, New York, September, 19-20.

Lide, D. R., Jr. (1981) Critical data for critical needs. *Science* 212:1343-1349.

Li Ming (1981) A promising young generation: an opinion poll of young. *Beijing Review* (formerly *Peking Review*) July 27:21-22.

Lindzey, G., and Guest, L. (1951) To repeat—check lists can be dangerous. *Public Opinion Quarterly* 15:355-358.

Link, R. (1980) The Literary Digest poll: Appearances can be deceiving. *Public Opinion* February/March:55.

Linton, M. (1975) Memory for real-world events. In D. A. Norman, D. E. Rumelhart, and the LNR Research Group. *Explorations in Cognition.* San Francisco: W. H. Freeman.

Lipset, S. M., Lazarsfeld, P. F., Barton, A. H., and Linz, J. (1954) The psychology of voting: An analysis of political behavior. In G. Lindzey, ed., *Handbook of Social Psychology,* Vol. 2. Cambridge, Mass.: Addison-Wesley.

Lizotte, A. J. and Bordua, D. J. (1980) Firearms ownership for sport and protection: Two divergent models. *American Sociological Review* 45:229-244.

Lockwood, B. (n.d.) Untitled [Unpublished analyses and tabulations of data from Stouffer (1955)]. Department of Sociology, University of Arizona.

Lodge, M., Cross, D. V., Tursky, B., and Tanenhaus, J. (1975) The psychophysical scaling and validation of a political support scale. *American Journal of Political Science* 19:611-649.

Loevinger, J. (1947) A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs* 61, no. 285.

Loevinger, J. (1965) Person and population as psychometric concepts. *Psychological Review* 72:143-155.

Loftus, E. F. (1975) Leading questions and the eyewitness report. *Cognitive Psychology* 7:560-572.

Loftus, E. F. and Marburger, W. (1983) Since the eruption of Mt. St. Helens, did anyone beat you up? Improving the accuracy of retrospective reports with landmark events. *Memory and Cognition* 11:114-120.

Loftus, E. F. and Palmer, J. C. (1974) Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior* 13:585-589.

Lord, F. M. (1975) The "ability" scale in item characteristic curve theory. *Psychometrika* 40:205-217.

Lord, F. M., and Novick, M. R. (1968) *Statistical Theories of Mental Test Scores.* Reading, Mass.: Addison-Wesley.

Louis Harris and Associates, Inc. (1975) *A Survey of Public and Leadership Attitudes toward Nuclear Power Development in the United States.* New York: Ebasco Services Inc.

Lowell, A. L. (1913) *Public Opinion and Popular Government.* New York: Longmans and Green.

Lowrance, W. W. (1976) *Of Acceptable Risk.* Los Altos, Calif.: Kaufmann.

Luce, R. D. (1959) *Individual Choice Behavior.* New York: Wiley.

Luce, R. D. (1979) The choice axiom after twenty years. In R. K. Merton, J. S. Coleman, and P. H. Rossi, eds., *Qualitative and Quantitative Social Research.* New York: Free Press.

References

Luce, R. D., and Tukey, J. W. (1964) Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology* 1:1-27.

MacBride, S. (1980) The power that public opinion wields. *Times* (London) March 21.

Machlup, F. (1962) *The Production and Distribution of Knowledge in the United States*. Princeton, N.J.: Princeton University Press.

MacKuen, M. (1981) Social communication and the mass policy agenda. In M. MacKuen and S. Coombs, eds., *More Than News: Media Power in Public Affairs*. Beverly Hills, Calif.: Sage.

Mac Lane, S. (1978) Statement. Presented at Social Science Research Council review symposium on *Science Indicators 1976*, Washington D. C., May 13.

Madansky, A. (1968) Latent structure. In D. Sills, ed., *International Encyclopedia of the Social Sciences*, Vol. 9. New York: Macmillan.

Mahalanobis, P. C. (1946) Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society* 109:326-378.

Mandelbrodt, B. B. (1977) *Fractals: Form, Chance and Dimension*. San Francisco: W.H. Freeman.

Mannheim, K. (1931/1936) Wissenssoziologie. In A. Vierkandt, *Handwörterbuch der Soziologie*. Stuttgart: F. Enke. (Tran. by L. Wirth and E. Shils, as Chapter 5 of Mannheim, K., *Ideology and Utopia*. Garden City, N.Y.: International Library of Psychology, Philosophy, and Scientific Method, Harcourt, Brace and World.)

Marketing News (1981) Marketing research industry survey finds increase in phone interviewing. January 9:20.

Marks, L. E. (1974) *Sensory Processes*. New York: Academic Press.

Marquis, K. H., and Cannell, C. F. (1969) *A Study of Interviewer-Respondent Interaction in the Urban Employment Survey*. Ann Arbor: Institute for Social Research, University of Michigan.

Marsh, C. (1976) *Guidelines for Commissioning an Interview Survey from a Research Company*. London: Social Science Research Council Survey Unit.

Marsh, C. (1979) How would you say you felt about political opinion surveys? Would you say you were very happy, fairly happy, or not too happy? Paper presented at British Sociological Association and Social Science Research Council Conference on Methodology and Techniques of Sociology, Lancaster, England, January.

Marsh, C. (1982) *The Survey Method*. London: George Allen and Unwin.

Martin, E. (1980) The effects of item contiguity and probing on measures of anomia. *Social Psychology Quarterly* 43:116-120.

Martin, E., McDuffee, D., and Presser, S. (1981) *Sourcebook of Harris National Surveys*. IRSS Technical Papers 6. Chapel Hill: Institute for Research in Social Science, University of North Carolina.

Martin, E. G. (1971) New McCarthyism? Blackball of scholars for links to the war stirs heated debate. *Wall Street Journal*, November 18.

Mauser, G. A. (1972) A structural approach to predicting patterns of electoral substitution. In A. K. Romney, R. N. Shepard, and S. B. Nerlove, eds., *Multidimensional Scaling*, Vol. 2. New York: Seminar Press.

Mauser, G. A. (1979) A technology for marketing political candidates. In A. D. Shocker, ed., *Analytic Approaches to Marketing Planning*. Cambridge, Mass.: Marketing Science Institute.

Mauser, G. A. (1983) *Political Marketing: An Approach to Campaign Strategy*. New York: Praeger.

McCombs, M., and Shaw, D. (1972) The agenda-setting function of mass media. *Public Opinion Quarterly* 36:176-187.

McConahay, J., Mullin, C., and Frederick, J. (1977) The uses of social science in trials with political and racial overtones: The trial of Joan Little. *Law and Contempory Problems* 41:205-229.

McDonald, R. (1980) Special report on annual revenues and profits of several organizations doing social science research. National Opinion Research Center, University of Chicago. Nov. 6 memorandum.

McFarland, S. (1981) Effects of question order on survey responses. *Public Opinion Quarterly* 45:208-215.

McGuire, W. J. (1960) A syllogistic analysis of cognitive relationships. In M. J. Rosenberg, C. I. Hovland, W. J. McGuire, R. P. Abelson, and J. W. Brehm, eds., *Attitude Organization and Change*. New Haven: Yale University Press.

McGuire, W. J. (1969) Suspiciousness of experimenter's intent. In R. Rosenthal and R. L. Rosnow, eds., *Artifact in Behavioral Research*. New York: Academic Press.

McGuire, W. J., and Padawer-Singer, A. (1976) Trait salience in the spontaneous self-concept. *Journal of Personality and Social Psychology* 33:743-754.

McKennell, A. C. (1974) *Surveying Attitude Structures*. Amsterdam: Elsevier.

McNemar, Q. (1946) Opinion-attitude methodology. *Psychological Bulletin* 43:289-374.

Mead, G. H. (1934) *Mind, Self, and Society*. Chicago: University of Chicago Press.

Metzger, D., and Williams, G. E. (1962) Patterns of primary reference in a Tzeltal community. Stanford: Anthropology Research Projects. Ms. no. 10.

Metzger, D., and Williams, G. E. (1963a) A formal ethnographic analysis of Tenejapa Ladino weddings. *American Anthropologist* 65:1076-1101.

Metzger, D., and Williams, G. E. (1963b) Tenejapa medicine I: The curer. *Southwestern Journal of Anthropology* 19:216-234.

References

Meyers, S., and Oliver, J. (1978) Privacy and hostility toward government as reasons for nonresponse in the National Medical Care Expenditure Survey. *Proceedings of the American Statistical Association (Social Statistics Section)* 1978:509-513.

Miller, T. W., and Wasserman, S. S. (1979) Review of "Mathematical models for social psychology," by W. Kempf and B. Repp. *Journal of the American Statistical Association* 74:936-937.

Miller, W. E. (1979) Crisis of confidence II: Misreading the public pulse. *Public Opinion* 2(Oct.-Nov.).

Minsky, M. (1975) A framework for representing knowledge. In P. H. Winston, ed., *Psychology of Computer Vision*. New York: McGraw-Hill.

Mitchell, R. C. (1980a) Polling on nuclear power: A critique of the polls after Three Mile Island. In A. Cantril, ed., *Polling on the Issues*. Cabin John, Md.: Seven Locks Press.

Mitchell, R. C. (1980b) Addendum to: "Polling on nuclear power: A critique of the polls after Three Mile Island." In A. Cantril, ed., *Polling on the Issues*. Cabin John, Md: Seven Locks Press.

Mitofsky, W. J. (1981) The 1980 pre-election polls: A review of disparate methods and results. *Proceedings of the American Statistical Association (Survey Research Methods Section)* 1981:47-52.

Moser, C. A., and Kalton, G. (1972) *Survey Methods in Social Investigation*. 2d ed. New York: Basic.

Mosteller, F., Hyman, H., McCarthy, P., Marks, E., and Truman, D. (1949) *The Pre-Election Polls of 1948*. Bulletin 60. New York: Social Science Research Council.

Mosteller, F., and Tukey, J. W. (1977) *Data Analysis and Regression*. Reading, Mass.: Addison-Wesley.

Mueller, J. E. (1973) *War, Presidents, and Public Opinion*. New York: Wiley.

Murphy, J. M., Murfin, G. D., Jamieson, N. L., Rambo, A. T., Glenn, J. A., Jones, L. P., and Leighton, A. H. (1974) *The Effects of Herbicides in South Vietnam, Part B: Working Papers: Beliefs, Attitudes, and Behavior of Lowland Vietnamese*. Washington D.C.: National Academy of Sciences.

Myers, D. G. (1980) Polarizing effects of social interaction. In H. Brandstatter, J. H. Davis, and G. Stocker-Kreichgauer, eds., *Contemporary Problems in Group Decision-Making*. New York: Academic Press.

Myers, V. (1977) Toward a synthesis of ethnographic and survey methods. *Human Organization* 36:244-251.

National Council on Public Polls (1979) *Principles of Disclosure*. Washington, D.C.: National Council on Public Polls.

National Council on Public Polls (1980a) Press Release: Letter to Roone Arledge at ABC-TV. October 31. Washington, D.C.

National Council on Public Polls (1980b) Press Release: Polling association cautions League of Women Voters on use of polls for debates decision. August 22. Washington D.C.

National Employers Life Assurance Company Limited v. Advisory, Conciliation and Arbitration Service (ACAS) and Association of Scientific, Technical and Managerial Staffs (ASTMS) (1979) Judgment of Mr. Justice Browne-Wilkinson, High Court of Justice, Chancery Division, February 23.

National Opinion Research Center (NORC) (1972a) *Manual of Procedures for Hiring and Training Interviewers*. Rev. ed. Chicago: National Opinion Research Center.

National Opinion Research Center (NORC) (1972b) *A Brush-Up on Interviewing Techniques*. Chicago: National Opinion Research Center.

National Opinion Research Center (NORC) (1973) *Southern Schools*. 2 vols. Chicago: National Opinion Research Center.

National Opinion Research Center (NORC) (1978) *General Social Surveys, 1972-1978: Cumulative Codebook*. Chicago: National Opinion Research Center.

National Opinion Research Center (NORC) (1980) *General Social Surveys, 1972-1980: Cumulative Codebook*. Chicago: National Opinion Research Center.

National Research Council (1977) *Community Reactions to the Concorde*. Washington, D.C.: National Academy of Sciences.

National Research Council (1979) *Privacy and Confidentiality as Factors in Survey Response*. Washington, D.C.: National Academy of Sciences.

National Science Board (1973) *Science Indicators: 1972*. Washington, D.C.: U.S. Government Printing Office.

National Science Board (1975) *Science Indicators: 1974*. Washington, D.C.: U.S. Government Printing Office.

National Science Board (1977) *Science Indicators: 1976*. Washington D.C.: U.S. Government Printing Office.

Navazio, R. (1977) An experimental approach to bandwagon research. *Public Opinion Quarterly* 41:217-225.

Neuman, W. R. (1979) Candidates and issues in the mass media: A trend analysis of media content and public opinion, 1952-1976. Paper presented at the Annual Meeting of the American Political Science Association, Washington, D.C.

Newton, I. (1687) *Mathematical Principles of Natural Philosophy*. (preface).

New York Times (1975) Gannett chain in pact for Louis Harris firm. March 19:79.

Neyman, J. (1934) On the two different aspects of the representative method: The method of stratified sampling and the method of purposive sampling. *Journal of the Royal Statistical Society* (Series A) 97:558-606.

Neyman, J. (1938) Contribution to the theory of sampling human populations. *Journal of the American Statistical Association* 33:101-116.

References

Nicholls, W. (1980) A project plan for a program of research into and development of a computer-assisted data collection system and assistance in implementation planning of a CATI system. U.S. Bureau of the Census, Center for Social Science Research.

Nie, N. H., with Andersen, K. (1974) Mass belief systems revisited: Political change and attitude structure. *Journal of Politics* 36:540-591.

Nie, N. H., and Rabjohn, J. N. (1979) Revisiting mass belief systems revisited: Or, doing research is like watching a tennis match. *American Journal of Political Science* 23:139-175.

Nie, N. H., Verba, S., and Petrocik, J. R. (1979) *The Changing American Voter.* Enlarged ed. Cambridge, Mass.: Harvard University Press.

Nisbet, R. (1975) Public opinion versus popular opinion. *Public Interest* 41:166-192.

Nisbett, R. E., and Bellows, N. (1977) Verbal reports about causal influences on social judgments: Private access versus public theories. *Journal of Personality and Social Psychology* 35:613-624.

Nisbett, R. E., and Wilson, T. D. (1977) Telling more than we can know: Verbal reports on mental processes. *Psychological Review* 84:231-259.

Noelle, E., and Neumann, E. (1967) *The Germans: Public Opinion Polls 1947-1966.* Allensbach, West Germany: Verlag für Demoskopie.

Noelle-Neumann, E. (1973) Return to the concept of powerful mass media. In *Studies in Broadcasting,* Radio and TV Culture Research Institute Nippon Hoso Kyokai.

Noelle-Neumann, E. (1974) The spiral of silence: A theory of public opinion. *Journal of Communication* 24:43-51.

Noelle-Neumann, E. (1977) Turbulences in the climate of opinion: Methodological applications of the spiral of silence theory. *Public Opinion Quarterly* 41:143-158.

Noelle-Neumann, E. (1979) Public opinion and the classical tradition: A re-evaluation. *Public Opinion Quarterly* 43:143-156.

North American Newspaper Alliance (NANA) (1978) Press release: Survey of New Right poorly worded and meaningless: NANA poll. By R. J. Wagman and S. D. Engelmayer. New York: NANA.

Nunnally, J. C. (1978) *Psychometric Theory.* 2d ed. New York: McGraw-Hill.

Office of State Budget and Management (1979) *Preliminary Report to the Governor: 1979 Spring Citizen Survey.* August 1, Raleigh: State of North Carolina.

Office of State Budget and Management (1980) The North Carolina Citizen Survey. Memorandum, June. Raleigh: State of North Carolina.

O'Neill, H. W. (1980) The pollster and the advocate. In A.H. Cantril, ed., *Polling on the Issues.* Cabin John, Md.: Seven Locks Press.

Opinion Outlook (1981) Inside opinion: The professional poll respondent. June 22:7.

Organization for Economic Cooperation and Development (OECD) (1973) *List of Social Concerns Common to Most OECD Countries*. Paris: Organization for Economic Cooperation and Development.

Organization for Economic Cooperation and Development (OECD) (1976) *Measuring Social Well-Being*. Paris: Organization for Economic Cooperation and Development.

Paletz, D., Short, J., Baker, H., Campbell, B., Cooper, R., and Oeslander, R. (1980) Polls in the media: Content, credibility, and consequences. *Public Opinion Quarterly* 44:495-513.

Panel on Incomplete Data, ed. (1983) *Incomplete Data in Sample Surveys*. 3 vols. New York: Academic Press.

Parsons, T. (1937) *The Structure of Social Action*. New York: McGraw-Hill.

Parten, M. (1950) *Surveys, Polls, and Samples*. New York: Harper.

Payne, G. (1979) Social research and market research. *Sociology* 13:307-313.

Payne, S. (1951) *The Art of Asking Questions*. Princeton, N.J.: Princeton University Press.

Penick, B. K., and Owens, M., eds. (1976) *Surveying Crime*. Washington, D. C.: National Academy of Sciences.

Pepitone, A., and DiNubile, M. (1976) Contrast effects in judgments of crime severity and the punishment of criminal violators. *Journal of Personality and Social Psychology* 33:448-459.

Perline, R., Wright, B. D., and Wainer, H. (1979) The Rasch model as additive conjoint measurement. *Applied Psychological Measurement* 3:237-255.

Peterson, I. (1979) Carter speech scores in Midwest TV poll. *New York Times*, July 16.

Pettigrew, T. F. (1964) *A Profile of the Negro American*. New York: Van Nostrand.

Pfanzagl, J. (1968) *Theory of Measurement*. New York: Wiley.

Piazza, T. (1980) The analysis of attitude items. *American Journal of Sociology* 86:584-603.

Pineo, P. C. (1977) The social standing of ethnic and racial groupings. *Canadian Review of Sociology and Anthropology* 14:147-157.

Plott, C. R., and Levine, M. E. (1978) A model of agenda influence on committee decisions. *American Economic Review* 68:146-160.

Polanyi, M. (1964) *Personal Knowledge*. New York: Harper.

Poulton, E. C. (1968) The new psychophysics: Six models for magnitude estimation. *Psychological Bulletin* 69:1-19.

Poulton, E. C. (1977) Quantitative subjective judgments are almost always biased, sometimes completely misleading. *British Journal of Psychology* 68:409-425.

References

Powley *v.* Advisory, Conciliation and Arbitration Service (ACAS) and Legal and General Life Assurance Company and Association of Scientific, Technical and Managerial Staffs (ASTMS) (1977) *Industrial Cases Reports* (U.K.) 1978:123.

President's Commission on a National Agenda for the Eighties (1980) *The Quality of American Life in the Eighties.* Final report of the panel on the quality of American life. Washington, D.C.: U.S. Government Printing Office.

Presser, S. (1980) Comment on Nelson and on Baloyra. *American Political Science Review* 74:459-460.

Presser, S. (1981) The Harris Data Center, Harris national surveys, and trend analysis. In E. Martin, D. McDuffee, and S. Presser, eds., *Sourcebook of Harris National Surveys.* IRSS Technical Papers 6. Chapel Hill: Institute for Research in Social Science, University of North Carolina.

Presser, S., and Schuman, H. (1980) The measurement of a middle position in attitude surveys. *Public Opinion Quarterly* 44:70-85.

Price, D. O., and Searles, R. (1961) Some effects of interviewer-respondent interaction on responses in a survey situation. *Proceedings of the American Statistical Association (Social Statistics Section)* 1961:211-221.

Pryor, J. B., and Kriss, M. (1977) The cognitive dynamics of salience in the attribution process. *Journal of Personality and Social Psychology* 35:49-55.

Pryor, J., Gibbons, F., Wicklund, R., Fazio, R., and Hood, R. (1977) Self-focused attention and self-report validity. *Journal of Personality* 45:513-527.

Rainwater, L. (1974) *What Money Buys.* New York: Basic.

Rao, J. N. K., and Scott, A. J. (1981) The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association* 76:221-230.

Rasch, G. ([1960]1980) *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Denmarks Paedagogiske Institut. Expanded ed., Chicago: University of Chicago Press.

Rasch, G. (1966) An individualistic approach to item analysis. In P. F. Lazarsfeld and N. W. Henry, eds., *Readings in Mathematical Social Science.* Chicago: Science Research Associates.

Reeves, T. (1979) Ballotting in trade union recognition disputes. Paper presented at British Sociological Association Annual Conference, April.

Reiss, A. J., Jr. (1980) Exploring the central paradox in methods of social science inquiry. Paper presented at annual meeting of American Sociological Association, New York.

Reiss, A. J., Jr., with O. D. Duncan, P. K. Hatt, and C. C. North (1961) *Occupations and Social Status.* New York: Free Press.

Riesman, D. (1979) Ethical and practical dilemmas of fieldwork in academic settings: A personal memoir. In R. K. Merton, J. S. Coleman, and P. H. Rossi, eds., *Qualitative and Quantitative Social Research.* New York: Free Press.

Roberts, F. (1979) Measurement theory: With applications to decision making, utility, and the social sciences. *Encyclopedia of Mathematics and its Applications*, Vol. 7. Reading, Mass.: Addison-Wesley.

Robinson, C. E. (1934) Straw vote. In E. Seligman, ed., *Encyclopaedia of the Social Sciences*, Vol. 14. New York: Macmillan.

Roll, C., and Cantril, A. (1980) *Polls: Their Use and Misuse in Politics*. Reprint. Cabin John, Md.: Seven Locks Press.

Romney, A. K., Shepard, R. N., and Nerlove, S. B. (1972) *Multidimensional Scaling*. Vol. 2. New York: Seminar Press.

Roper, B. W. (1980) The impact of journalism on polling. In A. H. Cantril, ed., *Polling on the Issues*. Cabin John, Md.: Seven Locks Press.

Rose, A. M. (1951) Discussion of "Measuring changes in ethnic reactions," by E. S. Bogardus. *American Sociological Review* 16:52-53.

Rosenberg, M. J. (1968) *The Logic of Survey Analysis*. New York: Basic.

Rosenberg, M. J., Verba, S., and Converse, P. E. (1970) *Vietnam and the Silent Majority: The Dove's Guide*. New York: Harper and Row.

Rosenthal, R., and Rosnow, R. (1969) The volunteer subject. In R. Rosenthal and R. Rosnow, eds., *Artifact in Behavioral Research*. New York: Academic Press.

Ross, L. (1977) The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz, ed., *Advances in Experimental Social Psychology*, Vol. 10. New York: Academic Press.

Rossi, P. H. (1979) Vignette analysis: Uncovering the normative structure of complex judgments. In R. K. Merton, J. S. Coleman, and P. H. Rossi, eds., *Qualitative and Quantitative Social Research*. New York: Free Press.

Rossi, P. H., Waite, E., Bose, C., and Berk, R. (1974) The seriousness of crimes: Normative structure and individual differences. *American Sociological Review* 39: 224-237.

Rothwell, N. D. (1971) Some notes about what could be learned as a result of using Form S-155, investigation of questions on Pilot Cities Victimization Survey. Response Research Branch Report no. 71-10. Bureau of the Census. May 3.

Rothwell, N. D. (1979) Mock interview project report. Bureau of the Census. July 16 memorandum.

Rothwell, N. D., and Tracey, E. (1966) Observation of interviews conducted during the Community Action Project (CAP) survey in the Cardozo area. Bureau of the Census. March 3 memorandum.

Rugg, W. D. (1941) Experiments in wording questions: II. *Public Opinion Quarterly* 5:91-92.

Rugg, W. D., and Cantril, H. ([1944]1972) The wording of questions. In H. Cantril and associates, *Gauging Public Opinion*. Reprint. Port Washington, N.Y.: Kennikat Press.

References

Rugg W. D. (1971-72) Interviewer opinion on the "salesman as interviewer" problem. *Public Opinion Quarterly* 35:625-626.

Rustemeyer, A. (1977) Measuring interviewer performance in mock interviews. *Proceedings of the American Statistical Association (Social Statistics Section)* 1977: 341-346.

Sabato, L. J. (1981) *The Rise of Political Consultants*. New York: Basic.

Safarov, R. (1981) Questions of theory: The study and efficacy of public opinion. (In Russian.) *Pravda* September 25. Excerpted in *Current Digest of the Soviet Press* (1981) 33:23.

Sagar, H. A., and Schofield, J. W. (1980) Racial and behavioral cues in black and white children's perceptions of ambiguously aggressive acts. *Journal of Personality and Social Psychology* 39:590-598.

Saks, M. J. (1976) Social scientists can't rig juries. *Psychology Today* 9 (January).

Salancik, J. (1974) Inference of one's attitude from behavior recalled under linguistically manipulated cognitive sets. *Journal of Experimental Social Psychology* 10:415-427.

Sanoff, A. P. (1981) The perils of polling 1980: How the takers of the nation's pulse missed the beat. *Washington Journalism Review* Jan.-Feb.:32-35.

Schaeffer, N. C. (1980) Evaluating race-of-interviewer effects in a national survey. *Sociological Methods and Research* 8:400-419.

Schanck, R. L. (1932) A study of a community and its groups and institutions conceived of as behaviors of individuals. *Psychological Monographs* 43 (whole no. 195).

Schank, R. C., and Abelson, R. P. (1977) *Scripts, Plans, Goals, and Understanding*. Hillsdale, N.J.: Erlbaum.

Schönborn, M. (1981) Perceptions of national security of the U.S. and the Federal Republic of Germany: Stability and change in public opinion. Paper presented at 36th Annual Conference of the American Association for Public Opinion Research, Buck Hill Falls, Pa., May 28-31.

Schulman, J., Shaver, P., Colman, R., Emrich, B., and Christie, R. (1973) Recipe for a jury. *Psychology Today*, May.

Schuman, H. (1972) Two sources of antiwar sentiment in America. *American Journal of Sociology* 78:513-536.

Schuman, H. (1974) Old wine in new bottles: Some sources of response error in the use of attitude surveys to study social change. Paper presented for the Research Seminar Group in Quantitative Social Science, University of Surrey, England.

Schuman, H. (1982) Artifacts are in the mind of the beholder. *American Sociologist* 17:21-28.

Schuman, H., and Converse, J. M. (1971) The effects of black and white interviewers on black responses in 1968. *Public Opinion Quarterly* 35:44-68.

Schuman, H., and Duncan, O. D. (1974) Questions about attitude survey questions. In H. L. Costner, ed., *Sociological Methodology 1973-74*. San Francisco: Jossey-Bass.

Schuman, H., and Gruenberg, B. (1970) The impact of city on racial attitudes. *American Journal of Sociology* 76:213-261.

Schuman, H., and Johnson, M. P. (1976) Attitudes and behavior. *Annual Review of Sociology* 2:161-207.

Schuman, H., and Presser, S. (1977) Question wording as an independent variable in survey analysis. *Sociological Methods and Research* 6:151-170.

Schuman, H., and Presser, S. (1977-78) Attitude measurement and the gun control paradox. *Public Opinion Quarterly* 41:427-438.

Schuman, H., and Presser, S. (1979) The assessment of "no opinion" in attitude surveys. In K. F. Schuessler, ed., *Sociological Methodology 1979*. San Francisco: Jossey-Bass.

Schuman, H., and Presser, S. (1980) Heads or tails, he wins. *New York Times*, July 17:A-19.

Schuman, H., and Presser, S. (1981) *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. New York: Academic Press.

Schwartz, S. (1977) Normative influences on altruism. In L. Berkowitz, ed., *Advances in Experimental Social Psychology*, Vol. 10. New York: Academic Press.

Scitovsky, T. (1976) *The Joyless Economy*. New York: Oxford University Press.

Scott, C., and Sabagh, G. The historical calendar as a method of estimating age: The experience of the Moroccan multi-purpose sample survey of 1961-63. *Population Studies* 24:93-109.

Sears, D. O. (1968) Political behavior. In G. Lindzey and E. Aronson, eds., *Handbook of Social Psychology*, Vol. 5. 2d ed. Reading, Mass.: Addison-Wesley.

Sears, D. O., and McConahay, J. (1973) *The Politics of Violence*. Boston: Houghton Mifflin.

Selltiz, C., Wrightsman, L. S., and Cook, S. W. (1976) *Research Methods in Social Relations*. 3d. ed. New York: Holt, Rinehart and Winston.

Seltzer, R., with Turner, C. F. (1981) Using social science skills in political struggle: Racism, police brutality and the trial of Terrence Johnson. *Insurgent Sociologist* 10:67-74.

Selvin, H. C. (1968) Survey analysis: Methods of survey analysis. In D. Sills, ed., *International Encyclopedia of the Social Sciences*, Vol. 15. New York: Macmillan.

Sewell, W. H., Jr. (1982) Occupational status in nineteenth-century French urban society. In R. M. Hauser, D. Mechanic, and A. O. Haller, eds., *Social Structure and Behavior*. New York: Academic Press.

Shepard, R. N., Romney, A. K., and Nerlove, S. B. (1972) Multidimensional Scaling. Vol. 1. New York: Seminar Press.

References

Sherif, M., and Hovland, C. I. (1961) *Social Judgment.* New Haven: Yale University Press.

Shye, S., ed. (1978) *Theory Construction and Data Analysis in the Behavioral Sciences.* San Francisco: Jossey-Bass.

Sieber, S. D. (1973) The integration of fieldwork and survey methods. *American Journal of Sociology* 78:1335-1359.

Sigelman, L. (1981) Question-order effects on presidential popularity. *Public Opinion Quarterly* 45:199-207.

Silk, A. J. (1969) Preference and perception measures in new product development: An exposition and review. *The Industrial Management Review* 11:21-37.

Simon, G. (1974) Alternative analyses for the singly-ordered contingency table. *Journal of the American Statistical Association* 69:971-976.

Singer, E. (1978) Informed consent: Consequences for response rate and response quality in social surveys. *American Sociological Review* 43:144-162.

Singer, E. (1979) Telephone interviewing as a black box: A discussion of response styles in telephone and household interviewing. Unpublished ms., Center for the Social Sciences, Columbia University.

Skvoretz, J., Windell, P., and Farraro, T. J. (1974) Luce's axiom and occupational prestige: Test of a measurement model. *Journal of Mathematical Sociology* 3: 147-162.

Slovic, P., Fischhoff, B., and Lichtenstein, S. (1979) Rating the risks. *Environment* 21:14-20, 36-39.

Slovic, P., Fischhoff, B., and Lichtenstein, S. (1980) Facts and fears: Understanding perceived risk. In R. C. Schwing and W. A. Albers, Jr., eds., *Societal Risk Assessment.* New York: Plenum.

Smith, M. B., Bruner, J. S., and White, R. W. (1956) *Opinions and Personality.* New York: Wiley.

Smith, T. W. (1978) In search of house effects: A comparison of responses to various questions by different survey organizations. *Public Opinion Quarterly* 42: 443-463.

Smith, T. W. (1979) Happiness: Time trends, seasonal variations, intersurvey differences, and other mysteries. *Social Psychology Quarterly* 42:18-30.

Smith, T. W. (1980a) The 75% solution: An analysis of the structure of attitudes on gun control, 1959-1977. *Journal of Criminal Law and Criminology* 71:300-316.

Smith, T. W. (1980b/1982) House effects: A comparison of the 1980 General Social Survey and the 1980 American National Election Study. GSS technical paper no. 23, Chicago: National Opinion Research Center. [Revised version (1982): *Public Opinion Quarterly* 46: 54-68.]

Smith, T. W. (1981) Can we have confidence in confidence? Revisited. In D. Johnston, ed., *Measurement of Subjective Phenomena.* Washington, D.C.: U.S. Government Printing Office.

Snyder, M. (1982) When believing means doing: Creating links between attitudes and behavior. In M. P. Zanna, T. Higgins, and C. P. Herman, eds., *Consistency in Social Behavior: The Ontario Symposium*, Vol. 2. Hillsdale, N.J.: Erlbaum.

Snyder, M., and Swann, W., Jr. (1976) When actions reflect attitudes: The politics of impression management. *Journal of Personality and Social Psychology* 34: 1034-1042.

Social Research Association (1980) *Terms and Conditions of Social Research Funding in Britain*. London: Social Research Association.

Southwick, J. C. (n.d.) *Survey Data for Trend Analysis*. Williamstown, Mass.: Roper Public Opinion Research Center.

Spaeth, M. (1979) Telephone interviewing facilities at survey research organisations. *Survey Research* 11:21-25.

Speier, H. (1949) Historical development of public opinion. *American Journal of Sociology* 55:376-388.

Spence, J. (1973) *Political Opinion Polls*. London: Hutchinson University Library.

Starr, C. (1969) Social benefit versus technological risk. *Science* 165:1232-1238.

Starr, J. (1978) Mass or public? The American people as revealed through polls. Paper presented at the annual meeting of the American Sociological Association, San Francisco, September.

Steeh, C. G. (1981) Trends in nonresponse rates, 1952–1979. *Public Opinion Quarterly* 45:40-57.

Steele, R. (1978) American popular opinion and the war against Germany: The issue of negotiated peace—1942. *Journal of American History* 65:704-723.

Stefansson, V. (1936) *Adventures in Error*. New York: R. M. McBride.

Stefflre, V. J. (1965) Simulation of people's behavior toward new objects and events. *American Behavioral Scientist* 8(9):12-15.

Stefflre, V. J. (1968) Market structure studies: New products for old markets and new markets (foreign) for old products. In F. M. Bass, C. W. King, and E. A. Pessemier, eds., *Applications of the Sciences in Marketing Management*. New York: Wiley.

Stefflre, V. J. (1972) Some applications of multidimensional scaling to social science problems. In A. K. Romney, R. N. Shepard, and S. B. Nerlove, eds., *Multidimensional Scaling*, vol. 2. New York: Seminar Press.

Stefflre, V. J. (1979) New products: Organizational and technical problems and opportunities. In A. D. Shocker, ed., *Analytic Approaches to Product and Marketing Planning*. Cambridge, Mass.: Marketing Science Institute.

Stefflre, V. J., Reich, P., and McClaran-Stefflre, M. (1971) Some eliciting and computational procedures for descriptive semantics. In P. Kay, ed., *Explorations in Mathematical Anthropology*. Cambridge, Mass.: MIT Press.

Steiner, G. (1963) *The People Look at Television*. New York: Knopf.

References

Stember, C. (1955) The effects of field procedures on public opinion data. Ph. D. dissertation, Columbia University.

Stephan, F. (1948) History of the uses of modern sampling procedures. *Journal of the American Statistical Association* 43:12-39.

Stephan, F., and McCarthy, P. (1958) *Sampling Opinions.* New York: Wiley.

Stevens, G., and Featherman, D. L. (1981) A revised socioeconomic index of occupational status. *Social Science Research* 10:364-395.

Stevens, S. S. (1946) On the theory of scales of measurement. *Science* 103: 677-680.

Stevens, S. S. (1966) A metric for the social consensus. *Science* 151:530-541.

Stinchcombe, A. L. (1964) *Rebellion in a High School.* Chicago: Quadrangle.

Stinchcombe, A. L., Adams, R., Heimer, C. A., Scheppele, K. L., Smith, T. W., and Taylor, D. G. (1980) *Crime and Punishment—Changing Attitudes in America.* San Francisco: Jossey-Bass.

Stinchcombe, A. L., Jones, C., and Sheatsley, P. (1980) Non-response bias for attitude questions. Unpublished manuscript, Department of Sociology, University of Arizona.

Storer, N. W. (1969) *Survey Applications of Social Psychological Questions.* Working Paper No. 29. Washington, D.C.: Bureau of the Census.

Stouffer, S. A. (1954) 1665 and 1954. *Public Opinion Quarterly* 18:233-238.

Stouffer, S. A. ([1955]1963) *Communism, Conformity, and Civil Liberties.* Reprint. Gloucester, Mass.: Peter-Smith.

Stouffer, S. A., and DeVinney, L. (1949) How personal adjustment varied in the Army—by background characteristics of the soldiers. In S.A. Stouffer, E. Suchman, L. DeVinney, S. Star, and R. Williams. *The American Soldier: Adjustment During Army Life.* Princeton, N.J.: Princeton University Press.

Stouffer, S. A., and others (1949-1950) *Studies in Social Psychology in World War II.* 4 vols. Princeton, N.J.: Princeton University Press.

Sudman, S., and Bradburn, N. M. (1974) *Response Effects in Surveys.* Chicago: Aldine.

Sullivan, J. L., Piereson, J. E., and Marcus, G. E. (1978) Ideological constraint in the mass public: A methodological critique and some new findings. *American Journal of Political Science* 22:233-249.

Sullivan, J. L., Piereson, J. E., Marcus, G. E., and Feldman, S. (1979) The more things change, the more they stay the same: The stability of mass belief systems. *American Journal of Political Science* 23:176-186.

Sulzberger, A., Jr. (1979) Polling association adopts public disclosure code. *New York Times,* December 18.

Survey Research Center (1976) *Interviewer's Manual.* Rev. ed. Ann Arbor: Institute for Social Research, University of Michigan.

Sussman, B. (1978) Jury is still out on Carter: Distortion in popularity polls. *Washington Post* February 12:A-1.

Sutcliffe, J. P. (1965) A probability model for errors of classification: I. General considerations. *Psychometrika* 30:73-96.

Swafford, M. (1980) Three parametric techniques for contingency table analysis: A nontechnical commentary. *American Sociological Review* 45:664-690.

Taietz, P. (1962) Conflicting group norms and the "third" person in the interview. *American Journal of Sociology* 68:97-104.

Tanur, J. (1981) Advances in methods for large scale social surveys. Paper prepared for the Committee on Basic Research in the Behavioral and Social Sciences, National Research Council, and the Social Science Research Council, Washington, D.C.

Tarshis, B. (1979) *The "Average American" Book.* New York: Atheneum/SMI.

Taylor, D. G., and Smith, T. W. (1979) Evaluation of the June, 1978, and January, 1979, Milwaukee Surveys. Unpublished report to American Federation of State, County, and Municipal Employees, Chicago, Ill.: NORC, April.

Taylor, S. E., and Crocker, J. (1981) Schematic bases of social information processing. In E. T. Higgins, C. P. Herman, and M. P. Zanna, eds., *Social Cognition: The Ontario Symposium*, Vol. 1. Hillsdale, N.J.: Erlbaum.

Tesser, A. (1978) Self-generated attitude change. In L. Berkowitz, ed., *Advances in Experimental Social Psychology*, Vol. 11. New York: Academic Press.

Tesser, A., and Leone, C. (1977) Cognitive schemas and thought as determinants of attitude change. *Journal of Experimental Social Psychology* 13:340-356.

Thomas, W. I. and Znaniecki, F. (1918) *The Polish Peasant in Europe and America.* 2 vols. Boston: R. G. Badger.

Thornton, M. (1980) Pros still argue if pollsters can affect results. *Washington Star*, January 11.

Thurstone, L. L., and Chave, E. J. (1929) *The Measurement of Attitude.* Chicago: University of Chicago Press.

Tjur, T. (1980) A connection between Rasch's item analysis model and a multiplicative Poisson model. Preprint No. 6. Copenhagen, Denmark: Institute of Mathematical Statistics, University of Copenhagen.

Toops, H. A. (1948) The use of addends in experimental control, social census, and managerial research. *Psychological Bulletin* 45:41-74.

Torgerson, W. S. (1968) Scaling. In D. Sills, ed., *International Encyclopedia of the Social Sciences*, Vol. 14. New York: Macmillan.

Traugott, M. W., and Katosh, J. P. (1979) Response validity in surveys of voting behavior. *Public Opinion Quarterly* 43:359-377.

Treiman, D. J. (1977) *Occupational Prestige in Comparative Perspective.* New York: Academic Press.

References

Triandis, H. C., and Triandis, L. M. (1965) Some studies of social distance. In I. D. Steiner and M. Fishbein, eds., *Current Studies in Social Psychology*. New York: Holt, Rinehart and Winston.

Tuchman, S., and Coffin, T. (1971) The influence of election night television broadcasts in a close election. *Public Opinion Quarterly* 35:315-326.

Tukey, J. W. (1977) *Exploratory Data Analysis*. Reading, Mass.: Addison-Wesley.

Turner, C. F. (1980) Discussion of effect of question on survey response. *Proceedings of the American Statistical Association (Section on Survey Research Methods)* 1980:43-45.

Turner, C. F. (1981a) Surveys of subjective phenomena: A working paper. In D. Johnston, ed., *Measurement of Subjective Phenomena*. Washington, D.C.: U.S. Government Printing Office.

Turner, C. F. (1981b) Patterns of disagreement: A reply to Angus Campbell. In D. Johnston, ed., *Measurement of Subjective Phenomena*. Washington, D.C.: U.S. Government Printing Office.

Turner, C. F., and Kiesler, S. B. (1981) The impact of basic research in the social sciences: The case of education. *Scientometrics* 3:177-190.

Turner, C. F., and Krauss, E. (1978) Fallible indicators of the subjective state of the nation. *American Psychologist* 33:456-470.

Tversky, A., and Kahneman, D. (1973) Availability: A heuristic for judging frequency and probability. *Cognitive Psychology* 5:207-232.

Tversky, A., and Kahneman, D. (1974) Judgment under uncertainty: Heuristics and biases. *Science* 185:1124-1131.

Tversky, A., and Kahneman, D. (1981) The framing of decisions and the psychology of choice. *Science* 211:453-458.

U.K. Central Statistical Office (1970-1981) *Social Trends* (annual publication). London: Her Majesty's Stationery Office.

U.K. Office of Population Censuses and Surveys, Social Survey Division (1973–1981) *The General Household Survey: Introductory Report* (annual publication). London: Her Majesty's Stationery Office.

Union Carbide (1980a) *The Vital Consensus: American Attitudes on Economic Growth*. New York: Union Carbide.

Union Carbide (1980b) *Adversaries or Allies? American Attitudes on Business, Government and Growth*. New York: Union Carbide.

Union Carbide (1981) *The Emerging Consensus: Public Attitudes on America's Ability to Compete in the World*. New York: Union Carbide.

United Nations (1975) *Toward a System of Social and Demographic Statistics*. Department of Economic and Social Affairs. New York: United Nations.

United States v. Ahmad et al. (1972), 347 F. Supp. 912 (M. D. Pa. 1972).

United States v. Hiss (1947), 185 F. 2d. 152 (2d Cir, 1947).

Upshaw, H. S. (1969a) Stimulus range and the judgmental unit. *Journal of Experimental Social Psychology* 5:1-11.

Upshaw, H. S. (1969b) The personal reference scale: An approach to social judgment. In L. Berkowitz, ed., *Advances in Experimental Social Psychology*, Vol. 4. New York: Academic Press.

Urban, G. L., and Hauser, J. R. (1980) *Design and Marketing of New Products.* Englewood Cliffs, N.J.: Prentice-Hall.

U.S. Department of Commerce (1977) *Social Indicators, 1976.* Office of Federal Statistical Policy and Standards and Bureau of the Census. Washington, D.C.: U.S. Government Printing Office.

Valentine, C. A., and Valentine, B. L. (1971) Missing men: A comparative study of underenumeration and related problems. Unpublished manuscript, Bureau of the Census.

Vanderbilt Television News Archive (1972-1981) *Television News Index and Abstracts.* Nashville, Tenn.: Vanderbilt Television News Archive.

Van Dusen, R. A., and Zill, N. (1975) *Basic Background Items for U.S. Household Surveys.* New York: Social Science Research Council.

Velten, E. (1968) A laboratory task for induction of mood states. *Behaviour Research and Therapy* 6:473-482.

Verba, S., and Nie, N. H. (1972) *Participation in America.* New York: Harper and Row.

Vinokur, A., Oksenberg, L., and Cannell, C. F. (1977) Effects of feedback and reinforcement on the report of health information. In NCHSR Research Report Series, *Experiments in Interviewing Techniques.* Ed. C. F. Cannell, L. Oksenberg, and J. M. Converse. DHEW Pub. No. (HRA)78-3204. Washington, D.C.: Department of Health, Education and Welfare.

Viscusi, W. K. (1979) *Employment Hazards.* Cambridge, Mass.: Harvard University Press.

Von Hoffman, N. (1980) Public opinion polls: Newspapers making their own news? *Public Opinion Quarterly* 44:572-573.

Waldrop, M. (1981) Massive neutrinos: Masters of the universe? *Science* 211:470-472.

Wallis, W. A., and Roberts, H. V. (1956) *Statistics: A New Approach.* Glencoe, Ill.: Free Press.

Wall Street Journal (1975) Editorial: Loading a question. June 2:10.

Washington Post (1980) 55% favor use of force in Iran. April 20:A-1.

Watts, W. A., and Holt, L. E. (1970) Logical relationships among beliefs and timing as factors in persuasion. *Journal of Personality and Social Psychology* 16:571-582.

References

Webb, E. J., Campbell, D. T., Schwartz, R. D., and Sechrest, L. (1966) *Unobtrusive Measures: Nonreactive Research in the Social Sciences.* Chicago: Rand McNally.

Webb, N. (1980) The democracy of opinion polls. Paper presented to European Society for Opinion Surveys and Market Research seminar on opinion polls, Bonn, Federal Republic of Germany, January.

Weiss, C. (1968) Validity of welfare mothers' interview responses. *Public Opinion Quarterly* 32:622-633.

Weiss, D. J., and Davison, M. L. (1981) Test theory and methods. *Annual Review of Psychology* 32:629-658.

Wendt, J. (1979) Canonical correlation as an exploratory technique of attitude scale construction. *Public Opinion Quarterly* 43:518-531.

Westie, F. R. (1952) Negro-white status differentials and social distance. *American Sociological Review* 17:550-558.

Wheeler, M. (1976) *Lies, Damn Lies, and Statistics.* New York: Liveright.

Wicker, A. W. (1969) Attitudes versus actions: The relationship of verbal and overt behavioral responses to attitude objects. *Journal of Social Issues* 25:41-78.

Wicklund, R. A. (1982) Self-focused attention and the validity of self-reports. In M. P. Zanna, E. T. Higgins, and C. P. Herman, eds., *Consistency in Social Behavior: The Ontario Symposium,* Vol. 2. Hillsdale, N.J.: Erlbaum.

Wicklund, R. A. and Frey, D. (1980) Self-awareness theory: When the self makes a difference. In D. M. Wegner and R. R. Vallacher, eds., *The Self in Social Psychology.* New York: Oxford University Press.

Wigdor, A. K., and Garner, W. R., eds. (1982) *Ability Testing.* 2 vols. Washington, D.C.: National Academy Press.

Wilhoit, G. C., and Weaver, D. H. (1980) *Newsroom Guide to Polls and Surveys.* Washington, D. C.: American Newspaper Publishers Association.

Williams, F., and Cantril, H. (1945) The use of interviewer rapport as a method of detecting differences between "public" and "private" opinion. *Journal of Social Psychology* 22:171-175.

Williams, J. A. (1964) Interviewer-respondent interaction: A study of bias in the information interview. *Sociometry* 27:338-352.

Wilson, R. (1975) Examples in risk-benefit analysis. *Chemtech* 6:604-607.

Wilson, W. C. (1964) Belief in capital punishment and jury performance. Unpublished manuscript. [Cited in Hovey *v.* Superior Court of Alameda County, California Supreme Court, H-1440, decision of August 28.]

Wish, E. D., Robins, L. N., Helzer, J. E., Hesselbrock, M., and Davies, D. H. (1978) Monday morning quarterbacking on limiting call-backs: Evidence from a panel study of veterans. Paper presented at the annual meeting of the American Association for Public Opinion Research, Roanoke, Virginia.

Witherspoon v. Illinois. 391 U. S. 510; 1968.

Wolfgang, M. E. (1978) National Survey of Crime Severity. Unpublished report to U.S. Law Enforcement Assistance Administration.

Woltman, H. F., Turner, A. G., and Bushery, J. M. (1980) A comparison of three mixed-mode interviewing procedures in the National Crime Survey. *Journal of the American Statistical Association* 75:534-543.

Wood, G. S. (1981) Review of "Truth in History," by O. Handlin. *Journal of Modern History* 53:84-90.

Word, C., Zanna, M., and Cooper, J. (1974) The nonverbal mediation of self-fulfilling prophecies in interracial interaction. *Journal of Experimental Social Psychology* 10:109-120.

Yalch, R. F. (1976) Pre-election interview effects on voter turnout. *Public Opinion Quarterly* 40:331-336.

Yankelovich, D. (1980) How not to read the polls. *Time*, April 28.

Yates, F. (1949) *Sampling Methods for Censuses and Surveys*. London: C. Griffin.

Youden, W. (1961) How to evaluate accuracy. *Materials Research and Standards* 1:268-271.

Youden, W., and Steiner, E., eds. (1975) *Statistical Manual of the Association of Official Analytical Chemists*. Washington, D.C.: Association of Official Analytical Chemists.

Zajonc, R. (1980) Feeling and thinking: Preferences need no inferences. *American Psychologist* 35:151-175.

Zanna, M. P., Higgins, E. T., and Herman, C. P., eds. (1982) *Consistency in Social Behavior: The Ontario Symposium*. Vol. 2. Hillsdale, N.J.: Erlbaum.

Zeisel, H. (1968) Some data on juror attitudes toward capital punishment. University of Chicago Law School, Center for Studies in Criminal Justice. Mimeo. [Cited in Hovey *v.* Superior Court of Alameda County, California Supreme Court, H-1440, decision of August 28.]

Zeisel, H. (1980) Lawmaking and public opinion research: The president and Patrick Caddell. *American Bar Foundation Research Journal* 1980:133-139.

Zeisel, H., and Diamond, S. (1976) The jury selection in the Mitchell-Stans conspiracy trial. *American Bar Foundation Research Journal* 1976:151-174.

Name Index

AAPOR. *See* American Association for Public Opinion Research
ABC, 58n23, 75, 77, 308, 331n4, 332n5, 345, 352n10
Abelson, R. P., 96, 182, 200, 232, 277, 287, 289, 291
Abrams, M., 4, 39
Acland, H., 88
A. C. Nielsen Company, 350n2
Adams, E. W., 127n10
Adler, K., 35
Adorno, T., 123
Advisory, Conciliation, and Arbitration Service (ACAS), 45, 58n21
Agnew, P. G., 19n1
Ajzen, I., 289
Allen, A., 417
Alwin, D., 39–40, 42, 43
American Anthropological Association, 37
American Association for Public Opinion Research (AAPOR), 61, 63–64, 75, 78–79, 84, 90, 91n1, 92n8, 92n10, 308, 332n6
American Bar Association, 44
American Broadcasting Company. *See* ABC
American Institute of Public Opinion (AIPO), 150–151, 155, 158, 161, 377, 378–387, 389, 392, 401. *See also* Gallup Organization
American Nuclear Energy Council, 86
American Psychological Association, 93n17
American Statistical Association (ASA), 7, 68, 309, 313
American Telephone and Telegraph. *See* AT & T
Anastasi, A., 119
Andersen, K., 288
Andersen, R., 99, 117
Anderson, T. W., 3
Andrews, F., 5, 6
Apcut, 353
Arabie, P., 118
Arbitron, 351n2
Aronson, E., 298
Arthur D. Little Survey, 346
ASI Market Research, 351n2
Associated Press, 58n20, 71, 165, 332n4, 349
Atomic Energy Commission, 209
AT & T, 348, 350, 351n8
Audits & Surveys, 332, 345

Babbie, E., 130
Back, K. W., 258
Bacon, F., 97
Baddeley, A. D., 296–297
Bailar, B., 9, 33, 55, 67–68, 96, 110, 274, 278n3, 286, 312, 320, 322, 342–343, 411
Bailey, L., 110, 274, 278n3, 321, 411
Baker, H., 404
Bakker, T., 175n4

Note. The *name* index lists all individuals and organizations referred to in this volume. References to some individuals and organizations have also been included in the *subject* index when the reference extends beyond the citation of an individual's published work or the data produced by an organization (e.g., if the organization, its procedure, or interviewing staff is discussed).

Barath, A., 271
Barton, S., 19n1
Battey, E., 19n1
Baxter, R., 73, 268
Bellows, N., 298
Belson, W. A., 251–252, 282–283
Bem, D., 417
Bender, I., 420
Beniger, J. R., 53, 124, 240, 245, 325, 330
Berelson, B., 256n1
Berk, R., 44
Bershad, M., 108–109
Better Business Bureau, 73
Biderman, A. D., 58n18, 258, 296, 298
Bills, D. B., 193
Binet, A., 118
Bishop, G. F., 288, 312
Bishop, Y. M. M., 323
Black, M., 253
Blair, E., 271
Blumer, H., 13, 237
Boffey, P. M., 37
Bogardus, E. S., 188–190, 193
Bolinger, D., 284
Bonafede, D., 52
Bonora, B., 44
Boorstin, D., 330
Booth, C., 25
Booz, Allen & Hamilton, 342, 346–347
Bordua, D. J., 300
Boseman, A. T., 347
Bowen, J., 30, 350
Bower, G., 285
Bowers, D., 66
Braaten, K., 275
Bradburn, N. M., 13, 87, 113, 116, 137, 252, 271, 320
Bradley, G. W., 420
Brezhnev, L. I., 57n15
Brody, C., 200, 202, 207
Brody, R. A., 51, 145
Bronk, D., 19n1
Brook, L., 164–165
Brooks, C. A., 67
Brown, J., 404
Browne-Wilkinson, Justice, 45, 58n21
Bruner, J. S., 238–239
Bryce, Lord J., 240
Bryson, M. C., 54n1
Bureau of Applied Social Research, 150, 152
Bureau of Justice Statistics, 185
Bureau of the Census, 4, 10, 29–32, 57n16, 69, 87, 92n4, 108–110, 114, 115, 138, 150, 153, 156, 160, 185, 259–261, 269–270, 277n2, 278n6, 281, 311–312, 324, 338–339, 343, 347, 348–349, 351n4, 351n6, 352n11, 369, 400, 405
Burke International Research Corporation, 345
Bush, G., 404
Bushery, J. M., 277n1
Business Roundtable, 78
Butler, E. W., 110–111

Caddell, P., 52, 312
California Poll, 332n4
Cambridge Reports, 92n19
Cambridge Survey Research (CSR), 70, 91n3, 332n4
Campbell, A., 5, 6, 10, 12, 20n8, 142, 312, 324
Campbell, A. A., 134
Campbell, B. A., 271
Campbell, D. T., 117, 213
Campbell, N., 127n13
Canetti, E., 23
Cannell, C. F., 175n4, 258, 271, 275, 281–282, 295–297
Cannon, L. W., 196
Cantril, A. H., 19n1, 145–146, 154–155, 157–158, 174, 252, 276, 307, 322, 377
Carlsson, G., 325
Carmines, E. G., 288
Carroll, J. D., 118, 227n1
Carter, J., 46, 51, 52, 75, 314, 358, 404
CASRO. *See* Council of American Survey Research Organizations
CBS, 58n23, 82, 83–84, 92n3, 165, 323, 331n4, 332n5, 348–349
Census Bureau. *See* Bureau of the Census
Center for Coordination of Research on Social Indicators, 318
Center for Disease Control, 21n11
Center for Political Studies, 327
Center for Survey Research, 347
Chilton Research Service, 342, 345, 351n8
Chisolm, D., 5
Cicourel, A. V., 256n2
Clausen, A., 286

Clogg, C. C., 19n4, 93n16, 198–200, 203, 206–207, 225
Cochran, W. G., 19n1, 320
Cohen, S., 153–154, 176n20
Coleman, J. S., 127n10, 182
Collins, M., 164–165
Collins, S., 43–44, 58n19
Columbia Broadcasting System. *See* CBS
Combs, B., 416
Committee on National Statistics, 30, 153, 352n11, 400, 405
Conference Board, 331n4
Conklin, H., 253
Connally, J., 404
Connor, W. D., 38
Consumer Federation of America, 78
Consumer Product Safety Commission, 78
Consumer Protection Agency (CPA), 78, 92n10
Converse, J. M., 27, 258, 262–266, 268–269, 271–273, 275–277, 278n6, 282, 295, 300n1
Converse, P. E., 6, 10, 12, 20n8, 27, 29, 51, 52, 59n26, 144–146, 199, 236, 241, 288, 312, 324, 414
Cook, L., 128n14
Cook, S. W., 188
Coombs, C. H., 127n13
Cooper, A. D., 37
Cooper, J., 272
Council of American Survey Research Organizations (CASRO), 61, 64–66, 84, 91n2, 308, 332n6
Cowan, C. D., 44, 138, 160–161, 176n19, 277n1
Crain, R. L., 88
Crane, H., 1
Crocker, J., 291
Cronbach, L. J., 119, 224–225
Cross, T. S., 258
Crossley, A., 19n1, 314
Current Population Survey (CPS), 259–261, 265–266, 269, 273, 278n3, 281–282, 286, 343, 400, 432n1

Data Development Corporation, 351n2
Davis, D., 240
Davis, J. A., 5, 242
Davison, M. L., 229n12
Dawes, R. M., 127n13
Deci, E. L., 278n10
Decisions Center, 351n2
DeLuca, D., 70
DeMaio, T. J., 24, 268, 336, 420
Deming, W. E., 19n1, 97, 112, 320
Demographic Surveys Division, 351n4
Department of Defense, 37
Department of Energy, 35
Department of Health and Human Services, 35
DeVinney, L., 141
Dewey, T. E., 314, 383–385, 387–388
Dexter, L. A., 244
Diamond, S., 44
DiNubile, M., 292
Dionne, E. J., Jr., 47, 307, 332n7, 349
Direct Mail/Marketing Association, 73
Division of Program Surveys, 29
Dodd, S., 61–4, 332n6
Drew, E., 51
Duncan, B., 137, 266–267, 273
Duncan, B. L., 410
Duncan, O. D., 2, 24, 93n16, 96, 137, 145, 168, 169, 180, 192–193, 195, 200, 202–203, 205, 207–209, 212, 223, 227n1, 228n7, 251, 252, 266–267, 273, 285, 327, 337–338, 421n1
Durkheim, E., 12, 20n8, 185
Dutka, S., 55n4, 113

Easterlin, R., 5, 195
Eckler, A. R., 19n1
EEC. *See* European Economic Community
Ehrhardt Babic Associates, 351n2
Eisenberg, D., 16
Eisenhart, C., 97, 112
Elder, R. F., 19n1
Eliot, T. S., 231
Ellsworth, P., 44
Elrick & Lavidge, 346
Equifax, 346
Erbring, L., 330, 355
Erdelyi, M. H., 415
Ericsson, K. A., 13, 299, 412
Erikson, R., 38
European Economic Community (EEC), 38, 57n17
Executive Office of the President, 4, 29, 67, 332n6, 369

Farraro, T. J., 185
Fay, R. F., 227n1
Featherman, D. L., 192
Federal Trade Commission (FTC), 44–45
Fellegi, I. P., 227n1
Ferber, R., 7, 116, 275–276
Fernandez, E. W., 25
Ferreira, J., 249
Field, M., 332n4
Fields, J., 420
Fienberg, S. E., 9, 212, 223, 323
Fischer, E., 267–268
Fischhoff, B., 2, 232, 246–247, 291, 339
Fishbein, M., 289
Fisher, G., 175n4
Ford, G., 75
Fortune, 154, 377, 379–383, 388
Fowler, F. J., 282, 297
Frake, C., 253
Frankel, J., 315
Frankel, L. R., 55n4, 57n15, 96, 113, 332n9, 340
Franks, F., 16
Franzen, R., 19n1
Freedman, R., 31
Freeman, J. L., 110–111, 207
Frey, D., 299
Funkhouser, G. R., 53

Gallup, G., 19n1, 62, 146, 307, 312, 330
Gallup Organization (includes Gallup Poll), 11–12, 29, 47, 54, 66, 70, 133, 165, 195, 314, 327–328, 331n4, 347, 357, 361, 364. *See also* American Institute of Public Opinion
Gannet Co., 346
Garner, W. R., 119, 121
Gaudet, H., 256n1
General Accounting Office (GAO), 14, 32, 56n9, 87, 92n5, 369–370, 370n2
General Foods, 348
General Social Survey (GSS), 217, 219, 328, 392–395, 400, 401–405
George Fine Organization, 167
German Armed Forces Institute for Social Research, 37
Gibson, C. O., 138, 160, 265
Gilbert, J., 88
Gimbel, F. M., 312
Ginsberg, G. P., 275
Gitelman, Z. Y., 38
Glaser, B., 252
Glasser, G., 111
Gleser, G. C., 119
Glock, C., 39
Goffman, E., 291
Goldberg, D., 31
Goldberg, F., 44
Goldberg, L., 420
Goldberger, A. S., 227n1
Goldthorpe, J. H., 185
Gollin, A., 237, 330
Gonzales, M., 61, 67, 311
Goodman, L. A., 9, 169, 173, 198–201, 203, 206–207, 212, 218, 223, 225, 227n1, 323
Greenberg, D., 331n3
Groves, R., 55n4, 137, 277n1, 320
Gruenberg, B., 193–194
GSS. *See* General Social Survey
Guest, L., 134
Gulliksen, H., 119
Guttman, L., 144, 187–188, 198–199, 201, 211, 220, 227n1

Haberman, S. J., 200, 203, 206, 212, 223, 227n1, 323
Haller, A. O., 193
Hambleton, R., 128n14
Hamblin, R. L., 190
Hampapa, P., 226
Handlin, O., 125n3
Hansen, M., 19n1, 108–109, 320
Harman, H., 227n1
Harris Data Center. *See* Louis Harris Data Center
Harris, L., 53
Harris Survey. *See* Louis Harris and Associates
Harrop, M., 39
Hart, C., 19n1, 27
Hartigan, J. A., 254
Harvey, O., 242
Hastorf, A., 420
Hatchett, S., 136, 271
Hauser, P., 19n1
Hausser, D. L., 281
Heeringa, S., 312
Heisenberg, W., 97
Hennessy, M., 44
Henry, N. W., 225, 228n9
Henson, R., 275

Herman, C. P., 276
Hershey, J. C., 291
Higgins, E. T., 276
Hiss, A., 58n19
Hitch, G. J., 296
Ho, C. Y., 15
Hodge, R. W., 184
Holland, P. W., 323
Holmberg, S., 86
Holmes, D. S., 420–421
Holt, L. E., 290
Honomichl, J., 342, 347, 350n1
Hope, K., 185
Horvitz, D., 153
House, J. S., 268
Hovland, C. I., 19n1, 242, 292
Humphrey, H., 256
Hunter, J. S., 15, 17, 21n11, 21n12
Hurwitz, W., 108–109, 320
Hussey, G. F., 19n1
Hyman, H. H., 3, 136, 140, 258, 263, 271–272, 276–277, 320, 322, 327

IMS International, 351n2
Information and Education Division of the War Department, 27
Institute for Social Research (ISR), 41, 312, 344
Institute for Survey Research, 344
Inter-University Consortium for Political and Social Research (ICPSR), 42, 70, 327
Issacs, S., 37

Jackman, M., 267
Jackson, D., 16, 20n10
James, W., 412
Jensen, R., 56n14
Jessen, R., 19n1
Johnson, M. P., 289
Jones, C., 20n8
Jones, E. E., 420
Jöreskog, K. G., 198, 227n1
Jurow, G., 44

Kahn, R., 55n4, 137, 277n1, 320
Kahneman, D., 291, 299, 416
Kalton, G., 111, 164–165, 252
Kaplan, A., 14
Karatnycky, A., 52
Katosh, J. P., 286
Katz, D., 238
Katz, E., 256n1
Kelman, H. C., 238
Kempf, W. F., 226, 229n12
Kennedy, J. F., 425
Kennedy, T., 404
Kiesler, C. A., 286
Kiesler, S. B., 5
King, L., 294
Kinsey, A. C., 100
Kinsley, M., 92n11
Kish, L., 110, 322
Klapper, J. T., 240
Klockars, A., 294
Koch, G. G., 109, 207
Kohn, M. L., 182–183, 228n5
Koopmans, T., 19n1
Kornhauser, A., 80
Kramer, J., 38
Krantz, D. H., 127n13
Kraus, S., 240
Krauss, E., 5, 44, 136, 156, 164–165, 285, 327, 395, 403
Kraut, R. E., 286
Kriss, M., 284
Kruskal, J. B., 96, 126n6, 203, 227n1, 254
Kunreuther, H., 326

Ladd, E. C., Jr., 7
Landon, A. M., 54n1
Lane, R. E., 239
Lang, S., 7, 20n5, 244, 312
Lanouette, W., 77, 323
Lanphier, M., 68, 312
Lansing, J. B., 275
La Porte, T., 5
Lasswell, H. D., 238
Laumann, E. O., 188
Laurent, A., 296
Lawson, S. A., 275, 281
Lazarsfeld, P. F., 19n1, 27, 151, 190, 199, 225, 228n9, 252, 256, 256n1, 333n13
League of Women Voters, 53
Lee, W., 16, 20n10
Leggett, J., 267
Lehnen, R., 207
Lenski, G., 267
Leone, C., 290–291
Lepkowski, J., 297
Le Play, P. G. F., 25
Lessler, J. T., 322
Levine, D. B., 31, 339

Lewis, V., 297
Lichtenstein, S., 246–247
Lide, D. R., Jr., 15
Lieberson, S., 193
Light, R., 88
Likert, R., 19n1, 27, 187
Li Ming, 57n15
Lindzey, G., 134
Link, R., 54n1
Linton, M., 296
Lipset, S. M., 3, 7
Literary Digest, 26–27, 91
Litton Industries, 45
Liu, C., 16, 20n10
Lizotte, A. J., 300
Lockwood, B., 151
Lodge, M., 186
Loevinger, J., 211
Loftus, E. F., 283–285, 297
Lord, F. M., 128n14
Los Angeles Times, 332n4, 349
Louis Harris and Associates, 47, 66, 158, 165, 209, 285, 346, 357, 361, 364, 395, 421n1
Louis Harris Data Center, 42, 70–71, 263, 312, 327–328, 331n4, 343
Lowell, A. L., 242
Lowrance, W. W., 248
Lucas, D. B., 19n1
Luce, R. D., 127n13, 185
Lynch, J. P., 296

MacBride, S., 51
Mach, G., 226
Machlup, F., 329
MacKuen, M., 51, 53, 124–125, 158, 165, 232, 240, 245, 325, 330, 332n7
Mac Lane, S., 5
Madansky, A., 200
Madow, W., 320
Mahalanobis, P. C., 110
Mandelbrodt, B. B., 104
Mannheim, K., 329
Marburger, W., 297
Market and Opinion Research International (U.K.), 332n4
Market Facts, 342, 345
Marketing and Research Counselors, 345
Market Opinion Research, 342
Market Research Society, 350
Marks, L. E., 191
Marquis, K. H., 282, 296–297
Marsh, C., 24, 44, 53, 58n18, 77, 330, 336
Martin, E., 267, 295, 327, 336
Martin, E. G., 37, 70, 232, 240, 242
Mauser, G. A., 253, 255
McCarthy, J. R., 52
McCarthy, P., 3
McClaran-Stefflre, M., 253
McConahay, J. B., 44, 128n16, 286
McDonald, R., 347
McDuffee, D., 327
McFarland, S., 136
McGovern, G. S., 422–423
McGuire, W. J., 290, 293, 415
McKennell, A. C., 135, 252
McKenney, N. R., 251
McNemar, Q., 13, 95, 187, 229n15
Mead, G. H., 20n8
Meeks, J. G. T., 17
Meier, H., 312
Metzger, D., 253
Meyers, S., 153–154
Miaoulis, G., 111
Miles, C., 13
Miller, A., 312
Miller, P. V., 297
Miller, T. W., 226
Miller, W. E., 312
Mills, J., 298
Minsky, M., 291
Mitchell, R. C., 86, 323
Mitofsky, W. J., 314, 349
Moore, J. C., 298
Moore, T., 110, 274, 322, 411
Moser, C. A., 252
Mosteller, F., 27, 88, 314, 323
Mueller, J. E., 133
Murphy, J. M., 37, 138, 160–161, 176n19
Murphy, L. R., 277
Myers, D. G., 421
Myers, V., 252

NAACP (National Association for the Advancement of Colored People), 43
National Academy of Sciences, 36. *See also* National Research Council
National Analysts, 342, 346
National Broadcasting Company. *See* NBC
National Bureau of Standards (NBS), 16
National Center for Health Services Research, 154

National Center for Health Statistics, 154
National Council of Public Polls (NCPP), 53, 61, 66–67, 69, 72, 84, 91n3, 92n3, 308, 314, 328, 332n6
National Education Association, 332n4
National Family Opinion, 331n4, 346
National Opinion Research Center (NORC), 6, 20n7, 29, 39, 41, 84, 88, 119, 135, 140, 143, 150–154, 158, 161, 167–168, 169, 170–173, 176n21, 176n22, 177n22, 183–184, 227n1, 228n8, 259–261, 266, 269–270, 273, 277n2, 280, 312, 316, 327, 344, 377, 383, 389, 395–400, 401–405, 421n1
National Research Council (NRC), 3, 5, 30, 87, 153, 176n15, 269–270, 278n5, 338, 352n11, 411. *See also* National Academy of Sciences
National Science Board, 4–5, 19n3
National Science Foundation, 20n7
NBC, 46, 58n23, 91n3, 92n3, 165, 332n4, 349, 351n9
NCPP. *See* National Council of Public Polls
N.E. Field Facts, 342
Nerlove, S. B., 227n1, 232
Neuman, W. R., 241
Newman, S. J., 9, 33, 55n8
Newton, I., 148
New York Times, 82, 83–84, 92n4, 165, 323, 331n4, 332n4, 332n5, 348–349
Neyman, J., 320
Nicholls, W., 347
Nie, N. H., 203, 241, 243, 288, 312
Nimmo-Smith, M. I., 297
Nisbett, R. E., 13, 298–299, 412, 420
Nixon, R., 256, 266, 327, 422–423, 429
Noelle-Neumann, E., 29
NORC. *See* National Opinion Research Center
North American Newspaper Alliance (NANA), 82, 83–84, 323
North Atlantic Treaty Organization, 35
NPD Research, 351n2
NRC. *See* National Research Council
Nunnally, J. C., 130

OECD (Organization for Economic Cooperation and Development), 38, 57n16
Office of Consumer Affairs, 78
Office of Management and Budget. *See* OMB
Office of Public Opinion Research (OPOR), 154, 155, 377, 378, 383–389
Office of State Budget and Management (North Carolina), 55n5
Oksenberg, L., 258, 275, 282, 295, 297
Oldendick, R. W., 288
Oliver, J., 153–154
OMB (Office of Management and Budget), 32–33, 34, 35, 36, 56n9, 56n13, 67, 69, 332n6, 369–370
O'Neill, H. W., 86
Opinion Research and Communication (ORAC), 80, 81
Opinion Research Centre of U.K., 332n4
Opinion Research Corporation (ORC), 84, 171, 172–173, 268, 331, 342, 346
Organization for Economic Cooperation and Development. *See* OECD
Owens, M., 296, 411

Padawer-Singer, A., 293
Page, B. I., 51
Paletz, D., 309, 331n4, 332n4, 367n2
Palmer, J. C., 283–284
Parsons, T., 11
Parten, M., 26, 29, 54n1
Payne, G., 39
Payne, S., 131–132, 146
Penick, B. K., 296, 411
Pepitone, A., 292
Perline, R., 211
Petrocik, J. R., 203, 241, 288
Pettigrew, T. F., 272
Pfanzagl, J., 127n10
Piazza, T., 144
Pilate, P., 97–98
Pineo, P. C., 189–190
Politz, A., 19n1
Poulton, E. C., 292
President's Commission on a National Agenda for the Eighties, 5
Presser, S., 13, 39–40, 41, 70–71, 72, 74, 131–135, 140, 142, 146, 175n2, 243, 267, 293, 312, 320, 323, 327, 413
Press Intelligence, 353
Price, D. O., 272
Pritzker, L., 108
Pryor, J. B., 284, 299

Rabjohn, J. N., 288, 312
Rae, S., 62, 330
Rainwater, L., 11, 20n8, 195
Rand Corporation, 344
Rao, J. N. K., 227n1
Rasch, G., 106, 128n14, 180, 185, 198, 210–213, 215–218, 220–226, 227n2, 228n9, 229n13, 229n17
Reagan, R., 52, 75, 77, 314, 357, 404
Reeves, T., 46
Reich, P., 253
Reiss, A. J., Jr., 14, 183–184
Reliance Group, 346
Repp, B. H., 226, 229n12
Republican National Committee, 75
Research Triangle Institute (RTI), 150, 153–154, 344
Roberts, F., 127n10
Roberts, H. V., 54n1
Robinson, A., 71
Robinson, C. E., 25
Rockefeller Foundation, 3
Rodgers, W. L., 6, 10
Roger Seasonwein and Associates, 70, 92n12, 371
Romney, A. K., 227n1
Roosevelt, F. D., 52, 54n1, 155, 379–383, 385, 387–388
Roper, B. W., 58n20, 107
Roper, E., 19n1, 43, 131, 152
Roper Organization, Inc., 165, 351n2
Roper Public Opinion Research Center, 42, 66, 70, 73, 150–152, 158, 268, 327, 331n4, 367n1, 393–394, 402
Rose, A. M., 188–189
Rosenberg, M. J., 52, 59n26, 181, 227n1, 228n7, 333n13
Ross, L., 420
Rossi, P. H., 184, 190–191
Rothwell, N. D., 9, 33, 55n8, 278n6, 282, 343, 411
Royal Statistical Society, 314
RTI. *See* Research Triangle Institute
Rugg, W. D., 131, 132, 145–146, 268
Rustemeyer, A., 281
Ruz, J. E., 325

Sabato, L. J., 52
Safarov, R., 56n15
Sagar, H. A., 411
Saks, M. J., 44
Salancik, J., 299
SAMI, 351n2
Sanoff, A. P., 308
Sawyer, D. O., 198–199
Schaeffer, N. C., 271
Schanck, R. L., 257–258, 421
Schank, R. C., 291
Schofield, J. W., 411
Schönborn, M., 37
Schulman, J., 44
Schuman, H., 13, 96, 131–136, 137, 140, 142, 146, 161, 175n2, 193, 194, 203, 205, 243, 251, 252, 262–269, 271–273, 276–277, 278n6, 285, 289, 293, 320, 323, 327, 332, 340, 413, 420
Schwartz, S., 289
Scitovsky, T., 8
Scott, A. J., 227n1
Searles, R., 272
Sears, D. O., 128n16, 288
Seasonwein. *See* Roger Seasonwein and Associates
Selltiz, C., 188
Seltzer, R., 44
Selvin, H. C., 227n1
Sewell, W. H., Jr., 193
Sharp, H., 31
Sharp, L. M., 58n19, 315
Sheatsley, P., 20n8, 140
Shepard, R. N., 227n1
Sherif, M., 242, 292
Shewhart, W., 19n1
Shoemaker, P. J. H., 291
Shye, S., 227n1
Sibley, E., 19n1
Sieber, S. D., 252
Siegel, P. M., 184
Siegelman, L., 136
Simon, G., 203
Simon, H. A., 13, 299, 412
Singer, E., 137, 269
Skvoretz, J., 185
Slesin, L., 249
Sloane, D. M., 145, 200, 202
Slovic, P., 246–247, 416
Smith, M. B., 238–239
Smith, T. W., 5, 9, 55n8, 96, 131, 136, 143, 146, 158, 164–165, 176n18, 203, 241, 285, 288, 327, 332n9, 336, 340, 391, 392, 394–396, 398–400, 403

Snyder, M., 289
Social Research Association, 58n18
Social Science Research Council (SSRC), 3, 27, 31, 42, 176, 314, 318
Social Security Administration, 35, 55n7
Sörbom, D., 227n1
Southwick, J. C., 29
Spaeth, M., 342, 347, 351n3
Speier, H., 238
SRC. *See* Survey Research Center
SRI International, 345, 347
Stanton, F., 19n1
Starch, INRA, Hooper, 351n2
Starr, C., 248
Statistics Canada, 108, 110
Steele, R., 51
Stefflre, V. J., 253
Steiner, E., 16, 319
Steiner, G., 152
Steinhaus, H., 104
Stember, C., 151, 175n9
Stephan, F., 3, 19n1, 320
Stephens, S., 39–40, 42, 43
Stevens, G., 192
Stevens, J., 278n3
Stevens, S. S., 118, 184–186
Stimson, J. A., 288
Stinchcombe, A. L., 20n8, 252
Stock, J. S., 19n1
Storer, N. W., 316–317, 337
Stouffer, S. A., 3, 19n1, 27, 37, 52, 141, 149–150, 175n8, 242
Stowell, R., 111
Strauss, A., 252
Sudman, S., 116, 137, 271, 275–276
Sullivan, J. L., 288, 312
Survey Research Center (SRC) (U. of California), 347
Survey Research Center (SRC) (U. of Michigan), 6, 39, 110, 135, 150, 152–153, 158, 167–168, 169, 170, 174, 175n12, 176n21, 176n22, 177n22, 257n, 259–261, 266, 268–269, 273, 275, 277n2, 278n3, 280–282, 286, 313, 338, 349, 352n11, 392–394, 396–399, 402–405, 421n1
Survey Research Laboratory, 347
Sussman, B., 332n7
Swafford, M., 228n7
Swan, J., 44
Swann, W., Jr., 289

Taietz, P., 273
Tanur, J., 8
Tarshis, B., 29
Taylor, S. E., 291
Teeter, R., 53
Tesser, A., 290–291
Thatcher, M., 48, 79
Thielens, W., Jr., 151
Thompson, W., 44
Thornton, M., 53
Thurstone, L. L., 187–190
Tjur, T., 211
Toops, H. A., 229n18
Tracey, E., 278n6
Tracy Locke Co., 345
Trades Union Congress, 48
Traugott, M. W., 286
Treiman, D. J., 184, 192–193
Triandis, H. C., 190
Triandis, L. M., 190
Truman, H. S., 314
Tuchfarber, A. J., 288
Tukey, J. W., 127n13, 323
Turner, A. G., 277
Turner, C. F., 2, 5, 6, 20n6, 24, 44, 51, 96, 125, 136, 156–158, 164–165, 171, 285, 327–328, 332n7, 336, 391, 395, 400, 403
Tversky, A., 127n13, 291, 299, 416

U.K. Central Statistical Office, 29
UNESCO, 51
Union Carbide Corporation, 70, 84–86, 92n11, 336, 371, 373, 375
United Nations, 4, 38
United States Information Agency (USIA), 35–36
Upshaw, H. S., 292–293
U.S. Department of Agriculture (USDA), 26
U.S. Department of Commerce, 4, 29, 57n16
U.S. Department of Transportation, 87

Valentine, B. L., 270
Valentine, C. A., 270
Vanderbilt Television News Archives, 59n23, 328
Van Dusen, R. A., 319
Varvra, J., 6
Velten, F., 285
Verba, S., 52, 59n26, 203, 241, 243, 288

Vinokur, A., 282
Von Hoffman, N., 51

Wainer, H., 211
Walker Research, 30, 351n2, 352n11
Wallace, G., 256
Wallis, W. A., 54n1
Warwick, C., 19n1
Washington Post, 167–168, 169, 170, 176n21, 176n22, 177n22, 332n5, 349
Wasserman, S. S., 226
Watkins, S., 325
Watts, W. A., 290
Weaver, D. H., 307
Webb, E. J., 9
Weiss, C., 268
Weiss, D. J., 229n12
Wendt, J., 144
Wertman, D., 35–36
Westat, Inc., 345
Westie, F. R., 190
Wheeler, M., 47, 91n1, 107
White, R. W., 238–239
Wicker, A. W., 289
Wicklund, R. A., 289, 299
Wiener, J., 138, 160–161, 176n19
Wigdor, A. K., 119, 121
Wilbur, W., 19n1
Wilhoit, G. C., 307
Wilks, S. S., 19n1, 154, 176n15
Williams, F., 276
Williams, G. E., 253
Williams, J. A., 136, 271
Willkie, W. L., 155, 379–382
Wilson, R., 249
Wilson, T. D., 13, 298, 412
Wilson, W. C., 44
Windell, P., 185
Wish, E. D., 262
Wish, M., 254
Withey, S., 5, 6
Wolf, S., 268
Wolfgang, M. E., 185–186
Woltman, H. F., 277n1
Wood, G. S., 125n3
Word, C., 272
Wright, B. D., 211, 228n12
Wrightsman, L. S., 188

Yalch, R. F., 286
Yankelovich, D., 66, 307
Yankelovich, Skelly and White, 331n4, 346
Yates, F., 320
York, R. L., 88
Youden, W., 16, 97, 107, 319
Yule, G. U., 206

Zajonc, R., 299
Zanna, M. P., 272, 276
Zeisel, H., 44, 326
Zentralarchiv für Empirische Sozialforschung, 42
Zill, N., 319

Subject Index

ABC (American Broadcasting Company), 58, 331n4; disputed poll conducted by, 75, 77, 308
ABC News/*Washington Post* poll, 332n5, 352n10
Abortion, 77, 133, 135, 139, 140, 199, 202, 228n8, 293, 418–419
Academic Mind survey, 150, 151–152
Academic survey organizations, 342–343
Academic surveys, 27, 39, 89, 316
A. C. Nielsen Company, 350n2
Advertising, 45, 84; and survey results, 85–86
Advisory, Conciliation, and Arbitration Service (ACAS), 45–46, 58n21
Advocacy polling, 77, 86
Age, 99, 112
Age heaping, 103, 126n6
Agenda setting, 53, 124, 240, 325, 330
Agnew, P. G., 19n1
American Association for Public Opinion Research (AAPOR), 52, 308; code of ethics and professional practice, 63–64, 78–79, 84, 90, 91n1, 92n8, 92n10, 332n6
American Bar Association, 44
American Broadcasting Company. *See* ABC
American Institute of Public Opinion. *See* Gallup Organization
American Nuclear Energy Council, 86
American Statistical Association, 7, 68, 309, 313, 332n8
American Telephone and Telegraph Company. *See* AT&T
Annual Housing Survey, 32, 348
AP/NBC News poll, 71, 165, 332n4
Arbitron Company, 351n2
Archives, 40, 41–42, 62, 226, 327, 328, 330
ASI Market Research, 351n2
Associated Press (AP), 71, 349. *See also* AP/NBC News poll
Association, patterns of, 197, 203–205; column-effects model, 205; diagonal-effects model, 205; row-effects model, 205; uniform association model, 205, 218, 229n17
AT&T (American Telephone and Telegraph Company), 348, 350, 351n8
Attention, 289, 299, 413
Attitude-behavior relation, 88, 157, 233, 243, 276, 289, 298, 299, 417
Attitude centrality, 238, 241
Attitude change, 286, 289, 298
Attitude crystallization, 145–146, 160, 202
Attitude polarization, 209, 280, 291
Attitude scaling. *See* Scaling
Attitude strength, 146, 286
Attitude structure, 123–124, 241, 287, 288, 291

Note. References to some individuals and organizations have been included in the *subject* index when the reference extends beyond the citation of an individual's published work or the data produced by an organization (e.g., if the organization, its procedures, or interviewing staff is discussed). The *name* index contains a comprehensive listing of individuals and organizations referred to in this volume.

Attitudes: motivations for, 238–239, 240–241; public vs. private, 239, 243, 257, 262, 263, 264, 274, 276, 418; symbolic, 238, 239, 241, 243, 244
Attributions, 298, 415–417, 419–421
Audits & Surveys, 332n4, 345
Auspices, 37, 153, 270, 338, 339
Authoritarian personality, 123–124, 128n16
Auxiliary data, 88, 125, 193, 241, 318, 325–329
Axiom of choice, 185

Bandwagon effects, 53
Barton, S., 19n1
Basic Background Items for U.S. Household Surveys, 319
Battey, E., 19n1
Bias, 101, 102, 105, 112, 147, 276. *See also* Coder variability; Context effects; Interviewer variability; Nonresponse; Question bias; Sample bias
Binet, A., 118
Birth expectations, 31, 157, 343
Bogardus, E. S., 188, 189
Booth, C., 25
Booz, Allen, and Hamilton, 342, 346, 347
Bronk, D., 19n1
Bureau of Applied Social Research (Columbia U.), 27, 150, 152
Bureau of Justice Statistics, 176, 185. *See also* National Crime Survey
Bureau of the Census, 4, 27, 108, 109, 110, 113, 251, 311, 343, 349, 369; census undercount, 87, 270, 278n7; lawsuit over 1980 Census, 87; policy on subjective measurements, 31–32, 339. *See also* Annual Housing Survey, Current Population Survey
Burke International Research Corporation, 345
Business, attitudes toward, 85, 372–375
Business Roundtable, 78

Cable television used for conducting surveys. *See* QUBE
Caddell, P., 52
Calibration, 16, 116, 174, 225, 226, 319
Call-in poll, 75, 77, 308
Cambridge Reports, 92n12. *See also* Cambridge Survey Research
Cambridge Survey Research, 70, 91n3, 332n4. *See also* Cambridge Reports
Campaign contributions, 53
Candidate preferences, 26, 48, 53, 120, 155, 253, 254, 255
Cantril, H., 19n1, 154, 174
Carey et al. v. *Klutznick et al.*, 87
CASRO. *See* Council of American Survey Research Organizations
Catastrophe, 249–250
Categorical data analysis. *See* Cross-classification
CATI (computer-assisted telephone interviewing), 30, 50, 55n4, 318, 319, 343
CBS (Columbia Broadcasting System), 58, 92n3, 332n5; establishment of poll by, 348–349. *See also New York Times*/CBS News survey
Census Bureau. *See* Bureau of the Census
Census of Population and Housing. *See* Bureau of the Census
Census undercount, 87, 270, 278n7
Center for Study and Evaluation of Survey Methodology, 314, 322–323
Center for Survey Research (U. of Massachusetts and Joint Center for Urban Studies, MIT and Harvard U.), 347
Chilton Research Services, 342, 345
China, Peoples Republic of. *See under* Surveys
Clippings, exploratory study of newspaper, 30, 47–50, 329–330, 353–368
Clustering, 227n1, 254, 256n1
Cochran, W. G., 19n1
Code of Standards for Survey Research. *See under* Council of American Survey Research Organizations (CASRO)
Coder variability, 107, 108, 109, 111, 134, 150, 154, 161
Codes. *See* Standards, codes of
Columbia Broadcasting System. *See* CBS
Commercial survey organizations, 342–343
Commercial surveys, 45–50, 84, 328, 348–349, 350n2
Communism, Conformity, and Civil Liberties, 52. *See also* Stouffer Survey
Comparability, 139, 140, 166, 281, 286, 412

Computer-assisted telephone interviewing. *See* CATI
Concatenation, 99, 100, 106, 127n13
Concepts: and normative criteria, 237, 238, 242, 244, 245, 249; lay, 233, 235, 237, 245–246, 250, 410–412, 417; scientific, 233, 235, 237, 245–246, 250, 412, 417. *See also* Frames of reference; Survey concepts and definitions
Conceptual ambiguity, 98, 113, 120, 122, 126n8, 138, 233, 236, 245, 248, 250, 410, 411, 417
Concorde, survey of reactions to noise from, 87
Conference Board, 331n4
Confidence, crisis of, 50, 312
Confidence in institutions or leaders, 5, 122, 158, 209, 210, 285, 322, 327
Confidentiality, 37, 65, 261, 270
Confidentiality guarantee, beliefs about, 269–270, 338
Conjoint measurement, 106, 127n13, 211
Consistency: issue, 202, 203, 287, 288; logical, 287–288, 290; situational, 257, 275–276, 293, 410, 417, 420. *See also* Response consistency
Construct validity, 106, 121, 123, 124, 147, 250, 408, 412
Consumer attitudes, 45, 46, 49, 78
Consumer Federation of America, 78
Consumer Protection Agency, disputed poll on establishment of, 78–79, 92n8–10
Content analysis, 240, 254–255, 325, 329, 330, 367n2
Content validity, 121, 122, 124
Context effects, 5, 45, 135, 139–140, 160–161, 167–173, 285, 287–295, 410, 411
Continuous Longitudinal Manpower Survey. *See* Longitudinal Manpower Survey
Contrast effects, 292; interpretive contrast, 293–294; perceptual, 292–293; response, 293; semantic, 294
Cooper-snoopers, 37
Corrective advertising, 45
Council of American Survey Research Organizatons (CASRO), 64–66, 91n2, 308; Code of Standards for Survey Research of the, 64–66, 84, 91n2, 332n6
Court cases, surveys in, 43–44, 58n21, 87; admissibility, 43; jury selection, 44; regulatory cases, 44–46
Crime: seriousness of, 185–187, 191, 292; victimization, 110, 131, 160–161, 296, 410, 411, 414, 415
Cross-classification, analysis of, 199–207, 209–210, 215, 216, 217, 323, 333n13
Crossley, A., 19n1
Cross-national comparisons, 184, 190, 192, 195
Culture. *See* Frames of reference
Current Population Survey (CPS), 26, 70, 116, 235, 421n1; costs and sample size, 343, 348, 349; errors by interviewers in, 281–282; interviewing practices, 259–261, 266, 269, 273, 277n2, 278n3; measurement artifacts in, 265 , 286; Methods Development Program of, 343; refusal rate in, 338

Data Development Corporation, 351n2
Data matrix, 181, 182
Decisions Center, 351n2
Deming, W. E., 19n1, 97
Democracy, 237–238, 239, 242, 243, 244, 330
Denial, 20n9, 431
Detroit Area Study (DAS), 202, 205, 251, 252, 262, 266, 267, 268, 272, 273
Dichotomous variables, 104–105, 203
Differential response, 191, 207–208
Direct Mail/Marketing Association, 73
Disclosure, 47, 62, 63, 65–66, 69, 71–72, 308, 309, 310, 328, 340; committee on, 66
Dispersion, 101, 102
Division of Program Surveys, U.S. Department of Agriculture, 26–27, 29
Documentation, 42, 63, 68, 135, 309, 310, 332n5
Dodd, S., survey standards proposed by, 61–64, 79, 332n6
"Don't know" (DK) responses, 134, 142, 146, 147, 150, 152, 153, 158, 161, 215, 280, 300n1

Eckler, A. R., 19n1
ECTA, 223
Education level, 69, 70, 109, 116

EEC (European Economic Community), collaborative survey programs of, 38, 57n17
Effective sample size, 107, 111
Effects of Herbicides in South Vietnam, Part B: Beliefs, Attitudes, and Behavior of Lowland Vietnamese, 37
Ehrhardt-Babic Associates, 351n2
Elder, R., 19n1
Election forecasts, 26, 53, 54n1, 255–256, 285–286, 314
Election surveys, 285–286, 314, 354; evaluation of, 314–315. *See also* Political polling
Elections, 48; 1936 presidential election, 26, 54, 314; 1948 presidential election, 27, 253, 254, 314; 1964 presidential election, 286; 1968 presidential election, 253; 1980 presidential election, 48, 53, 75, 77, 314. *See also* Presidential debates; Voting
Elite samples, 209, 244, 326
Elrick and Lavidge, 346
Emergency School Assistance Act, 88–89, 93n16
Employment Protections Act (U.K.), 45
Error, 8, 14, 198, 295; observed value, 99; random, 103–104, 145; systematic, 101–104, 112, 145, 174n1. *See also* Measurement; Validity; Reliability; True value
Error difference distributions, 100–101
Ethnic distance, 188–191, 193
Ethnic group prestige, 189–190
Ethnic segregation, 193
Ethnicity, 251
Ethnographic semantics, 236, 241, 253, 256. *See also* Respondent debriefing
Eurobarometer, 57n17
European Economic Community. *See* EEC
Evaluation. *See* Review process; Surveys, critiques and reviews
Existential beliefs, 415
Expectations, questions about, 415
Expected values, 102, 106
Experimental design, 136, 318, 323, 332n11–12. *See also* Split-ballot experiments; Interviewer variance studies
Exploratory data analysis, 323

Factor analysis, 197, 227n1
Factorial design. *See* Experimental design; Survey design
Fatigue, 264
Federal Trade Commission (FTC), 44–45
Forbid/allow wording variation. *See* Speeches against democracy
Forecasts, 9. *See also* Election forecasts
Form-resistant correlations, 141, 210, 225
Fortune poll, 154
Frames of reference: and policy issues, 245, 246, 247; in survey questions, 79, 134, 135, 183, 246, 251, 291, 300, 322, 410–413; social and cultural, 10, 12, 120, 122, 183, 184, 191, 195, 197, 236, 241, 242, 291, 411, 413; techniques for exploring, 251–256. *See also* Concepts
Franzen, R., 19n1
Free speech, 130, 202, 291, 419
FREQ, 206
F-scale, 123–124
Fundamental measurement, 127n13
Funding for surveys, 42–43, 315

Gallup, G., 19n1, 62
Gallup Organization (includes AIPO, Gallup Poll), 29, 47, 66, 70, 141, 150, 151, 161, 328, 331n4, 347; Gallup Opinion Index, 54; treatment of "Don't Know" responses by, 150–151, 161
General Accounting Office (GAO), 87; recommendation for increased policing of attitude surveys, 14; survey approval by, 32, 56, 92n5, 369–370
General Foods, 348
General Social Survey. *See under* National Opinion Research Center
George Fine Organization, 167
German Armed Forces Institute for Social Research, 37
Germany. *See under* Surveys
Government, attitude toward, 85, 338, 372–375
Government surveys, 26–27, 31, 34, 36, 55n4, 68, 338, 343, 348, 369–370; OMB review, 32, 56n9–11, 67, 69, 369; standards, 61, 67, 311, 312, 340; subjective measures in, 4, 29, 31, 33, 35, 56n13, 67, 337, 339–340; voluntary response in, 153, 261, 269, 338, 339. *See also*

Bureau of the Census; Current Population Survey; Health Interview Survey; National Crime Survey
Graphical analysis, 323
Great Britain. *See under* Surveys
Gun control, 13, 130
Guttman scaling, 187, 198–199, 211, 221, 227n1

Hansen, M., 19n1
Happiness, 5, 6, 10, 12, 19n4, 32, 157, 167–171, 195–196, 203, 207, 301n3, 419
Harris, L., 53, 165
Harris Data Center. *See* Louis Harris Data Center
Harris Survey. *See* Louis Harris and Associates
Hart, C., 19n1, 27
Hauser, P., 19n1
Health, 35, 117, 153, 258, 265, 271, 282, 296, 348
Health Interview Survey (HIS), 33, 35, 258, 265, 282, 297, 348; context effects in, 265
Heisenberg uncertainty principle, 97
House effects. *See* Intersurvey comparisons
Household Survey for Monroe County, Georgia, 35
Housing characteristics, 32, 109, 348
Hovey v. *Superior Court of Alameda County*, 44
Hovland, C. I., 19n1
Hussey, G. F., 19n1

IMS International, 351n2
Income, 11, 109, 116, 195, 278n3; satisfaction with, 11–12, 195
Independence, assumption of, 108, 214
Index of inconsistency, 108
Individuation, 277
Informed consent, 269, 338, 339
Institute for Social Research (U. of Michigan), 41, 312; size and revenues of, 344. *See also* Survey Research Center (U. of Michigan)
Institute for Survey Research (Temple U.), 344
Intelligence. *See* Mental testing
Intentions, 9, 123, 415
Interdisciplinary research, collaborative, 236, 252, 253, 256, 270, 284, 287, 298, 320, 321, 322, 324
Internal Revenue Service, 116, 370n1
Interpenetrating interviewer design, 110
Intersurvey comparisons, 5, 63, 135, 149–162, 170, 176n17, 318–319, 326, 327–328, 377–406. *See also* Measurement variability
Inter-university Consortium for Political and Social Research (ICPSR), 42, 327
Interview, 257–279, 280–283, 321–322; as citizen referendum, 80, 174n1, 263; as intimate conversation, 262–263; as meaningless exercise, 263–265; as test, 93n14, 265–268; conditioning effects of, 284–286, 287, 288, 289, 290; follow-up, 252, 258; in-depth, 112, 113, 239, 244–245, 252, 326; length of, 33, 272; telephone, 137, 140, 277n1
Interview setting, 3, 257, 259, 262, 272, 274, 276–277, 322, 413; presence of other people, 272–273
Interviewer, 111, 258–262, 263–278, 280–283, 321–322; approach to respondent, 258–262; characteristics of, 140, 267, 271, 274; errors made by, 281; expectations, 271–272, 274, 278n3; experience, 281; feedback to respondents, 281, 282, 297; influence on responses, 261, 265, 271, 282; instructions to, 266, 280; probing by, 134, 150, 161, 280, 281; rapport, 262, 268, 269, 272, 275; supervision, 111, 281; training, 87, 261, 300n1. *See also* Race-of-interviewer effect
Interviewer manuals, 259, 260, 261, 273, 277–278n2
Interviewer-respondent communication, 279, 280, 282, 283
Interviewer-respondent intimacy, 262–263, 272
Interviewer variability, 107, 108, 109, 110, 186, 274, 321–322, 411; for factual and attitude questions, 110
Interviewer variance studies, 110, 274, 321, 411
Introspection, 12–13, 283, 298–300, 412
Issue consistency. *See* Consistency, issue

Issue salience, 53
Item-centered analysis, 179–183, 187, 189, 190, 197, 207, 208, 228n5
Item characteristic curve, 220, 221
Item nonresponse, 153, 158, 268, 278n3

Japan. *See under* Surveys
Jessen, R., 19n1
Joint SSRC-NRC Committee on the Measurement of Opinion, Attitudes, and Consumer Wants, 3, 19n1–2, 176n15
Journal of Physical and Chemical Reference Data, 15, 16
Journalists' commentary: as an alternative to surveys, 17; on surveys, 52, 92n11, 307–308. *See also* Surveys, critiques and reviews of

Koopmans, T., 19n1

Labor force, 9, 265, 411
Labor unions, 44, 48, 50
Laboratory studies of survey measurement, 272, 326
Language spoken in the home, 113–116
Latent-class model, 199, 200
Latent-distance model, 225
Latent-structure models, 199–200, 225
Latent-trait model, 210–227
Lazarsfeld, P. F., 19n1, 27, 190, 333n13
League of Women Voters, 53
Le Play, family budget studies by, 25
Likert, R., 19n1, 27
LISREL, 198
Literary Digest poll, 26, 27, 54n1, 91, 314
Litton Industries, 45
Local independence postulate, 213, 225, 226
Log-linear models, 201, 203, 206, 207, 212, 227n1; hierarchical and nonhierarchical models, 201, 227n1; saturated and unsaturated models, 205
Log-multiplicative models, 206–207
Longitudinal Manpower Survey, 348, 351n5
Longitudinal surveys. *See* Panel surveys
Los Angeles Times, polling by, 332n4, 342
Louis Harris and Associates (includes Harris Survey), 47, 66, 70, 158, 285, 328, 343, 345
Louis Harris Data Center (U. of North Carolina), 42, 263, 327
Lucas, D. B., 19n1

Magnitude estimation, 186, 292
Mail solicitation, 75–76, 306, 315
Marital happiness, 167–171
Market and Opinion Research International (U.K.), 332n4
Market Facts, 342, 345
Market Opinion Research, 342
Market research. *See* Commercial surveys
Market Research Society (U.K.), 350
Marketing and Research Counselors, 345
Mass media. *See* News media; Agenda setting
McNemar's test for symmetry, 229n15
Mean square error, 101–102
Measurement, 15, 17, 105, 118, 125, 126n10, 127n11, 174, 224, 233, 306, 317, 319, 320, 325; categorical, 126n10; effect of, 285–286, 287, 288, 289, 290; intrusive, 233, 283, 285; numerical, 100, 118, 127n13; ordinal, 118; reactive, 117
Measurement and theory, 9–10, 105, 106, 123, 211–212, 233, 235, 324, 325
Measurement error, 16, 102, 103, 104, 106, 121; error models, 108, 109, 198; error profiles, 322; total survey error, 103, 322
Measurement variability: comparing measurement modes, 98, 99, 270; coordinated measurement programs, 16–17, 107, 166–174, 319, 328; reproducibility, 148–149, 161, 166, 174
Memory, 284, 296, 321, 415, 416
Mental testing, 118–122, 227. *See also* Psychometric models
Methodological research, 16, 147, 316, 317, 324; center for, 91, 314, 317, 319, 322, 323
Milwaukee Journal, poll by, 312
Minneapolis Tribune poll, 312
Misleading advertising, 45
Mode of interviewing, 137, 140, 227n1
Monetary incentives, 275–276, 278n10
Monthly Report on the Labor Force. *See* Current Population Survey
Mood, 285, 419

Motives, 9, 13, 298, 299, 300, 412. *See also* Attitudes, motivations for
Multidimensional scaling, 227n1, 254, 255
Multivariate normality, assumption of, 198

NAACP (National Association for the Advancement of Colored People), 43
National Academy of Sciences (NAS), 19n1, 19n2, 36–37. *See also* National Research Council
National Analysts, 342, 346, 347
National Assessment of Educational Progress, 122
National Broadcasting Company. *See* NBC
National Bureau of Standards (NBS), 16
National Center for Health Services Research, 154
National Center for Health Statistics (NCHS), 27, 31, 154. *See also* Health Interview Survey
National Council on Public Polls (NCPP), 53, 77, 308; standards of disclosure of, 61, 66–67, 69, 72, 84, 91n3, 314, 328, 332n6
National Crime Survey (NCS), 110, 185, 274, 297, 410, 411, 431n1; context effects in, 138, 160–161; costs and size of, 348
National Family Opinion, 331n4, 346
National Life Assurance Company Ltd. v. *Advisory, Conciliation, and Arbitration Service (ACAS), et al.*, 58n21
National Longitudinal Surveys, 348
National Medical Care Expenditure Survey, 150, 153–154
National Opinion Research Center (NORC), 27, 39, 41, 70, 88; General Social Survey (GSS), 12, 20n7, 42, 140, 182, 312, 316, 328, 421n1; interviewing practices of, 259–261, 266, 269, 270, 273, 277n2, 280–281; size and revenues of, 344; treatment of "don't know" responses by, 150–154, 158, 161
National Research Council (NRC), 3, 5, 19n1, 87, 153, 176n15, 269, 278n5, 338, 352n11, 411. *See also* National Academy of Sciences
National Science Board, 4, 5
National Science Foundation (NSF), 20n7
NATO (North Atlantic Treaty Organization), 35–36
NBC (National Broadcasting Company), 46, 58n23, 91n3; establishment of poll by, 349, 351n9. *See also* Associated Press/NBC News poll
NCPP. *See* National Council on Public Polls
N.E. Field Facts, 342
Neighborhood services, dissatisfaction with, 193, 194
Network sampling, 182, 240
News media, 46, 47, 75, 77, 86, 328, 330; coverage of polls and surveys, 47, 53–54, 59n24, 81, 86, 309, 329; in-house survey units, 30, 348–349. *See also* Agenda setting; Reporting practices
New York Times, 92n3, 165, 309; establishment of poll by, 348–349; poll reports in, 307–308, 309, 331n4
New York Times/CBS News survey, 82, 165, 323, 331n4
New York Times Information Bank, 329, 331n4, 367n2–3
Neyman, J., 26
Nonattitudes, 145, 146, 199, 203, 215, 264, 413
Nonresponse, 26, 54n1, 90, 93n18, 126n10, 153, 154, 175n12, 268, 270, 286, 315, 338; item, 153, 158, 161, 175n12, 268, 269, 278n3
Nonsampling error, 7, 20n6, 67, 69, 71, 107, 166, 291, 317, 332n11–12
NORC. *See* National Opinion Research Center
North American Newspaper Alliance (NANA), 82, 83, 323
North Carolina v. *Joan Little*, 44
North-Hatt occupational prestige scale, 183–184, 192
NPD Research, 351n2
Nuclear power, 86, 209–210

Objective phenomena, 8–9, 11, 98, 408–411; defined, 4, 8, 408
Observation error, 99, 102
Occupation, 109, 117

Occupational mobility, 206
Occupational prestige, 123, 183–185, 186, 191–193
OECD (Organization for Economic Cooperation and Development), 38, 57n16
Office of Education, 116
Office of Management and Budget. *See* OMB
Office of Public Opinion Research, 154, 155
Office of Radio Research, 27
OMB (Office of Management and Budget), 32, 33, 35, 56n9–11, n13, 67, 69; survey approval by, 67, 69, 369–370. *See also Standards and Guidelines for Federal Statistics* (Circular A-46)
Opinion Research and Communication (ORAC), 80, 81
Opinion Research Centre (U.K.), 332n4
Opinion Research Corporation (ORC), 268, 331n4, 342, 346
Order effects, 5, 45, 132, 140, 167, 171, 294–295. *See also* Context effects; Response categories; Questions
Organization for Economic Cooperation and Development. *See* OECD
Outrigger principle, 325
Overreporting, 114, 116, 117, 268, 286

Panel attrition, 286
Panel surveys, 3, 286, 326
Personality assessment, 119, 123–124
Physical measurements, 15, 16, 21n12, 97, 104, 107; interlaboratory comparisons, 16–17, 21n11, 107, 319; lead concentration, 21n11; polywater, 16; thermal conductivity of copper, 15
Pilot testing, 135, 227, 251, 252, 315–316, 319
Pluralistic ignorance, 257
Poland. *See under* Surveys
Policy and program evaluation, surveys for, 56n13, 87, 88, 245–250, 339
Political behavior, 243, 244. *See also* Voting
Political efficacy, 202
Political ideology, 82, 288
Political polling, 48, 52–53, 244, 308, 309, 314, 330
Poll. *See* Straw vote; Surveys; Political polling
Poll reports, 30, 47–50, 58n23, 69, 309, 312, 329, 330, 332n5; topics of, 48, 49. *See also* Clippings, exploratory study of newspaper
Polytomous variables, 126n10, 203–207
Polywater, 16
Powley v. *Advisory, Conciliation and Arbitration Service (ACAS)*, 58n21
Predictive validity, 13, 62, 121–122, 123, 128n15, 256, 276, 277, 289, 417
Preferences, 418. *See also* Candidate preferences; Revealed preferences
Preferred measurement procedure, 98, 112
Presidential addresses, 46
Presidential debates, 53, 75, 77, 308
Presidential popularity, 48, 50, 125
Presidential use of polls, 50, 51, 52, 330
Pretesting, 63, 67, 79, 88, 236, 252, 283
Privacy, 273, 278n3
Privacy and confidentiality survey, 150, 152–153
Probing, 134, 150, 161, 280, 281
Problems and worries, 182, 418
Production and Distribution of Knowledge in the United States, 329
Projection: complementary, 431; defensive, 431; egocentric, 430
Proprietary surveys, 46, 47, 66, 311, 328, 340
Pseudo-surveys, 73, 74–77, 268, 306, 315
Psychological measurement. *See* Personality assessment; Psychometric models; Psychophysical scaling
Psychometric models, 118–119, 120–122, 211–212; item characteristic curve, 220, 221; paired comparison method, 185, 187. *See also* Rasch measurement model
Psychophysical scaling, 186, 191, 292; magnitude estimation, 186, 292
Public attitudes vs. private attitudes, 238, 239, 240, 243, 257, 262, 263, 274, 276, 277, 418
Public opinion, 236–245, 325, 330, 418; and social communication, 202, 237, 239, 241, 245, 256n1, 288; classical tradition, 237–243, 244–245; formation of, 202, 210, 239, 240, 241, 288, 329, 330. *See also* Agenda setting

Public opinion research, 28, 29, 236–237, 244, 245, 324, 326, 328, 330

Quality control, 16, 17, 68, 174. *See also* Review process for surveys
Quality of life, 4, 5, 10, 38, 418
Quasi-facts, 9, 32, 55n8, 138, 412
Quasi-independence, 199
QUBE, 46, 77
Question bias, 7, 14, 63, 77–84, 175n3, 225, 283–286
Question meaning, 63, 119, 130, 134, 147, 156–157, 212–213, 225, 246, 251, 265, 279, 281, 283, 284, 286, 288, 291, 294, 299, 310, 322, 407, 410, 411, 413, 419
Question wording, 77, 130–134, 142, 174n1, 175n3, 280–281, 283, 284, 286, 288, 291, 299, 312, 322, 327, 419; reporting of, 65, 66, 69, 72, 73, 130, 309; substantive and nonsubstantive changes, 130–134, 283, 284, 286
Questions: active vs. passive voice, 284; agree-disagree, 142, 144; alternatives to, 8–9, 17–18; ambiguous, 111, 156, 251, 279, 282–283, 291, 413, 415; classification of, 296, 321, 407–431; compilations of, 319, 327; contiguity of, 290, 295; double-barrelled, 78; factual, 110, 111, 138, 297, 407, 408, 410–411; general vs. specific, 135, 136, 168–171, 293, 299, 415, 416–417; hypothetical, 255, 416; objectives of, 281, 282; open-ended, 134–135, 150–151, 161, 252, 281; order of, 5, 140, 167, 171; quasi-factual, 32, 55n8, 138, 411–412; referents of, 283, 408–414, 415; retrospective, 111, 138, 296–298, 299, 413, 415; rules for asking, 266, 280–283; single vs. multiple, 77–78, 82, 119, 124, 130, 133, 138, 142, 180, 225, 308; task requirements, 295–300, 321, 408, 414–419
Quota sampling, 26, 154

Race-of-interviewer effect, 136–137, 151, 267, 271, 278n8
Racial attitudes, 88, 128n16, 136, 140, 143–144, 271, 410
Rand Corporation, 344
Rapport, 262, 268, 269, 272, 275
Rasch measurement model, 185, 198, 210–227; complete analysis, 223, 224; item and subject parameters, 180, 213, 224, 225, 226; item difficulty, 213, 216, 221; separability of parameters, 224, 226; stratification, 216, 222, 223. *See also* Psychometric models
Rebuttal poll, 82, 91, 323
Recall, 283–284, 285, 296–298, 349, 415; reference period, 296–297; temporal anchor, 297
Recency bias, 296, 415
Record-check studies, 116–117, 296
Reference period, 296–297
Refusals, 153, 154, 268, 270, 315, 338
Reinterview, 113, 114, 148, 258, 282
Reliability, 101, 120–121, 143, 145, 184, 408; internal consistency method, 121; parallel forms method, 121; reliability coefficient, 120, 121; test-retest method, 105, 108, 121, 146, 148. *See also* Error, random
Replication, 16, 121, 131, 139, 140, 149–154, 161, 166, 170–171, 251, 316, 326
Reporting practices, 62, 65, 69, 90, 130, 311; in journals, 71–73; in news media, 67, 69, 71, 75, 86, 309, 329, 331n4, 332n5; selective, 84–86, 212, 340. *See also* Review process for surveys; Standards for surveys
Republican National Committee, 75
Research Branch of the Information and Education Division of the War Department, 27
Research Triangle Institute (RTI), 150, 153–154, 344
Respondent: burden, 265, 315, 341, 349; cooperation, 158, 259–261, 264, 268, 275, 308, 315, 338, 339; debriefing, 63, 242, 252, 283, 417 (*see also* Ethnographic semantics); deference to interviewer, 267; error, 295, 321, 416; expectations, 258, 261, 262, 271; fatigue, 264; knowledge, 92, 109, 156, 266–267, 272, 290–291, 413–414; motivation, 259, 260, 262, 263, 265, 275–276, 296–297; race of, 267, 272; suspicion, 268–270, 275; tasks, 275, 295, 297, 321, 408, 414; understanding, 258, 266, 279, 282, 283
Respondent-centered analysis, 180, 181, 182, 208, 222

Respondent's role, 258, 262–270, 322; sources of variability in, 273–274
Response categories, 20n4, 45, 58n21, 77–78, 79–81, 134, 156–157, 251, 252–265, 280, 292, 293, 301n2; assigning scores, 93n16, 207; collapsing, 203, 226, 251; order of, 45, 131–132, 147. *See also* "Don't know" (DK) responses
Response consistency across items, 93n16, 287–290; models of, 200–203; response consistency parameter, 201, 202. *See also* Response set
Response distortion, 270, 273, 274, 284, 308
Response rate, 68, 90, 153, 154, 268, 269, 275–276, 338; and fear of crime, 268; reporting of, 42, 68
Response set, 264; acquiescence, 124, 142, 144, 267. *See also* Response consistency
Response structure, 118, 123, 179, 180, 197–210, 287, 292. *See also* Response consistency; Rasch measurement model
Response time, 290, 295, 297
Response variability, 105, 107, 108, 110, 111, 416
Response variance, correlated and simple, 108–110
Retirement History Survey, 351n7
Revealed preferences, 8, 17
Review process for surveys, 310, 311–314, 322, 340; by OMB, 369–370. *See also* Surveys, critiques and reviews of
Risk assessment, 236, 245–256, 291, 326; acceptable risk, 247; catastrophes and, 249; voluntary risk, 248, 250
Rockefeller Foundation, 3
Roger Seasonwein and Associates, 70, 92n12, 371
Roper, E., 19n1, 43, 131, 158
Roper Organization, 66, 151–152, 158, 165, 331n4, 351n2
Roper Public Opinion Research Center, 42, 70, 327
Rotation group bias, 286
Royal Statistical Society, election poll post mortems by, 314

Salience, effects of, 53, 284, 299, 416
SAMI, 351n2
Sample: design, 26, 68, 154, 184, 256n1; elite, 209, 244, 326; expert, 63; of items, 119, 120, 121, 122, 128n14, 144
Sample bias, 26, 45, 69, 70, 75, 87, 90
Sample frame, 26, 45, 87
Sample size, 316, 332n10; effective sample size, 111
Sampling error, 20n6, 102, 104, 107, 109, 111, 175n2, 332n10
Sampling theory, 3, 26, 320
Scalability, 186, 187, 198
Scaling, 118, 183–191, 198, 199, 254, 292; attitude indexes, 142–145; Bogardus scale, 188; Guttman scaling, 187, 198, 199, 211, 221, 227n1; Likert scale, 187; Thurstone scale, 187, 198. *See also* Psychometric models; Rasch measurement model
Schema, 290, 291
Scholastic Aptitude Test (SAT), 122
Science, attitudes toward, 4, 5, 19n3
Seasonwein. *See* Roger Seasonwein and Associates
Secondary analysis, 41, 42, 327
Selective reporting of survey results, 81, 84–86, 212, 340
Self-report, 8, 88, 117, 195, 412, 417, 419, 420–421
Self-selection bias, 26, 46, 75, 77, 92n7, 260, 278n4, 308
Sex-typing, 202
Shewhart, W., 19n1
Sibley, E., 19n1
Simple random sampling, assumption of, 175n2, n11, 176n20, 214, 227n1
Simple response variance, 108, 109–110
Social consensus: 10, 12, 183, 184, 185, 189, 191, 257, 410, 421. *See also* Frames of reference
Social desirability, 89, 267–268, 419, 430
Social distance between interviewer and respondent, 267, 268. *See also* Ethnic distance
Social indicators, 5, 10, 12, 20n8, 38, 57n16, 140, 324
Social Science Research Council (SSRC), 3, 31, 176n15, 318–319; review of 1948 election polls, 27–28, 174, 314
Social Science Research Council Survey Archive (U.K.), 42
Socioeconomic status scales, 192, 197–198

Socratic effect, 290
Soviet Union. *See under* Surveys
Sozialwissenschaftliches Institut der Bundeswehr, 37
Sparse tables, 223
Speeches against democracy, 131–132, 133, 139, 142, 419
Split-ballot experiments, 63, 82, 91, 130, 132, 133, 135, 136, 137, 138, 139, 140, 167–173, 174n2, 251–252, 285, 291, 293, 328; with abortion questions, 135, 293; with American/Communist reporters questions, 140–141, 143, 291; with Communist takeover in Vietnam questions, 133–134; with confidence in institutions questions, 285, 327; with happiness questions, 167–170, 171–173; with speeches against democracy questions, 131, 133, 139, 419
Sponsorship, 47, 84, 309, 310
SRC. *See* Survey Research Center
SRI International, 345, 347
Standard International Occupational Prestige Scale, 184
Standardization, 16, 21n12, 62, 121, 261, 275
Standard of living, 11, 195
Standards and Guidelines for Federal Statistics (Circular A-46), 67, 332n6
Standards for surveys, 16, 27–28, 53, 61–73, 309–310; attempts at setting, 62–67; codes of, 61, 62, 63, 64–65, 66–67, 79, 84, 90, 91n1, 309, 332n6; enforcement of, 14, 62, 64, 66–67, 73, 77, 79, 90, 91n1, n3, 306, 307, 308, 309, 310, 311, 312, 313, 314; for federal surveys, 61, 67, 311, 312; in academic journals, 71–73, 74, 90, 312, 313; indicators of survey quality, 68–73. *See also* Disclosure; Reporting practices; Review process for surveys; Surveys, critiques and reviews
Stanton, F., 19n1
Starch/INRA/Hooper, 351n2
Statistical analysis. *See* Survey analysis; Log linear models; Association, patterns of; Rasch measurement model; Psychometric models; Cross-classifications
Statistical interactions, 127n13, 224, 226
Statistics Canada, 108, 110
Steiner television survey, 150, 152
Stephan, F., 19n1
Stimulus range effects, 292
Stimulus scaling, 186, 191, 292
Stock, J. S., 19n1
Stouffer, S. A., 3, 19n1, 52
Stouffer Survey, 149–151. *See also Communism, Conformity, and Civil Liberties*
Straw vote, 25, 26
Structural equation models, 227n1
Subjective and objective phenomena, relationship of, 10–12, 124–125, 180, 186, 191–197, 410–412
Subjective phenomena, 10, 20n8, 98, 324, 412; and social facts, 10, 12, 20n8, 410; defined, 8, 9, 10, 409
Subjective social class, 196
Sugging (selling under the guise of a survey), 73, 268, 315
Survey, definition of, 7–8, 25, 370
Survey analysis, 179–180, 181–183, 197, 203, 208, 316, 318, 323, 332; classical approach, 181, 197, 227n1, 333n13; modernization of, 323–324. *See also* Association, patterns of; Log linear models; Rasch measurement model; Psychometric models; Cross-classifications
Survey auspices, 38, 153, 261, 270, 338, 339
Survey concepts and definitions, 110, 114, 251, 410–411. *See also* Concepts; Frames of reference
Survey costs, 30, 315, 343
Survey design, 68, 179, 315–318, 319, 323, 332n11–12
Survey documentation, 42, 63, 68, 135, 309, 310, 332n5
Survey industry, 30, 47, 54, 66, 341–352; number of interviews, 30, 33, 343, 349; revenues, 30, 342–348, 350n2; trends, 30
Survey measurements, topics of: abortion, 77, 133, 135, 199, 202, 228n8, 293; auto safety, 162, 164; changing how country is run, 251–252; Communist reporters in U.S.A., 140, 141, 160; confidence in leadership or institutions, 5, 157, 160, 162, 209–210, 285, 327; Consumer Protection Agency, establishment of, 78, 92n10; crime vic-

Survey measurements, topics of (cont'd.) timization, 110, 138, 160, 297, 414; criminals, punishment of, 160, 163, 164; desirable qualities in children, 182; economic issues, 10–12, 32, 79, 85–86, 132, 157, 162, 171, 195; ethnic distance, 188–190; family size, 157, 164; favorite actors or entertainers, 136–137; foreign language use, 113–116; foreign policy, 77, 133–134, 155, 327; freedom of speech, 131–132, 142, 202, 419; government spending, 5, 156, 160, 163, 182; gun control, 13, 130, 131, 164; happiness, 5, 10, 12, 19n4, 32, 55, 156–157, 162, 167–174, 195, 203, 207–208, 301n3, 412, 419; Hispanic origin, 251; legalization of marijuana, 160, 164; neighborhood services, 193–194; nuclear power, 86, 209–210; personality and personal outlook, 123–124, 128n16, 163, 202; political self-identification, 164; presidential popularity, 158, 165; prestige of occupations, 162, 183–185; racial intermarriage, 143–144; religion, 206, 338; Russian reporters (*see* Communist reporters in U.S.A.); seriousness of crimes, 185–186; sex roles, 119, 163, 202, 212–221; social class self-identification, 196–197; surveys, frequency of, 352n11; traffic noise, 162; trust in government, 338; voting preferences, 120, 160. *See also* 371–405; 420–431

Survey of American Professors (1977), 7, 20n5, 312

Survey of Fishing, Hunting, and Wildlife Recreation, 348

Survey of Indochinese Resettlement, 35

Survey of Travel to Work, 348

Survey organizations, 39–40, 47, 261, 300n1, 342; collaboration among, 174, 318–319, 322, 328; comparisons of, 69, 150, 153, 154, 161, 259, 260, 261, 300, 319, 343 (*see also* Intersurvey comparisons). *See also* Academic survey organizations; Commercial survey organizations

Survey Research Center (U. of California), 347

Survey Research Center (U. of Michigan), 27, 39, 70, 110, 150, 152, 158, 167–170, 281, 313, 349; estimates of voter turnout by, 286; interviewing practices of, 257, 259–261, 266, 269, 273, 277n2, 278n3, 280–282; survey refusal rates at, 268, 338

Survey Research Laboratory (U. of Illinois), 347

Survey standards. *See* Standards for surveys

Survey uses: in court, 43–44, 58n21, 87; in policy and program evaluation, 56n13, 87, 88, 245–250, 339; in regulation, 44–45; trends in, 39, 41, 50, 86. *See also* Academic surveys, Commercial surveys, Government surveys

Surveys: allocation of resources for, 315–316; conducted during Vietnam War, 36–37, 59n26; conducted in Germany, 37, 56n14; conducted in Great Britain, 30, 31, 37, 45, 47–50, 51, 57–58n18, 69, 79–80, 184, 350, 353–355; conducted in Japan, 56n14; conducted in People's Republic of China, 57n15; conducted in Poland, 52; conducted in Soviet Union, 56n15; conducted in tandem, 149–154, 158, 161, 166; contracting for, 39, 57–58n18; critiques and reviews of, 3, 19n1, 37, 53, 68–69, 79, 92n10, 307–310, 312, 322–323, 328, 331n4, 340; factual, 32–33; foreign, 35–37, 56n14; history, 23–28, 236–237; influence of, 50, 52–53, 54, 330; in-house, 348–349; interpreting, 28, 62, 64, 92n10, 119–120, 138–139, 141, 235, 237, 238, 244–245, 276–277, 286, 306, 307, 316, 323; misuse of, 73–90, 330; proprietary, 84, 311, 328, 339; public acceptance of, 306–308, 315, 338; public exposure to, 47, 59n24, 261, 263, 278n5, 303, 349, 352n11, 353–356; quasi-factual, 32–33; role in society, 50–54, 329–331; sponsorship, 47, 84, 309, 310. *See also* Academic surveys; Commercial surveys; Government surveys; Panel surveys

Suspicion, 270, 275

Tandem surveys, 149–154, 158, 161, 166

Tasks posed by survey questions, 184, 207, 418; evaluation, 417–419; generalization, 299, 415, 416 (*see also* Questions, general vs. specific); inference, 415–416. *See also* Recall

Teeter, R., 53
Telephone interviewing, 46, 55n4, 111, 113, 176n16, 277n1. *See also* CATI
Television, 30, 53, 58n23, 152, 328. *See also* ABC, CBS, NBC
Television surveys. *See* Steiner television survey; QUBE
Temporal anchor, 297
Test anxiety, 266, 275
Test conditions, 121
Thinking, effect of, 287, 289, 290, 291, 297
Three Mile Island accident, 86
Thurstone attitude scale, 187
Total survey error, 103, 322
Toward a System of Social and Demographic Statistics, 38
Trademark infringement, 44
Trait concepts, 417, 430
Transfer parameters, 226
Trends: compilations, 29; studies, 4, 5, 11, 139, 140–141, 144, 184, 189, 193, 195, 196, 251, 288, 325, 327, 328
True value, 97, 98, 104, 107, 112, 116, 117, 121, 147

Ulstrata, 199, 222, 223, 224, 225
Ulstrith, 229n19
Undercount, census, 87, 270, 278n7
Underreporting, 114, 116, 117, 268
Uniform association model, 205, 218, 229n17
Union Carbide, 70, 84–86, 91, 92n11–12, 332n4, 371–375
Union of Concerned Scientists, 76
United Nations, 4, 38
United States Information Agency (USIA), 35–36
United States v. *Hiss*, 58n19
Unobtrusive measures, 9
U.S. Census: of 1960, 109, 111, 113, 116–117; of 1970, 109, 113, 116–117; of 1980, 32, 55n8, 87, 100, 109, 113. *See also* Bureau of the Census
U.S. Department of Agriculture (USDA), 27
U.S. Department of Defense (DOD), 36–37
U.S. Department of Transportation (DOT), 87

Vague quantifiers, 13, 156
Validation, 5, 62, 88–89, 98, 99, 113, 125, 147, 180, 270, 325, 326, 407, 417, 420
Validity, 3, 88, 101, 120–125, 233, 245, 275–277, 298, 300, 408, 413; construct validity, 106, 121, 123, 124, 147, 250, 408, 412; content (face) validity, 121, 122, 124; predictive (criterion) validity, 13, 62, 121–122, 123, 128n15, 256, 276, 277, 289, 417
Valuation, 417–419
Values for children, 182–183
Vanderbilt Television News Archives, 59n23, 328
Variance, 101, 102, 106, 111–112, 156; estimation, 68–69
Variance-covariance matrix, 103
Variance decomposition, 102, 103, 104, 108, 109
Verification, 10, 98, 109, 113, 116
Victimization. *See* Crime
Vietnam and the Silent Majority: The Dove's Guide, 52, 59n26
Vietnam War, 36–37, 59n26, 133, 262
Voluntary response, 153, 261, 269, 338, 339
Voter preference. *See* Candidate preference
Vote-splitting, 255–256
Voting, 28, 53, 253, 256, 286, 343, 414

Walker Research, 30, 351n2, 352n11
Wall Street Journal, 78
War Department, 27
Warwick, C., 19n1
Washington Post, polling at, 349, 352n10. *See also* ABC News/*Washington Post* poll
Westat, 345
Wilbur, W., 19n1
Wilks, S. S., 19n1–2, 154, 176n15
Wishful thinking, 415
Wissensoziologie, 329
Women's working, 119, 212, 213

Yankelovich, Skelly, and White, 66, 331n4, 346
Yea-saying, 124, 142, 144, 267

Zentralarchiv für Empirische Sozialforschung, 42