Making Order of Scientific Chaos

The Explosion of Contemporary Science

"If I have seen further it is by standing on the shoulders of giants," Isaac Newton wrote to Robert Hooke in 1675/6. In assuming this modest pose, he alluded to a fundamental assumption that our culture makes about science, namely, that it is progressive and cumulative, a corollary of which is that forays into the unknown by any researcher, however brilliant, are merely extensions of the knowledge amassed up to that time. For centuries it has been an article of faith that scientists base their research on existing information, add a modicum of new and better data to it, and thereby advance toward an ever more profound, complete, and accurate explanation of reality.

But today we are experiencing a crisis of faith; many of us no longer feel sure that science, though growing explosively, is moving inexorably toward the truth. Indeed, "growing explosively" is an ominous oxymoron: "growing" implies orderly development, but "explosively" denotes disorder and fragmentation. Virtually every field of science is now pervaded by a relentless cross fire in which the findings of new studies not only differ from previously established truths but disagree with one another, often vehemently. Our faith that scientists are cooperatively and steadily enlarging their understanding of the world is giving way to doubt as, time and again, new research assaults existing knowledge.

In recent years, however, methodologists in a number of scientific disciplines have been developing an antidote to the increasingly chaotic output of contemporary research. Known as meta-analysis, it is a means of combining the numerical results of studies with disparate, even conflicting, research methods and findings; it enables researchers to discover the consistencies in a set of seemingly inconsistent findings and to arrive at conclusions more accurate and credible than those presented in any one of the primary studies. More than that, meta-analysis makes it possible to pinpoint how and why studies come up with different results,

and so determine which treatments—circumstances or interventions—are most effective and why they succeed.*

To appreciate how anarchic contemporary research has become and how needed this new methodology is, one has only to read the daily papers. Here, for instance, are two typical recent news stories:

> NEW STUDY FINDS VITAMINS ARE NOT CANCER PREVENTERS
>
> A new study [reported in *The New England Journal of Medicine*] has failed to find evidence that vitamin supplements protect against the development of precancerous growths in the colon. . . . Many [previous] studies had found that people who eat large amounts of fruits and vegetables had lower cancer rates, and fruits and vegetables are known for providing vitamins C and E.[1]

> STUDY SAYS EXERCISE MUST BE STRENUOUS TO STRETCH LIFETIME
>
> Moderate exercise may well be the route to a healthier life, but if living longer is your goal, you will have to sweat. A new Harvard study that followed the fates of 17,300 middle-aged men for more than 20 years has found that only vigorous and not nonvigorous activities reduced their risk of dying during the study period.[2]

In a follow-up, the writer adds: "The new finding . . . has surprised leading researchers in the field. They are striving to reconcile it with many other studies that point to a life-saving benefit from moderate exercise, and they are perplexed that the Harvard study failed to find the expected benefit."[3]

Some other instances of seeming disarray in recent scientific findings:

- Ten studies determine how much the risk of ischemic heart disease (blockage of heart arteries) is reduced when serum cholesterol is lowered by roughly one-tenth of the average levels in Western countries. All ten studies conclude that it does reduce the risk, but the reported reduction ranges from nearly 40 percent in one study to as little as 15 percent in another.[4]
- Twenty-one studies of the use of fluorouracil against advanced colon cancer all find it beneficial, but findings of its effectiveness vary so widely—from a high of 85 percent to a low of 8 percent—as to be meaningless and useless to clinicians.[5]

* The method has many other names, among them research synthesis, evaluation synthesis, overview, systematic review, pooling, and structured review. In general, I use the term meta-analysis, since it is the one most often used in journal titles, indexes, and data bases.

- In 1994 a study published by the National Task Force on the Prevention and Treatment of Obesity reported that "yo-yo dieting" (the repeated losing and regaining of weight) poses no significant health risks—a direct contradiction of the findings of previous studies that off-and-on dieting can disrupt the body's metabolism, increase body fat, lead to heart problems, and heighten the risks of suffering other health problem.[6]
- A recent major study of the effects of exposure to the electromagnetic fields that surround power lines and electrical equipment shows a stronger link between electromagnetic fields and brain cancer than any previous study—but also contradicts earlier studies by finding no evidence of increased risk of leukemia.[7]

Such cases are legion not only in medical and biological research but also in behavioral and social science research:*

- Many studies of the treatment of aphasia (loss of speech due to brain damage) by speech therapy find it valueless, while others find it distinctly effective.[8]
- A number of studies of the effect of coaching on Scholastic Aptitude Test scores have shown that it raises them significantly, others that it raises them only trivially.[9]
- A generation ago, the Department of Health, Education, and Welfare asked Richard J. Light, a statistician at Harvard University, to determine whether the Head Start program worked. Light found a wealth of research data in thirteen studies that had already evaluated the program. The first twelve all showed modest positive effects, but the thirteenth, far larger than any of the others, disconcertingly showed no effect.[10] "I had no idea what to do," Light recently told a reporter for *Science*; his bewilderment eventually motivated him to develop a way of combining disparate research results, a precursor of meta-analysis.[11]
- Some studies find school desegregation to improve the academic achievement of black students significantly; others find only modest gains; and still others observe hardly any improvement. Even more confusing, some social scientists present credible evidence that desegregation improves achievement, but others offer equally credible evidence that it diminishes achievement.[12]

* For brevity, the behavioral and social sciences are referred to hereafter as the social sciences.

- Do women in management have a different leadership style from men? The question has long been hotly debated: some management experts and social scientists claim the evidence shows they do differ, others that the data yield no clear pattern of differences in supervisory style.[13]
- A massive and influential review of the scientific literature on sex differences assembled during the heyday of feminism by two respected women psychologists found little evidence of such differences in any area of social behavior except aggression. But later studies have furnished experimental and observational evidence of sex differences in many kinds of social behavior, including helping, sending and receiving nonverbal messages, and conforming to group values.[14]
- The Department of Health and Human Services recently ordered a review of studies of the prevalence of alcohol, drug, and mental disorders among the homeless, expecting the information to help in developing sensible policies for reducing homelessness. A reviewer located eighty studies containing an abundance of data—but no answers. The estimates in the studies differed so widely as to be useless: alcohol problems in the homeless population, from 4 percent to 86 percent; mental health problems, from 1 percent to 70 percent; and drug problems, from 2 percent to 90 percent.[15]

And so on.

Why have the sciences apparently degenerated into an intellectual free-for-all in our time?

In truth, there never was a golden era of pure harmony and cooperation among scientists. There have always been and are always bound to be competition, disagreement, and conflict among those pursuing research in any given area, since no two researchers think, perceive, or conduct a study in exactly the same fashion, nor are any two laboratory experiments or field observations exactly alike. Even when two researchers use the same methods to study a phenomenon, normal sampling errors (akin to the chance variations in the sum of two identical dice thrown repeatedly), minor differences in the persons they are studying, and other random factors make it unlikely that they will get the same or even very similar results. Accordingly, comparable and even replicate studies of any subject almost never yield identical findings.

In the social sciences the possibility of disparity is far greater than in the physical and biomedical sciences.[16] So many interacting variables influence human behavior that no two groups of human subjects are identical, even if the groups are large and carefully equated. Moreover, human subjects, unlike cells in vitro or bacteria in a patient's body, often react to

experimental situations according to their own volitions and past experiences, thereby adding unique influences to the effects of the variables being examined by the researchers and supposedly under their control.

While discrepancies and contradictions have always existed in scientific research, today they are more numerous, well publicized, and disturbing than before. One obvious reason for the increase is the mushroom-like expansion of the sciences in the last half century. In medicine, for instance, during a single recent year the *New England Journal of Medicine* and the *British Medical Journal* alone devoted some 4,400 pages to 1,100 articles, and currently, throughout the world, over two million medical articles are published each year.[17] With such voluminous output and so many new areas of investigation, it is inevitable there should be more disagreement than ever.

The reward system of science greatly intensifies the potential for disagreement. Career success is contingent on publication, and publishers are most interested in those studies that present news—findings that challenge the previously accepted wisdom. Although the primary motive of researchers is new knowledge, they are bound to hope that the results of their investigations will correct, conflict with, or disprove the results of earlier studies and those of concurrent research by colleague-competitors; such hopes, as a wealth of experimental evidence has shown, often unconsciously affect the researchers' performance in ways that tend to produce the desired result.

The resulting intellectual melee does serious injury to science and society.

For one thing, it impedes scientific progress: As the volume of research grows, so does the diversity of results, making it all but impossible for scientists to make sense of the divergent reports even within their own narrow specialties. In consequence, their research tends to be based less on the accumulated knowledge of their field than on their limited and biased view of it.

For another, when legislators and public administrators seek, through hearings and staff research, to study a pressing issue, they can rarely make sense of the hodge-podge of findings offered them. In the recent congressional debates about smoking in public places, members of Congress were told by antismoking advocates that many studies, including a report by the surgeon general, found "passive smoking" (inhalation of others' smoke) to be a cause of lung cancer. Tobacco-state colleagues and tobacco-industry lobbyists told them, however, of other studies by qualified researchers showing little evidence of such a connection, and even, remarkably, of two small 1984 studies and a larger, more recent one—in the authoritative *New England Journal of Medicine*—that found *less* lung cancer in people exposed to others' smoke in the home than in

people not so exposed.[18] Who could blame a legislator for not relying on research to help him or her decide how to vote on the issue?

Lastly, the prevalence of scientific disparities is eroding public belief in and support for research. Many intellectuals see the conflicting findings as justifying "constructivism," a view now popular among the "postmodernist" academic Left that scientific discoveries are not objective truths but only cultural artifacts, not representations of reality but self-serving products of the system.[19] At a different intellectual level, many of the uninformed and gullible see the contradictory outcomes of current research as grounds for broadly rejecting scientific knowledge in favor of simpler and more coherent beliefs—or "faiths"—in the power of prayer, guardian angels, miracles, astrology, past-lives regression, channeling, back-from-death experiences, and assorted New Age psychic phenomena.[20]

The Classic—and Inadequate—Solution

In most fields of science the standard way of dealing with the multiplicity of studies and divergent findings has long been the "literature review" or "research review." Scientific reports customarily begin with a brief résumé of previous work on the problem being considered; in that tradition, for some decades journal editors have published occasional articles summarizing and evaluating recent studies in actively researched areas of their discipline, and nearly every field has a type of annual review journal consisting entirely of such résumés.

Review articles, according to Howard White of the College of Information Studies, Drexel University, "are generally admired as a force for cumulation in science."[21] From the vantage point of meta-analysis, however, that tribute seems unwarranted. It is true that a good review article can marshal and summarize recent work on a particular topic; it is true, too, that one can only admire those who perform the heroic task of reading scores of often dense, technical, and tedious studies and summing up each in a sentence or two. But anyone conversant with meta-analysis will question whether reading the desiccated summaries in such articles—not unlike chewing a mouthful of dry bran—yields a genuine integration of the new knowledge. Consider, for example, the following brief excerpt from a recent review article on individual psychotherapy:

> Some comparisons of psychotherapy and drug treatment have suggested that combined treatment may present definite advantages over either treatment alone (Frank & Kupfer 1987; Weissman et al 1987; Hollon et al 1988), others have shown no differences between psychotherapy and psychotherapy plus medication at termination (Beck

> et al 1985), and still others have shown advantages at follow-up for patients who received cognitive-behavior therapy (Simons et al 1986). In the comparison of a cognitive-behavioral (prescriptive) therapy and a dynamic-experiential (exploratory) treatment of depression and anxiety, Shapiro & Firth (1987) found a slight advantage for the prescriptive approach, especially on symptom reduction.[22]

This specimen exemplifies the typical achievement and typical failure of the research review article: Although it offers a handy list of items in a particular area of research, it does little to integrate or cumulate them. Some reviews do offer more combinatory conclusions, but not methodically or rigorously; a recent critique of fifty medical review articles said that most summarized the pertinent findings in an unsystematic, subjective, and "armchair" fashion.[23] In an even harsher appraisal of medical review articles, two leading medical meta-analysts, Thomas Chalmers and Joseph Lau, write,

> Too often, authors of traditional review articles decide what they would like to establish as the truth either before starting the review process or after reading a few persuasive articles. Then they proceed to defend their conclusions by citing all the evidence they can find. The opportunity for a biased presentation is enormous, and its readers are vulnerable because they have no opportunity to examine the possibilities of biases in the review.[24]

Such criticisms apply to review articles in other fields of science. In *Summing Up*, a handbook of meta-analysis, Richard Light and David Pillemer characterize traditional review articles as not only subjective and scientifically unsound but "an inefficient way to extract useful information" because they lack any systematic method of integrating the relationships among the variables in the different studies and of reconciling differences in the results.[25]

Most review articles do not subject the studies they examine to the relatively simple statistical tests that would estimate how likely it is they mistook chance results—chiefly, sampling error—for meaningful ones (a false positive conclusion) or used too small a sample so that chance factors concealed the important results (a false negative conclusion). Review articles, in short, offer knowledge without measurement, the worth of which was famously expressed long ago by Lord Kelvin:

> When you can measure what you are speaking about and express it in numbers you know something about it, but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of *science*, whatever the matter may be.

A Radical New Approach

In 1904 the British mathematician Karl Pearson invented a statistical method for combining divergent findings. At that time, the effectiveness of inoculation against typhoid fever was still unclear; not only did the results of different trials vary but the samples were so small that the results of any one trial might be partly or largely due to chance factors. Pearson's simple but creative idea was to compute the correlation within each sample between inoculation and mortality (a correlation is a statistic showing how closely one variable is related to another variable) and then average the correlations of all the samples; the result, balancing out the chance factors and idiosyncrasies of the individual studies, would be a datum more trustworthy than any of the individual statistics that went into computing it.[26]

Seven decades later, when meta-analysis finally caught on, its practitioners developed an array of more complicated and precise computations than Pearson's, but to this day the basic concept behind combining and reconciling studies remains as simple and radical as in 1904. Rather than reaching vague conclusions like those of review articles—"The majority of studies show that . . ." or "While some studies find the treatment effective, most fail to reach statistical significance"—the new approach asks, "How can we produce a precise, quantitative finding representing what the studies show when synthesized into one superstudy?"

Current meta-analytic techniques involve subtle and discriminating procedures, of which Pearson's averaging is one of the simplest. We will look at them later (in verbal, not mathematical, terms) when we peer over the shoulders of scientists as they conduct meta-analyses that resolve the ambiguities in bodies of important research data. For now, however, let us carry out a hypothetical experiment that will give us a first glimpse at how meta-analysts combine data, the central process in meta-analysis.

You are one of a hundred physicians taking part in the testing of a new fever-reducing medication, antipyron. You are to give the same specified dose of the drug to the next eight patients you treat for influenza, and to record their temperatures on taking the dose and again four hours later.

Your first case is a young man who has a fever of 104° F; four hours after taking the drug his temperature has plummeted to 98°, and you, naturally, are delighted and enthusiastic. Your second flu patient is a middle-aged man with a fever of 102°; disappointingly, in four hours the antipyron lowers his temperature only to 100°. You give it to six more patients—with varying results.

What can you conclude about the overall effectiveness of the drug? Can you average the data of your eight cases and arrive at a typical figure for fever reduction for the dosage given? Certainly. Indeed, the researchers with whom you are collaborating might then take your average and mathematically combine it with the averages turned in by the other ninety-nine physicians taking part in the study to arrive at an overall drug effect. They might discover, say, that antipyron reduced fever in flu patients by an average of 2.5° F in four hours, across one hundred medical trials and eight hundred patients. This is a rudimentary meta-analysis, indicating how effective the drug is on average.

But being a well-trained doctor, you do not accept your own finding—or the finding of all one hundred trials—as an adequate guide to treatment, since you know that all patients are not alike. For one thing, the young man who was your first case weighs 150 pounds, the older man who was the second case, 220 pounds, and, typically, the greater the body weight, the less effect a given dose will have, since its effect is more diffused in a larger system. For another, the men also differ in age; perhaps the drug works more swiftly in a young body, with its higher metabolic rate, than in an older one. Of course, the other ninety-nine physicians know this too. So they, like you, not only recorded the change in each patient's temperature but also their weight, age, and possibly some other data. The project researchers might have compiled your data, arranging the patients in the order of their weight:

Patient	Age	Weight	Temp. change
1	20	150	−6°
5	50	160	−5°
4	35	170	−4°
6	60	190	−4°
3	45	180	−3°
7	30	200	−3°
8	25	210	−2°
2	55	220	−2°

With the data arranged in this fashion, you would easily see a clear relationship between the weight of the patient and the effect of the drug; with one exception, patient 6, as the patient's weight goes up the effect gets smaller. It is less clear that age affects the outcome.

Using your data, it is possible to calculate the correlations between the variables of weight and temperature change and between age and temperature change. A correlation will be +1.00 if the relation between

two variables is perfect and positive (that is, as one variable goes up, so does the other), −1.00 if it is perfect but negative (as one goes up the other goes down), and zero if no relation exists at all.

When you receive the researchers' report, you learn that in your own set of eight patients the correlation between weight and the drug's effectiveness is −.93 and between age and effectiveness, −.17. But when the researchers meta-analyzed—combined—the correlations for all one hundred data sets, they found that the average correlation between weight and temperature change is −.85, a strong (but not perfect) negative relation. The average correlation between age and effectiveness is only −.13, a weak association. Based on these meta-analytic findings, they conclude that higher doses of the drug may be necessary when treating heavier patients but that age, being largely irrelevant, does not need to be taken into account.

This imaginary experiment is a simplified version of only one process central to meta-analysis, namely, *combining* the findings of different studies. But a second and equally central problem exists: *reconciling* differences among studies. Suppose that nearly all one hundred physicians reported average reductions in their patients' fever of 3° but that one doctor reported an average reduction of only 1° and another an average reduction of 5.5°. How do you reconcile these differences from the more general finding? A little detective work might reveal that the small reduction came from a clinic that treated obese patients while the large reduction involved athletes. The meta-analyst might then suggest that differences among the patients are the clue to reconciling the disparate outcomes.

With that brief rundown behind us, let's return to Pearson's first effort at combining the data of a set of studies. In 1904 the mainstream scientific community expressed little interest in Pearson's techniques. The time was not ripe, and the idea of synthesizing studies mostly languished for the next six decades. Only a handful of avant-garde scientists pursued Pearson's ideas in the first half of the century, sensing the impending need to synthesize studies.

In the first third of the century, for example, agricultural scientists conducted numerous experiments in farming techniques but were unable to draw general conclusions from them since the tests almost always involved differences in soil, agricultural practices, climatic conditions, and so on. The problem was how to reach any useful generalizations from these dissimilar studies. A statistician named Leonard H. C. Tippett found an answer. From each experiment, Tippett obtained three pieces of data: the size of the sample, the size of the difference in crop

yield between different farming techniques, and the amount of variation in yield that occurred by chance within any specific technique (for instance, in experiments using the same kind and amount of fertilizer, how much variation in yield occurred without any known cause?).

With this information, Tippett was then able, adjusting for sample size, to compare the difference *between* techniques to the difference *within* techniques, the latter being a measure of how much variation might occur by chance. This enabled him to calculate the likelihood that the difference in yield between farming techniques was due to chance—and conversely, the likelihood that the technique was causing real improvements in yield. He might, for instance, discover that, given a particular sample size and a certain amount of *within*-technique variation, only twelve times in one hundred would a *between*-technique difference of a prescribed size as large as existed in that set of studies occur solely by chance. (Today we call this number the "probability of a chance finding.")

Then Tippett made a notable leap: He worked out a statistical method of combining the probability values of the several studies. This statistic, bypassing all the differences among the studies, showed how likely it was that the results of the whole set of studies were due to chance and, conversely, how likely that the results arose from the new farming technique.[27]

Although a handful of other researchers working with agricultural studies soon used Tippett's method or their own variants of it, scientists in other disciplines did not adopt the approach. Nor did they accept the other methods of combining probabilities constructed by a few avant-garde statisticians working on research in education, psychology, and the social sciences. In 1937 an imaginative biostatistician, William G. Cochran, went off in a different direction and worked out a way of combining the sizes of the effects reported in studies rather than the correlations between treatment and effect; although this approach would become a key technique of meta-analysis, it also attracted little attention.[28]

But by the 1950s the sciences were growing explosively, and scientists increasingly needed to sum up the proliferating studies in their fields and reconcile their differences. A small but growing cadre of researchers began developing methods to combine the results of studies within medicine, psychology, and sociology.[29] By the early 1970s, others were designing methods for aggregating studies of teaching methods, television instruction, and computer-assisted instruction. Robert Rosenthal, a social psychologist at Harvard, was developing a technique for combining the effect sizes of psychological studies at the very same time that Gene V Glass, a professor of education at the University of Colorado, was working out

a remarkably similar method of combining studies of the effects of psychotherapy, though neither knew of the other's work.[30]

Finally came the event generally cited as the beginning of the meta-analysis movement. In April 1976, Glass, then president of the American Educational Research Association, delivered his presidential address at the annual meeting, held that year at the St. Francis Hotel in San Francisco. For this important event, he chose to highlight a new and higher level of scientific analysis to which he gave the name "meta-analysis." Glass, then in his mid-thirties and fully aware of the topic's potential importance, had labored and agonized over the paper for two years, during which, he recently said, "I was a basket case."* As the day of his address neared, he was desperately afraid that the audience of a thousand would either drift away, doze off, or deride his ideas. But Glass, who describes himself as a highly competitive person, stepped cockily to the podium and with seeming self-assurance gave a lucid, witty, and persuasive talk. The audience, recalls psychologist Mary Lee Smith (who was then his wife and had worked with him on the meta-analysis paper), was "blown away by it. There was tremendous excitement about it; people were awestruck." His address, published later that year in *The Educational Researcher*, was judged by many who read it to be a breakthrough applicable to all sciences.[31]

The meta-analytic process, briefly sketched in Glass's paper and later spelled out by him and two collaborators, as well as by others, has five basic phases that parallel the phases of conducting a new study. They can be summarized in a few phrases, though the details fill books:[32]

1. Formulating the problem: Deciding what questions the meta-analyst hopes to answer and what kinds of evidence to examine.
2. Collecting the data: Searching for all studies on the problem by every feasible means.
3. Evaluating the data: Deciding which of the gathered evidence is valid and usable; eliminating studies that do not meet these standards.
4. Synthesizing the data: Using statistical methods, such as the combining of probabilities and the combining of effect sizes, to reconcile and aggregate disparate studies.
5. Presenting the findings: Reporting the resulting "analysis of analyses" (Glass's phrase) to the wider research community, providing details, data, and methods used.

That sounds simple; in fact, the process is almost always complicated, tedious, and problematic. "The magnitude of the task cannot be

*Unpublished quotations throughout the text are from personal interviews conducted by the author.

overemphasized in my view," writes Nan Laird of the Harvard School of Public Health.[33] Daniel Druckman, an eminent political scientist who devoted three years of his evenings and weekends to single-handedly conducting a meta-analysis, says, "It drove me crazy. Never again!"

Nonetheless, from Glass's 1976 presentation to the present the prestige and practice of meta-analysis has grown steadily—at first slowly, then with increasing speed and spreading from one discipline to another. But not without initially encountering much scornful and even hostile opposition: In an article in *American Psychologist* in 1978, the distinguished English psychologist H. J. Eysenck contemptuously dismissed Glass's work as "an exercise in mega-silliness"; in 1979 a peer reviewer of a meta-analysis by Harris Cooper, which looked at studies of sex differences in conformity, wrote, "I simply cannot imagine any great contribution to knowledge from combining statistical results"; and when social psychologist Judith Hall gave a seminar on meta-analysis at the Harvard School of Public Health in 1980, to an audience made up primarily of natural scientists, she encountered such acerbic criticism and derision that, she recently said, "If they'd had any rotten tomatoes to throw at me, they would have."[34]

None of which prevented meta-analysis, an idea whose time had come, from winning converts in first one discipline, then another. Those who carry out meta-analyses, when asked what draws them to such work, offer a variety of motivations. One sees it as the "cutting edge" of science; another says he is "a compulsive data analyst" who has to solve the puzzles presented by disparate findings; a third calls herself a "neatnik" who likes "to bring order out of chaos, to tidy things up"; and nearly all share a desire to discover patterns in the seemingly hopeless jumble of dissimilar findings.

And so meta-analysis gained currency, with books on its methodology appearing in the 1980s and a few top-notch statisticians, among them Frederick Mosteller of Harvard, Ingram Olkin of Stanford, and Larry Hedges of the University of Chicago, refining the techniques. The best indicator of its success may be the frequency of its appearance in scientific journals. At first, editors, unsure that the method was either scientifically legitimate or a true contribution to knowledge, were reluctant to publish meta-analyses. But each year they published more than the year before. In 1977, five major data bases, ERIC (education), PsycINFO, Scisearch, SOCIAL SCISEARCH, and MEDLINE had zero listings of meta-analyses, but in 1984 they had a total of 108, in 1987, 191, and in 1994, 347, by which time the grand total in those five data bases was 3,444.[35] Recently, some journal editors have even plaintively asked contributors to ease up on submissions of meta-analyses.

A Sampler of Meta-Analytic Achievements

What has this flood of meta-analyses yielded? Not every effort has produced important findings; indeed, many have added little to the world's inventory of scientific knowledge. But a fair share have made important contributions and resolved long-standing uncertainties. In a number of cases, the validity of meta-analytic conclusions has been confirmed by later massive studies or clinical trials—which suggests that the latter were more or less unnecessary. Some of the more noteworthy cases are presented in chapters 2 through 6, but here is a handful of others in capsule form:

- Coronary artery bypass surgery to treat ischemic heart disease has been practiced for twenty-five years, but clinical studies of varied design have yielded widely divergent conclusions about its ability to reduce mortality as compared to medical treatment. There has even been some question as to whether bypass surgery, though it improves quality of life, has any life-extending value. A 1994 meta-analysis, collaboratively conducted by a dozen institutions in five countries, combined seven major trials that compared bypass surgery with medical therapy. It found that five years after treatment, the mortality rate of bypass patients was 10.2 percent while that of medically treated patients was 15.8 percent—half again as high—and that there was a comparable advantage for bypass patients at seven and ten years.[36]
- In the 1980s, calcium channel blockers were among the most commonly used drugs for acute myocardial infarction (heart attack), unstable angina, and certain other cardiovascular conditions. The National Heart, Lung, and Blood Institute and the Bowman-Gray School of Medicine cosponsored a meta-analysis of twenty-eight disparate studies. Its surprising conclusion, published in 1987, was that "calcium channel blockers do not reduce the risk of initial or recurrent infarction or death when given routinely to patients with acute myocardial infarction or unstable angina."[37]
- Between 1965 and 1980 at least fifty clinical trials sought to determine whether there was any benefit in giving preventive antibiotics to patients about to undergo colon surgery, rather than merely requiring the standard bowel-cleansing procedures. The clinical trials reported confusingly discrepant infection and mortality rates; a few even indicated better results for patients not treated with antibiotics. A meta-analysis conducted by a team at New York's Mount Sinai School of Medicine and published in 1981 clarified the issue: Combining the results of twenty-six trials that met the standards for meta-analysis, it showed

that antibiotic therapy reduced infection rates from 36 percent to 22 percent, and death rates from 11.2 percent to 4.5 percent.[38]

- In 1977, cimetidine, an H_2 blocker, dramatically changed the preferred treatment of peptic ulcers from surgery to pharmaceutical therapy, and over the next fifteen years two other H_2 blockers, famotidine and ranitidine, were introduced. Clinical studies differed as to which drug yielded the best results. A 1993 meta-analysis of sixteen trials directly comparing the drugs showed that famotidine taken at bedtime had significantly higher healing rates than either cimetidine or ranitidine and that the three did not differ significantly in terms of adverse reactions.[39]
- For many years researchers have debated whether chlorination of drinking water, which prevents many infectious diseases, is carcinogenic. Studies provided contradictory findings and the issue long remained unsettled. In 1992, however, a team at the Medical College of Wisconsin in Milwaukee meta-analyzed ten studies and reported that chlorination is correlated with slightly higher rates of rectal and bladder cancer but that "the potential health risks of microbial contamination of drinking water greatly exceed the risks" of the two cancers. The team also pointed out that the ten meta-analyzed studies were conducted in the 1970s; since then, federal standards have lowered the permissible level of chlorination, and the risk of the two cancers may now be lower.[40]
- Is intelligence related to the innate quickness of the individual's brain when reacting to external stimuli? The answer would cast light on the long-debated issue of the extent to which intelligence is determined by heredity rather than by experience and social influences. A good index of innate, unlearned mental speed is "inspection time" (IT), commonly measured by flashing two lines of different lengths on a screen immediately followed by a pattern to overcome the brief residual image in memory. An individual's IT is defined as the minimum time of exposure he or she needs to reliably discriminate between the two lines. Dozens of studies have yielded a mish-mash of answers, but a recent meta-analysis by psychologists John Kranzler and Arthur Jensen of the University of California–Berkeley found a strong negative correlation of about $-.54$ between IT and adult general IQ; that is, the longer the IT, the lower the individual's IQ.[41]
- Does alcohol cause aggressive behavior? Most studies have provided only correlational evidence—that is, alcohol and aggression tend to co-occur—but while correlation suggests some kind of link between the two, it does not prove causality; something else may cause the corre-

lation. The amount of time spent watching television, for instance, correlates with ill health, but TV watching does not itself impair health; sicker people watch more because they are less able to do other things.[42] Or, back to our example, it might be the case that people who lack the ability to control their impulses are more likely both to drink and to behave aggressively.

Experimental studies have yielded evidence for several competing causal explanations of the alcohol-aggression correlation. Among them: alcohol causes cognitive and emotional changes resulting in aggression; drinkers deliberately use alcohol so as to have an excuse for aggressive behavior; alcohol psychologically disinhibits persons who are predisposed to aggression; alcohol directly and physiologically causes aggression by anesthetizing the part of the brain that normally prevents such responses. Brad Bushman and Harris Cooper of the University of Missouri meta-analyzed thirty experimental studies in which the behavior of different kinds of drinkers was observed either after drinking alcohol, after drinking a placebo they thought was alcohol, or after drinking nothing. The meta-analysis revealed little or no support for any single causal theory but, synthesizing the results of the studies, Bushman and Cooper did conclude that alcohol definitely causes aggression, possibly through a combination of some of the hypothesized causal factors.[43]

- The extent to which violence on television stimulates aggressive, antisocial, or delinquent behavior has been a matter of controversy for over three decades. More than two hundred studies have yielded an array of answers; over the years, that lack of agreement has undoubtedly strengthened the hand of television programmers and weakened that of government regulators. A meta-analysis recently conducted for the National Research Council finally furnished a definitive answer: Viewers are more apt to commit aggressive or antisocial acts after seeing violence on TV (particularly violent erotica), the most common kind being physical violence against another person.[44]
- As mentioned earlier, the findings of studies of the prevalence of alcohol, drug, and mental health problems among the homeless have been extraordinarily inconsistent, ranging from 4 percent to 86 percent for alcohol problems, 2 percent to 90 percent for drug problems, and 1 percent to 70 percent for mental health problems. A meta-analysis by Anthony Lehman of the University of Maryland and David Cordray of Vanderbilt University made sense of these vast discrepancies, their synthesized findings being that 28 percent of the homeless have alco-

hol abuse problems, 10 percent have drug abuse problems, anywhere from 23 to 49 percent have mental disorders (depending on the category of disorder), and 11 percent have various combinations of the three. The figures represent the current prevalence of these problems among the homeless; far larger numbers have had such problems at some time in the past and presumably could have them again.[45]

- Fluoxetine (Prozac) came on the market in 1987, swiftly became the most prescribed antidepressant in the United States, and was hailed by the media as a wonder drug. Scores of studies said that it was far more effective than the tricyclics, the previous standard antidepressants. A team composed of researchers from the State University of New York at Syracuse and several other institutions carried out two meta-analyses of studies comparing the effects of Prozac with those of tricyclics and placebos under double-blind conditions (that is, with neither patient nor physician knowing what the patient was being given). The meta-analytic findings were illuminating: In many ostensibly double-blind trials, Prozac has less noticeable side-effects than tricyclics, which tend to cause dry mouth, blurry vision, and other obvious symptoms; as a result, patients and doctors were often able to guess correctly when Prozac was being administered and wishfully overevaluated its antidepressant effect. Correcting for this error and aggregating the results, the team found that Prozac was only a half to a quarter as effective as previously reported and no better than the tricyclics, except for the diminished side-effects.[46]
- A staggering amount of research is churned out annually on agricultural issues—far more than growers or agricultural administrators can master. An example: In 1992 alone, the National Agricultural Library added 363 new research items about strawberries. To demonstrate that meta-analysis can present both growers and administrators with easily comprehensible summary findings, Douglas Shaw, a professor of agriculture at the University of California–Davis, and statistician Ingram Olkin of Stanford University meta-analyzed a group of studies of chemical control and a group of studies of biological control that focused on the important strawberry pest *Tetranychus urticae* (a spider mite). The original studies in each group differed in their findings, but the meta-analysis clarified the matter: Although biological controls had a statistically significant effect, chemical controls were nearly four times as effective in terms of increased strawberry yield. The meta-analysis did not look at other benefits and harms that might be associated with yield.[47]

The Value of Meta-Analysis

Despite the track record of meta-analysis, a number of scientists still scorn and belittle it. Some deride it as "garbage in, garbage out," arguing that combining studies, even using fancy statistics, merges trashy research with sound research and therefore degrades the whole exercise. Others say meta-analysis "crowds out wisdom," since assistants usually do the tedious work of evaluating and compiling the data, and although they lack knowledge and mature judgment; the senior researchers, however, then meta-analyze the data as confidently as if they had compiled it themselves. Still others see meta-analysis as a fancy set of techniques for achieving ever greater precision in answering questions of possibly dubious merit.

David Sohn, a psychologist at the University of North Carolina, is even more caustic and rejecting. Asserting in a recent issue of *American Psychologist* that primary research is the only valid method of making discoveries, he ridicules the claim that meta-analysis is a superior mechanism of discovery:

> It is not reasonable to suppose that the truth about nature can be discovered by a study of the research literature. . . . Meta-analytic writers have created the impression, with a farcical portrayal of the scientific process, that the process of arriving at truth is mediated by a literature review. . . . After some critical mass of findings has been gathered, someone decides to see what all of the findings mean by doing a literature review, and thereby knowledge is finally established.[48]

Such is the minority view, however. The majority, as already documented, see meta-analysis as an important, even historic, advance in science. Below are a few of the major benefits that meta-analysis is widely agreed to yield:

- Physicians can now make decisions as to the use of therapies or diagnostic procedures on the basis of a single article that synthesizes the findings of tens, scores, or hundreds of clinical studies.
- Scientists in every field can similarly gain a coherent view of the central reality behind the multifarious and often discordant findings of research in their areas.
- Meta-analysis of a series of small clinical trials of a new therapy often yields a finding on the basis of which physicians can confidently begin using it without waiting long years for a massive trial to be conducted.
- In every science, meta-analysis can generally synthesize differing results, but when it cannot, it can often identify the moderator and me-

diator variables (about which more later) that account for the irreconcilable differences. By so doing, the meta-analysis identifies the precise areas in which future research is needed, a function of considerable value to science.

- On the pragmatic level, meta-analyses of a wide range of social problems have profound implications for social policy; their findings about such issues as the value of job training for the unemployed, the effects of drinking-age laws, and the rehabilitation of juvenile delinquents offer policy-makers easily assimilated syntheses of bodies of research they have neither the time nor the training to evaluate on their own.

Whether one sees meta-analysis as a set of recondite techniques for getting precise answers to irrelevant questions or as an epochal advance in scientific methodology, it unquestionably has come to occupy a major place in contemporary scientific research. Yet it is not itself a science and does not embody any scientific theory. It is, rather, a method or group of methods by means of which scientists can recognize order in what had looked like disorder. As Ingram Olkin writes, "I like to think of the meta-analytic process as similar to being in a helicopter. On the ground individual trees are visible with high resolution. This resolution diminishes as the helicopter rises, and in its place we begin to see patterns not visible from the ground."[49]

To use a different image, meta-analysis is a tool used by scientists. Far from being a belittling characterization, this is high praise, for as the illustrious physicist Freeman Dyson, a member of the Institute for Advanced Study at Princeton, recently commented, "The great advances in science usually result from new tools rather than from new doctrines."[50]