In recent years, social programs for the poor clearly have lost the support of the American public. At the same time, public opinion polls show that the vast majority of Americans remain sympathetic to the plight of the poor or, at least, to the plight of poor people who adhere to mainstream values and social norms. How are we to reconcile these two facts? The answer is simple. Most Americans are convinced that social programs simply do not work and that many existing programs encourage antisocial behavior and attitudes.

Certainly many programs have failed. Even worse, some of the failures have become apparent only after extravagant claims of success have been made. This has made the public exceedingly skeptical of all social programs and given rise to an increasingly popular myth that nothing works. This is simply not true.

A number of programs have had a substantial, positive impact on the lives of the people they have served and have benefited society as a whole. These programs are not miraculous. They do not completely solve any social problem. They do not help all of the people they serve or anything close to 100 percent. But they do substantially reduce the rates and severity of particular social problems among participants.

This volume describes some of the very best programs and documents their benefits. Chapters 2 through 10 each presents a description and formal evaluation of a single program. These particular programs were chosen after reviewing several hundred formal evaluations, because, according to a subjective interpretation of objective evidence, they stand out in one or more of five ways.1

Each program offers at least one of the following: (a) extraordinarily large benefits per dollar of cost (chapters 3, 5, 8, and 9); (b) unusually convincing evidence that the program delivers substantial benefits, regardless of cost (chapter 2); (c) convincing evidence of
long-term effects (chapters 4, 5, 6, and 9); (d) evidence of cost-effec-
tiveness on a national scale (chapter 7); and (e) new hope of making
progress to solve a seemingly intractable social problem (chapter 10).

Chapter 11 is different from the rest. It does not evaluate a single
program. No work and training program “really works,” because even
the best ones produce only small benefits. Yet this volume discusses
them because they play a key role in the Personal Responsibility and
the national welfare system, this is one of the most important pieces
of social legislation ever passed in this country. In chapter 11,
Lawrence Mead challenges the consensus by reinterpreting old data
and presenting new data from numerous work and training programs.
His arguments have important implications for the welfare revolution
that is just beginning.

Thirteen criteria are used to determine whether a particular social
program meets one or more of the five standards. They are also used
to interpret the policy relevance of evaluation results. These criteria
are detailed in the following section of this chapter. Although used
to judge the objective findings of the evaluations, most of these
criteria cannot be applied without some degree of subjective inter-
pretation.

CRITERIA
The first criterion is whether a program has statistically significant ef-
fects on the treatment group. This is a necessary, but clearly not a suf-
cient, condition for a program to be judged successful. It is neces-
sary because we cannot be confident that a finding is valid unless it is
statistically significant. But it is insufficient because a significant ef-
fect can be so small that it has no practical importance or policy rel-
ance whatsoever. All findings presented in this chapter are statisti-
cally significant, unless otherwise noted. Therefore, few references are
made to this criterion.

The second criterion is effect size. A program that works must have
“substantial” effects on participants’ lives. Unfortunately, the defini-
tion of what constitutes a substantial effect size is completely subjec-
tive. Therefore, each reader must ultimately decide whether the effect
of a particular program is substantial. As a broad rule of thumb, the
following changes can be considered substantial: a reduction in the
incidence of a serious social problem by more than 20 percent, a
change in a behavior or an outcome of more than 0.25 standard devi-
ation, and an income gain of more than two thousand dollars a year. In cases where more than one of these standards is relevant, the effect should meet the most stringent one. The first standard is a 20 percent change, not a 20 percentage point change. Thus, for example, reducing the proportion of individuals in a group who have committed a felony from 30 to 20 percent is only a 10 percentage point reduction, but it is a 50 percent reduction, which is substantial. Small percentage point reductions sometimes seem trivial, but they can be important when dealing with a national problem, because the total number of people involved is large. For instance, reducing the use of hard drugs from 2 to 1 percent is only a 1 percentage point decrease. But in a population of 250 million, this 50 percent reduction would imply that 2.5 million fewer people were using hard drugs.

The third criterion is a program’s cost-benefit relationship. If the monetary value of all costs and benefits could be determined for all programs being compared, it would be possible to ascertain which ones produce the most dollars of benefit for each dollar invested. In theory, this is the best way to judge programs and the best criterion for determining which programs should be funded. Unfortunately, determining the cost-benefit relationship is easier said than done. Although the costs are usually easy enough to measure, determining the monetary value of benefits is often difficult. Formal cost-benefit analyses have not been done for most of the programs presented here, but they are discussed where they are available. For the other programs, cost estimates and qualitative judgments are presented about the size of the benefits and their overall relationship to cost. In some cases, these judgments are necessarily quite rough, but in others the costs are either so large or so small that we can be fairly confident about whether the benefits exceed them or not. Some might question the value of imprecise estimates. However, in cases where the costs are extreme, in one direction or the other, subjective judgments about whether benefits exceed costs have a high probability of being correct. The opportunity costs of ignoring such information simply because we cannot generate a precise numerical estimate are potentially very large. Thus, despite the lack of precise estimates, the relationship between costs and benefits is the single most important criterion used here to judge how well each program works.

The fourth criterion is the length of time that effects last. This is directly related to the third criterion. The longer an effect lasts, the more benefits it is likely to generate. It is common for short-term ef-
fects to fade out over time. Some effects, such as temporary increases in academic achievement, may have no benefit at all. However, this does not imply that all short-term effects are worthless. Temporary academic gains that reduce grade retentions and special education placements clearly do save money. In chapter 5, Lawrence Schweinhart and David Weikart argue that although academic gains disappear, they contribute indirectly to other long-term benefits. Short-term health improvements reduce suffering and save money. Temporary reductions in drug use probably reduce crime and improve health outcomes. Nevertheless, all other things being equal, the longer effects last, the better.

The fifth criterion is the quality of the evaluation. No matter how large or long term a particular program effect is, it is meaningless if its estimate was not generated in a methodologically sound way. A pure experimental design in which evaluation participants are randomly assigned to treatment and control groups is the universally accepted gold standard in evaluation research. Results generated from studies using this design have a lot of credibility, although even they may suffer from problems such as attrition bias or issues of external validity. When random assignment is not feasible, evaluators typically select a comparison group, using any one of a number of techniques designed to make it as similar as possible to the treatment population. Unfortunately, it is impossible to be sure that the distinctions between the treatment and comparison groups are produced by the program and not by initial differences in the makeup of the groups. Even if the two groups appear to be similar or if statistical controls are used, unobservable differences could be generating variable outcomes. Thus nonexperimental evaluations always carry less weight. However, if there is evidence of equivalence between the comparison and treatment groups, and if other factors lend credence to the results, nonexperimental results should be taken seriously.

The sixth criterion is the relationship between the evaluator and the program. Evaluations done by program developers tend to yield larger effects than ones done by independent evaluators. Therefore, results are always more credible when produced by independent researchers. This is not to say that we should discount evaluations done by program developers. But large investments in such programs should be made only after the results have been verified by someone else.
The seventh criterion is replication. No matter how impressive the results, no matter how good the study design, and no matter how objective the researchers, a single evaluation of a single program is never enough to prove that it really works. The value of replication is a matter of simple multiplication. If the odds of generating a particular result are 1 in 20, the odds are 1 in 400 of generating it twice, 1 in 8,000 of generating it three times, and 1 in 160,000 of generating it four times. If a program is replicated many times, and the results are consistently the same, we can be very confident that the findings are real, unless they can be traced to a common design flaw in every evaluation.

The eighth criterion is a variation of the seventh. Some studies have not been replicated per se, but they can be compared to programs that are similar enough to be considered “near replications.” If such near replications produce similar effects, the original findings become more credible. Naturally, real replications are preferable, because what appear to be small differences among similar programs may turn out to be critical ones.

The ninth criterion is uniqueness. It is in a sense the opposite of the eighth. Several of the programs presented here have been included precisely because there are no similar programs with comparable results. They are uniquely successful. In statistical terms, they are outliers. This begs the question of whether their success may simply be a product of good luck. Even if all programs were completely ineffectual, it is likely that some of them would appear to be successful by random chance alone. To confirm that an unusually good result is valid, we should be able to distinguish features of the program that could explain its unique success—something that shows the program to be an orange in a distribution of apples, rather than just an apple that by chance is unusually juicy. Unfortunately, this is a highly subjective exercise. It is always possible to point to certain unique components or combinations of features that could account for a uniquely successful result. So the seventh criterion supersedes the ninth one. All results, no matter how impressive, need to be replicated.

The tenth, eleventh, and twelfth criteria are all related to internal consistency. Certain patterns of findings can strengthen our confidence in the overall results.

The tenth criterion involves the relationship between effect size and program intensity or time of participation. A positive relationship between the size of effect and the intensity of the program or the amount of time participants spend in it lends credence to an overall
finding of beneficial effects. In other words, if more of the program works better than less of it, it tends to confirm that the program really does work. Of course, a program’s benefits may reach a saturation point. Thus the absence of a relationship is not necessarily a problem, but it clearly is a problem if more participation or greater intensity reduces the effects. If more of a program is worse than less of it, any positive effects are called into question.

The eleventh criterion has to do with implementation fidelity. No replication of a program is ever a perfect copy of the original model. Whenever there are many replications, there is always some variation in the degree to which the copies remain faithful to the original design, when they are implemented. If a program really works, there should be a positive relationship between implementation fidelity and the similarity of the replication’s results to the original’s.

The twelfth criterion involves the relationship between the effect size and “site experience,” that is, the time a program has been in existence at a particular site. If a program really works, more experience with applying it could improve its effectiveness. So a positive relationship between a site’s experience with a program and the effect size is a good indicator of the program’s overall effectiveness. This criterion may, in part, be a variation on the eleventh. In many replications, it takes time to implement a program fully or to learn how to use it correctly. So site experience may sometimes be a proxy for implementation fidelity.

The thirteenth and final criterion has to do with scale. Even if a program really works at one or two sites, it will not necessarily be successful if replicated on a substantially larger scale. Although success on a small scale does have some value, particularly if it proves a concept or suggests a promising strategy, the ultimate goal of social program research is to develop good programs that can benefit lots of people. Programs that have already been “upscaled” successfully have a leg up on pilots, because they have shown that they can work outside a “hothouse” environment. This final criterion is not the same as the seventh. Although upscaling almost always involves replication, it presents the program with new obstacles and problems that only emerge once a certain size has been reached.

Having detailed the criteria for judging success, we can now apply them, in the following section, to determine whether the social programs presented in this volume really work. We can also use the criteria to decide whether each program merits the investment that would be required to scale it up substantially.
THE PROGRAMS
Success for All, presented in chapter 2, is aimed at improving the reading skills of children in the early grades of elementary school. Although the program is multifaceted, its main component is individualized tutoring, with one tutor teaching one child. This is crucial, because individualized tutoring is the key ingredient that makes the program successful.

A huge number of programs have been developed to raise the academic achievement of disadvantaged children. It is probably the single most common goal of social programs. For the most part, these programs have proven disappointing. There are isolated cases of gains, but very little in the way of consistent patterns of success that would demonstrate convincingly that a particular strategy works. *There is one clear exception:* individualized tutoring has consistently raised children’s reading levels. Three different tutoring programs have improved the reading test scores of thousands of children at scores of sites.

Reading Recovery was the first of these programs, and it served as the model for the other two. It remains perhaps the best known and most widely acclaimed of all reading programs. It probably deserves that acclaim, although Success for All is even more effective.

In Success for All, tutors work with individual students in twenty-minute sessions. The tutors are certified teachers who have experience working with disadvantaged children. They cover the same material and teach the same concepts addressed in the regular curriculum, although they often use different approaches.

The program has components besides tutoring, but individualized tutoring appears to be a necessary element for a successful reading program. Many programs include one or more of the other components of Success for All, but only those programs with tutoring consistently increase reading test scores by substantial amounts. What we do not know is whether tutoring is sufficient. Each of the three successful tutoring programs provides additional services. Until we vary the ancillary services in some kind of systematic experiment, there is no way to know for sure whether any of these services or any of the mixes of services unique to each program add to the overall impact.

What is the overall impact of Success for All? This has been assessed many times in many places with various kinds of children. These evaluations demonstrate a consistent pattern of positive and
statistically significant effects on reading test scores. In chapter 2, Robert Slavin and his coauthors report the results of a meta-analysis of Success for All. A meta-analysis is a study that attempts to determine the average effect size of a large number of individual evaluations. The meta-analysis presented here includes evaluations of Success for All in nineteen schools, involving roughly four thousand children (and a comparable number in comparison schools that did not use the program).

The average effect of Success for All was to increase children’s reading test scores by about half a standard deviation. This effect is not huge, but it is enough to bring a child reading in the middle of the bottom half of the distribution up close to average. For example, an increase of half a standard deviation would bring a child from the twenty-fifth percentile up to the forty-third. It would not bring a child at the very bottom to anywhere near the average.

However, the impact of Success for All on the very poorest readers was much larger. The average effect size for children reading below the twenty-fifth percentile ranged from 1 standard deviation for first graders to 1.7 standard deviation for fourth graders. A full standard deviation is a big effect by any measure, and a 1.7 standard deviation effect is huge. Consider, for example, a child reading at the tenth percentile, close to the very bottom. An increase of 1 standard deviation would boost him to the thirty-eighth percentile, while an increase of 1.7 standard deviation would catapult him to the sixty-fifth percentile, almost into the upper third of the population. Follow-up studies suggest that the effects are maintained through grade seven, at least two years after the final treatment. No data are available yet for older children.

Impacts of this size are impressive. But large effects do not in and of themselves guarantee that a program really works. The evaluations did not assign children randomly to treatment and control groups. Instead, they selected comparison groups composed of children from schools that were chosen because they were similar to the Success for All schools in terms of demographics and test scores. This kind of comparison is a lot better than none at all, but it is not as good as one predicated on random assignment. It increases the chances that systematic differences between the Success for All and comparison children could bias the results. In addition, because children were the unit of analysis, while schools were the unit of assignment, significance tests should have been adjusted for within-school correlations,
but they were not. Also, Success for All is only adopted in a school if 80 percent of teachers vote to have it. Thus there is a chance that the treatment schools have teachers who are more motivated on average than the comparison schools. This, too, could bias the evaluation results, making the program look more effective than it really is.

However, a number of factors lend credence to the positive results. First, the program has been successfully replicated and upscaled. It has consistently generated statistically significant effects of substantial size across numerous evaluations, sites, cohorts, and grade levels. Second, the results have been strongly significant in most cases, so it is unlikely that the bias in the significance tests have affected their outcomes. Third, the two similar tutoring programs that have been tested on a large scale have also consistently generated positive results. And fourth, there is evidence of a learning curve in the implementation process. The effects are larger in schools that have used the program longer. An increase occurs each year from the first through the fourth year of implementation. In this case, site experience may be a proxy for implementation fidelity, although there is no direct analysis of the relationship between implementation fidelity and the effect size.

In short, the lion’s share of the evidence suggests that Success for All is a program that really works. However, at least two issues remain. One is that most of the evaluations were done by the program’s founders or their protégés, at schools in which they were closely involved. We would have even greater confidence in the program’s effectiveness if more studies had been done by independent evaluators. The second, and more important, issue is that we do not know if the benefits of Success for All exceed its costs.

It is hard to value the benefits of educational programs that have not followed participants into adulthood, because the main outcome variables are test scores. Surprisingly, very little research has been done on the effects of test scores on long-term adult outcomes, such as earnings. Jencks and Phillips (1996) find that the effects of increases in high school mathematics scores on earnings can be fairly large, particularly at the bottom of the distribution. However, they only find significant effects of reading gains for girls. This work, although important, is too preliminary and too removed from the kinds of effects discussed here to use in estimating benefits.

The only direct monetary value that raising childhood test scores has is that it typically reduces grade retentions and special education...
placements, both of which are quite expensive. Success for All does reduce special education placements, but that alone is by no means enough to pay for the program in and of itself. The program may have long-term benefits. The cognitive gains may endure into adulthood and contribute to increases in income, reductions in welfare dependency, and other positive outcomes. Even if the reading gains fade away, valuable behavioral changes may emerge later. If so, the total benefits may ultimately exceed the costs, even when the long-term gains are discounted at a reasonable interest rate. Unfortunately, we simply do not know right now.

We have a little more information on the program’s costs. Slavin and his coauthors do not provide data on the total cost of Success for All. But based on experience from his own program, George Farkas (chapter 3) conservatively estimates that it costs at least three thousand dollars a year to deliver a full complement of tutoring sessions to the average child. The other services provided by the program add to this cost. If this were a one-time expenditure that yielded permanent effects, it would almost certainly be worth the investment. However, if that amount had to be spent several times in a child’s school years and if the gains completely faded out before adulthood, the program almost certainly would not be cost-effective. At this point, we do not know if effects tend to cumulate for children who receive multiple years of treatment, how much would have to be spent per child to generate permanent effects, or indeed if the program can generate enduring changes at all.

Until these questions have been answered, it will be impossible to determine whether the benefits of Success for All outweigh the costs. However, Slavin and his coauthors note in chapter 2 that because virtually all schools who adopt Success for All reallocate Title I resources to it, the incremental cost of the program is quite small. Evaluations suggest that Title I programs have, on average, small effects at best and perhaps none at all (Arroyo and Zigler 1993; Slavin, Karweit, and Madden 1989). This is not to say that no Title I programs work, but on the whole, Title I is not an effective social program. It is likely that replacing Title I with a national Success for All program would improve educational outcomes for poor children.

It is also possible that Success for All might be too expensive to provide to all disadvantaged children but might still be cost-effective for the bottom quartile of readers. If the large gains for the poorest readers hold up over time and contribute to improvements in earnings and other adult outcomes, they might very well justify a large investment.
At the very least, Success for All offers great hope. It demonstrates convincingly that an educational program can raise the academic achievement levels of disadvantaged children by substantial amounts. Moreover, a new program suggests that its principles can be applied in a more general way to completely restructure schools. Roots and Wings, discussed briefly in chapter 2, is a complete elementary school model that evolved out of Success for All. It is too early to pass judgment, but the preliminary results are extremely promising. They suggest that elementary schools with large proportions of disadvantaged children can raise achievement levels at least close to average in every subject.

Despite Success for All’s promise, the issues of cost and permanency are problematic. They are also intimately related. Most social programs attempt to deliver a one-shot intervention that will have permanent effects. There is evidence, described here, that such a strategy can reduce criminal behavior and other social problems. But this approach has never succeeded in generating long-term gains in academic achievement or cognitive skills. A different approach may be needed. Rather than attempting to “inoculate” students against academic failure, perhaps we should try to raise educational productivity on an ongoing basis. In other words, we may need to increase the amount that disadvantaged children learn every year that they are in school or at least increase it several times during their educational careers.

To be cost-effective, such an approach would have to be relatively inexpensive in terms of expenditures per child per year. As George Farkas shows in chapter 3, Reading One-to-One successfully delivers individualized tutoring at a fraction of the cost of Success for All. It does so by using part-time workers (often college students) as tutors, rather than experienced teachers.

Given its reliance on less-experienced tutors, one would expect the effects of Reading One-to-One to fall short of those of Success for All. But the evidence suggests otherwise. On average, each Reading One-to-One tutoring session increased students’ reading achievement 0.073 grade levels. Thus sixty sessions, which was about the average amount received by full-time program participants, generated gains of 0.44 grade equivalents (0.51 standard deviation). This is roughly equivalent to the average effect of Success for All. If the marginal effectiveness of the tutoring sessions held up as the number of sessions increased, then 100 sessions generated increases of
0.73 grade equivalents (0.85 standard deviation). These are large gains, although not as large as the ones that Success for All generated for the poorest readers.

What is most impressive about Reading One-to-One is that it delivers a large “bang for the buck.” The total cost, including administrative overhead, is about $8.38 per session. Thus sixty sessions of tutoring cost just $503, and 100 sessions cost $838. As noted, it is hard to do a true cost-benefit analysis of educational programs, because it is difficult to determine the monetary benefit of improved test scores. But if these numbers are accurate, Reading One-to-One delivers far greater gains in academic achievement per dollar spent than have ever been documented before.

These results suggest a number of exciting possibilities. At this rate, for example, one could deliver a sixty-session tutoring module to a child every year from grade one to grade twelve at a total cost of six thousand dollars, roughly the cost of one year of public school. If the marginal effectiveness of the program held up every year, the total increase in reading achievement would be five to six years by the time the child graduated. This would constitute a huge improvement in the productivity of the overall educational system for the child, clearly worth the 8.3 percent increase in overall cost.5

Unfortunately, we do not know whether the effectiveness of tutoring sessions would remain at such a high level as the total number of sessions increased and as the child moved past the third grade. The marginal effectiveness of tutoring is calculated from a sample of students who received an average of sixty sessions. So we cannot even be completely confident of the estimates for one hundred sessions in one year, much less for the seven hundred and twenty sessions over twelve years. However, even if we remain cautious in extrapolating from Reading One-to-One’s basic results, they suggest that the reading skills of disadvantaged children can probably be improved at a reasonable cost.

The fundamental question is whether these basic results are credible. Farkas was not able to assign students randomly to treatment and control groups. He did not attempt to establish a nonrandom comparison group. Instead, he measured effectiveness by analyzing the relationship between variation in the number of tutoring sessions and changes in the level of reading achievement. This is a systematic way of using the relationship between time of participation in the program and the effect size (the tenth criterion) as an evaluation tool. The validity of this approach depends on the assumption that the number of
tutoring sessions is not correlated with the students’ abilities or motivation.

Farkas argues that this assumption is warranted, because the number of tutoring sessions that students received was determined by “accidents of scheduling” and thus was determined in an essentially random way. His case is strong, because he finds no relationship between the number of tutoring sessions and initial reading ability. Of course, there could still be unobserved differences in potential among students of comparable initial ability. Farkas eliminated one possible difference by having tutors rate students’ levels of concentration in sessions and then controlling for this variable in his estimates. But it is impossible to eliminate all conceivable differences.

For example, one possibility is that frequent absence was associated with both fewer sessions and lower potential, even among those with comparable initial ability. But offsetting this concern is a selection process for participation in the program that may have biased estimates of the program effects in the opposite direction. When regular tutees were unavailable because of accidents of scheduling, alternates were chosen. Farkas notes that these alternates were typically the students who received the fewest number of sessions in the sample. Teachers selected both the regular tutees and the alternates. If teachers used subtle unobserved differences to select those most in need of tutoring as the primary tutees and those somewhat less in need as alternates, the estimated effect size would have been biased downward.

All in all, Farkas’s evaluation design is probably as good as one using a matched comparison group, but clearly not as good as one using a randomly assigned control group. The results presented here are reinforced, however, by related ones. Similar tutoring programs, such as Reading Recovery and Success for All, have also been successful. But, of course, because the unique use of low-cost tutors makes Reading One-to-One potentially more cost-effective than its peers, the relevance of results from the other programs is limited. Reading One-to-One itself has been replicated and upscaled. Farkas has analyzed different sites, cohorts, and time periods, with consistently positive results. Although the relationship between implementation fidelity and effectiveness has not been analyzed directly, the effects have grown over time as sites have gained more experience with the program. This is a good sign. Nevertheless, a good deal more work
needs to be done to prove the concept, all the more so because the program is so extraordinarily promising.

Three separate lines of research, with different time frames, should be established as quickly as possible. The first line of research should be aimed at determining whether Reading One-to-One, in its present form, merits being turned into a national-scale program over the next several years. As Farkas notes in chapter 3, we spend more than five billion dollars a year on educational programs for poor children through Title I, even though Title I is not an effective program. Farkas calculates that if we used all of the Title I money, we could provide sixty Reading One-to-One sessions to ten million children a year (or one hundred sessions to six million students a year). Therefore it would be quite feasible simply to replace Title I with a national Reading One-to-One program. This would be a bold step, but since the current program is ineffective, the risk is relatively small.

Nevertheless, we would naturally want to be as confident about the program’s effectiveness as we reasonably could be before making such a major investment. Therefore, numerous evaluations of different sites and cohorts should be undertaken. As many evaluations as possible should be done by independent evaluators using a pure experimental design. If these evaluations yielded consistently positive results, we would probably have enough evidence in two or three years to justify making a major investment to expand the program.

Even with the best of results, the program should be ratcheted up in steps. Even if the program works extremely well at its present size, it might lose its effectiveness if its scale were increased substantially. It has already grown successfully from a single site to thirty-two schools and twenty-four hundred children at its height. However, that is a long way from being a national-scale program serving millions of children. To reduce the risk of wasting large amounts of money, the program should grow in stages. Evaluations should be done at each stage, and ensuing steps should be taken only if those evaluations indicate that the program still works.

The second and third lines of research should have longer time frames. The second should be made up of long-term studies that follow various cohorts from numerous sites into adulthood. This would enable us to determine whether Reading One-to-One’s effects endure or fade out.

The third line of research should be structured to push the limits of Reading One-to-One’s potential to raise the productivity of the
American educational system. This would involve developing equally inexpensive versions of Reading One-to-One for children in grades four through twelve. Planned variation studies should then be structured to determine how the program’s impact varies with the duration and timing of interventions over the course of twelve years. At the earliest sign of success, we would probably also want to develop a “Math One-to-One” program and do similar planned variation studies on it, if the initial results prove promising.

One could certainly argue that no major investments should be made in Reading One-to-One before results from the long-term studies in the second and third lines of research have generated positive results. After all, numerous other programs have raised academic achievement in the short run only to see those gains fade out. There is a good chance that the second line of research will show that the gains of Reading One-to-One fade out sooner or later. It would be prudent to wait to make major investments in the program, were it not for two facts. First, the low cost of Reading One-to-One offers hope that we can intervene whenever gains begin to fade out. And second, in the plan described here, money would be reallocated from an ineffective program (Title I), so there is little risk of making the current situation worse. This seems like a good risk to take when measured against the opportunity cost of delaying implementation for the ten to fifteen years that it would take for the long-term follow-up studies to yield results.

One reason we can be hopeful that multiple years of intervention can produce long-term gains is that this has been the case with the Chicago public schools’ Child-Parent Center (CPC) and Expansion Program. The CPCs offer a combination of educational enrichment and family support services to children between the ages of three and nine (preschool through third grade). The key components of the program are small class sizes, parental involvement, an emphasis on basic skills, and most notably, the opportunity to receive up to six years of services.

The effects of the CPCs are evaluated by Arthur Reynolds using the Chicago Longitudinal Study in chapter 4. This study included all 1,152 children enrolled in twenty CPC preschools and kindergartens between 1983 and 1986 as well as a comparison group. The children in the comparison group attended different full-day kindergartens at six randomly selected schools in high-poverty areas of Chicago. The study continues to follow both groups as they move through the school system.
Reynolds finds that the CPCs generate long-term gains in academic achievement, but only for children who participated for at least three years. The effects for those children who participated for six years are quite substantial. By eighth grade, the effect for six-year participants was 0.56 standard deviation in reading and 0.50 standard deviation in math. These long-term participants gained about one full grade level in achievement, making up roughly half the gap to the national average. Only 11 percent were held back a grade in school versus 34 percent in the comparison group. They also averaged half a year less in special education and developed more practical life skills.

Gains tended to fade out over time, but only partially. Effect sizes fell between third and fifth grade but stabilized thereafter. There is no way to know yet whether the gains achieved through eighth grade are permanent, but by that point the gains had survived for five years after the final intervention.

The most striking feature of these results is the consistent relationship between the size of gains and the duration of participation. Each extra year of participation after the first tends to raise performance on each one of five measures of achievement. The correlation is not perfect, but we would not expect it to be, given random error. The positive trend is quite strong, particularly when all five measures are considered. This pattern is important for two reasons. It suggests that long-lasting improvements in academic performance can be achieved through sustained intervention, and it lends a great deal of credibility to the results. Thus the program confers tangible benefits that accrue slowly, but consistently, over time.

This pattern is all the more important, because the evaluation did not have a pure experimental design. Reynolds and Temple (1995), who are both independent evaluators, go to great lengths to identify and adjust for any biasing effects that nonrandom selection might have on these findings. The results appear to be robust, but without random assignment it is hard to be completely confident. Also, the significance tests are biased, because they do not adjust for within-school correlations.

Reynolds analyzes the CPC data as if it were all from one medium-size program, in part because it provides him with enough cases and variation with which to assess the relationship between duration of participation and effect size. But the CPCs have essentially been replicated and scaled up. There are twenty sites in Chicago, which have served thousands of children. An analysis that defines each cohort at
each site as a single case (as is done by Slavin and his coauthors in chapter 2) should be carried out to replicate the basic results. It would be preferable for such a study to use different cohorts, so that those findings are completely independent of the ones presented here. That study could also examine the relationship between effect size and implementation fidelity or site experience.

Most programs similar to the CPCs have not yielded long-term gains in academic achievement. This fact begs the question of whether CPC results are those of a random outlier. In chapter 4, Reynolds attempts to identify a number of components of CPCs that could explain their relatively unique success, but each of the elements he cites, such as parental involvement, has been used by other programs that did not generate enduring gains. Therefore, if future studies replicate Reynolds’s results, we would have to ascribe the program’s success to either the unique mix of components or the duration of intervention itself or both.

The CPCs are expensive. They cost $3,600 more than the regular educational program for the primary grades and $4,180 more for kindergarten. Because five or six years of intervention are needed to generate substantial long-term gains, the overall cost is in the range of $18,000 to $22,000. This is mitigated somewhat by reductions in grade retentions and special education. We will not be able to determine whether the benefits outweigh the costs until the current cohorts reach adulthood, although it is quite possible that they will, given the gains in both academic achievement and practical life skills. Offhand, it might seem that the total cost would be politically and financially infeasible. Yet the Chicago public school system has provided this program to thousands of children over many years, so we should not be quick to write off the program as being too costly. Long-term interventions may be necessary to generate enduring academic gains, so it is vital that we follow up with these cohorts to determine whether the investment pays off in the end. However, until we can replicate these results and measure the long-term benefits, it is not advisable to spend large amounts on upscaling this program.

Although sustained intervention may be necessary to raise academic achievement, results from evaluations of the High/Scope Perry Preschool Program suggest that the “inoculation” approach may be extremely effective in dealing with other important social problems. The Perry Preschool is described by Lawrence Schweinhart and David Weikart in chapter 5. It is, in many senses of the word, the mother of
all social programs. Initiated in 1962, it was one of the first rigorously planned social experiments. Its evaluation was one of the first to use a pure experimental design. It was a model for Head Start and, in a looser sense, for much social policy and experimentation that evolved over the ensuing decades. The consistency and duration of follow-up studies of the long-term effects of the program are unparalleled. In fact, the participants and controls are still being followed today, more than thirty years after the experimental treatment ended. Thus it offers as much or more evidence on the question of the permanence of effects than any program ever has. The evaluation’s only real weakness is that it was not carried out by independent analysts.

The program itself was the archetypal early intervention program for preschoolers. Three- and four-year-old children participated for one or two school years. The program applied classic principles of child development to create a nurturing environment fostering social, emotional, and cognitive growth. Student-teacher ratios were very low, with four teachers in each class of twenty to twenty-five children. Teachers made weekly ninety-minute home visits to every family. All teachers were certified in both early childhood and special education.

In the early years after the treatment was completed, the program was deemed a failure, because most of the early academic gains faded out in the later grades. But long-term follow-ups revealed successes in other areas.

The most recent data were generated when the participants were twenty-seven years old. By that age, there were a number of large effects, several of which were truly dramatic. The program quadrupled the proportion of individuals who were earning more than two thousand dollars a month, from 7 to 29 percent. It almost tripled home ownership, raising it from 13 to 36 percent. It reduced the proportion who had ever received welfare by a quarter, from 80 to 59 percent. And in what is probably the single most extraordinary gain ever achieved by a social program, it reduced the rate of hard-core criminality (defined as having five or more arrests) by four-fifths, from 35 to 7 percent. For males, the reduction was from 49 to 12 percent. There were no significant differences between children who spent one year in the program and those who spent two years, so time of participation had neither a positive nor a negative influence.

The program was expensive, but cost-effective. The total cost was $12,356 per participant in constant 1992 dollars, but the savings to the public, in terms of lower taxes and reduced crime, was $88,433
(assuming an annual discount rate of 3 percent). In other words, the program saved the general public $7.16 for each dollar invested. And this does not even take into account any psychological benefits that accrued to the participants and their families, which common sense tells us were huge.

If these results are replicable, the policy prescription is clear. We should turn Head Start into a copy of the Perry Preschool Program and make it available to every poor child. This would involve a substantial up-front investment that would pay for itself many times over by significantly reducing the incidence of some of America's worst social problems down the road.

The problem is that we cannot yet be sure that the results are replicable, and we need to have a high degree of confidence to justify a multibillion dollar investment. Despite the high quality of the evaluation design, a single study is never enough to be sure of a program's efficacy. And because the program was done at a single site, we do not know whether it would work as well for poor children in other places and situations or whether quality would suffer in the process of upscaling it. The lack of replication also means that there is no information on the relationship between effect size and either implementation fidelity or site experience.

Recently, however, Schweinhart and Weikart (1997) have published results from one near replication, as part of the High/Scope Preschool Curriculum Comparison. In that study, children were randomly assigned to three preschools, each with a different curriculum design. One was based on the High/Scope (Perry Preschool) model, a second was a more intensive academic program called Direct Instruction, and a third was a traditional nursery school. In short, the study was structured specifically to determine whether the unique design of the Perry Preschool was responsible for its outcomes. Unfortunately, although the original program was replicated, the original evaluation cannot be. There was no pure control group of individuals who did not receive any treatment at all.

The basic result was that by age twenty-three, the High/Scope group had better outcomes on a number of social (but not academic) measures than the Direct Instruction group. As with the original Perry Preschool children, the largest and most consistent effects were in the area of crime and delinquency. For example, the High/Scope individuals had less than half the total number of arrests. And just 10 percent of the High/Scope group had been arrested for a felony compared
with 39 percent for the Direct Instruction group. But there was no clear pattern of statistically significant differences between the High/Scope group and the traditional nursery school group, with respect to crime or any other outcomes. Therefore, given the absence of a pure control group, it is hard to know whether the High/Scope and traditional nursery schools reduced crime, whether the Direct Instruction Program increased it, or both. As it happens, the arrest rates at age twenty-three were roughly the same for the original Perry Preschool group as for the new High/Scope group. They were also roughly the same for the no-treatment control group in the original Perry study and the new Direct Instruction group. These facts hint that the later results may reflect improvement in the High/Scope and traditional nursery groups rather than deterioration in the Direct Instruction group. However, there is no sound methodological basis for making such comparisons.

The similarity in outcomes between the High/Scope and the traditional nursery school groups is also an important result. On the one hand, it is bad news in the sense that it calls into question whether the unique features of the Perry Preschool model add any value at all above and beyond what traditional nursery schools provide. On the other hand, it is good news in the sense that it suggests that any preschool offering a high-quality program emphasizing social and emotional development may reduce crime and delinquency among individuals who grow up poor.

A major research effort is needed to determine the effects of various types of preschools on crime and delinquency over the life cycle. Part of this research should be devoted to developing new studies similar to the High/Scope Preschool Curriculum Comparison (but with a no-treatment control group in which parents chose between the options available in the general community). But we cannot rely on this approach exclusively. It takes too long for the major potential benefits of such programs to manifest themselves. Fortunately, both the first and the second studies find that reductions in early adolescent misconduct presaged later reductions in crime and delinquency. Even so, if we started now, it would take a minimum of ten years to get any clear indication of effectiveness.

Thus another part of this research effort should be aimed at following up on adolescents and adults who were in other preschool studies. Numerous evaluations of various kinds of preschool programs were conducted over the past thirty years. Unfortunately, most
of them focused on academic outcomes. Follow-up studies that focus on criminality and other behavioral outcomes in adolescence and adulthood are needed on as many of these programs as possible. Even data from projects with poor evaluation designs would be of some value in meta-analyses. If such a research effort reveals a strong relationship between certain types of preschool and crime, a large-scale investment in early intervention programs might be justified. Depending on the nature of the results, this might take the form of expanding, standardizing, or intensifying Head Start.

The Perry Preschool was originally considered a disappointment because most of the early cognitive benefits it generated faded out over time. This is a consistent pattern in early intervention programs. They typically raise IQ (intelligence quotient) and achievement test scores by large amounts, but these gains disappear by the third or fourth grade.

The Abecedarian (ABC) Project, described by Craig Ramey, Frances Campbell, and Clancy Blair in chapter 6, is an exception. It increased cognitive test scores, and these increases held up through the latest evaluation, at age fifteen. ABC is somewhat unusual in that it intervened extremely early, at six weeks of age. Children who were randomly assigned to the control group received nutritional supplements and screening for developmental problems. Children who were randomly assigned to the treatment group received these services and were enrolled in an intensive educational day care program. On reaching kindergarten age, half of the treatment and control groups were randomly assigned to a follow-up home enrichment program that lasted for an additional three years. The other half received no further treatment and began attending public school kindergarten. Because time spent in the program and the age of intervention are confounded, there is no way to determine the relationship between effect size and duration of participation.

Seven to ten years after the final treatment, at age fifteen, those children who attended the preschool program had IQs that were 0.35 standard deviation (4.6 points) higher than those who did not. This difference is not large, but it is not trivial either. The program’s overall effect on achievement test scores was comparable. The effect on the IQs of the children of mothers with very low IQs was quite large (11.7 points), but there were just six such cases each in the treatment and control groups. The program reduced grade retentions by about 40 percent and special education placements by half. The follow-up
program had no effect on IQ or mathematics scores, but it did raise reading achievement by a small amount.

These findings are promising, but there are reasons to be cautious. The evaluation was not done by independent researchers, although the design was excellent. The program has not been formally upscaled or replicated. Because of the lack of replication, there is no information on the relationship between effect size and implementation fidelity or site experience. The most important reason for pause is that similar early intervention programs have not had consistent long-term effects on cognitive test scores (although they have reduced grade retentions and special education placements; Consortium for Longitudinal Studies 1983). It is possible that the Abecedarian Project is simply one of two random outliers (along with the Milwaukee Project, Garber 1988). One factor that may distinguish ABC as an “orange” in this distribution of apples is its initiation of intervention in early infancy. Most of the IQ difference between treatment and controls was already present by six months of age.

Even if ABC’s results are replicable, the program may not be cost-effective on the basis of cognitive gains alone. The annual cost of the program is six thousand dollars per child. Because the elementary school intervention had virtually no impact, the effective part of the program involved five years of intervention, for a total cost of thirty thousand dollars. Even if the cognitive gains hold up through adulthood, they would not justify that kind of investment, unless they spin off other positive outcomes. We should not rush out and spent lots of money replicating ABC on a larger scale, although it may hold some promise as a medium-scale program targeted specifically to mothers with very low IQs.

The next follow-up is scheduled to be done when the individuals are twenty-one years old. At that point, it will be important to determine whether economic and behavioral outcomes similar to those generated by the Perry Preschool have manifested themselves. If so, ABC might turn out to be highly cost-effective. Unfortunately, results from the early follow-ups on childhood misconduct show that the treatment group exhibited more problem behavior than the controls. So it may be that ABC had too strong an academic orientation, like the Direct Instruction model.

Regardless of whether we would ever want to replicate ABC on a large scale, the program is important because it suggests that we should not completely abandon the hope of inoculating children against cognitive deficits through early intervention. The results are
favorable enough to justify further research. In particular, we should continue examining the cognitive effects of very early intervention in infancy and the effects of intervention on the children of mothers with very low IQs. If we can identify the particular elements of the Abecedarian Project that generate the effects, we may be able to develop a program that can deliver the same impact for a fraction of the cost or a larger impact for the same price.

The programs discussed so far have all been small to medium scale. But at least one national-scale federal program can offer evidence of both substantial benefits and cost-effectiveness: the Special Supplemental Nutrition Program for Women, Infants, and Children, better known as WIC, which is analyzed by Barbara Devaney in chapter 7. The WIC program serves low-income pregnant women and mothers as well as children under the age of five. It provides its clients with vouchers for food supplements, nutrition education, and referrals to health care and social service providers. WIC is one of the largest social programs in existence. In fiscal 1993, it served almost six million people and cost 2.8 billion dollars.

WIC appears to have substantial impacts on gestational age and birthweight, outcomes that are associated with various health risks. Devaney, an independent evaluator, finds that the incidence of preterm births was 18 to 31 percent lower for the WIC participants. The incidence of low birthweight was 22 to 31 percent lower. And the average birthweight was 51 to 117 grams higher for WIC children.

WIC also seems to save more money than it costs. And unlike some other programs in which the benefits do not start to accrue until years after the investments are made, WIC appears to be cost-effective almost from the beginning. Devaney finds that in the first sixty days after birth, the Medicaid costs of WIC newborns and mothers were $277 to $598 lower than those of newborns and mothers who did not participate in WIC. Thus the program saved between $1.77 and $3.13 in Medicaid for every dollar spent on the program. Savings were even larger for a subset of infants whose mothers were not eligible for Medicaid during pregnancy, but who became eligible after birth because the newborns suffered health problems requiring expensive medical care.

We need to be cautious about these results. They were not generated from a controlled experiment, but rather from a comparison between WIC participants and nonparticipants on Medicaid. Because nonparticipants were eligible for WIC, selection was based largely on choice and access. Such nonrandom selection may bias the results, al-
though it is not clear in which direction. To the extent that WIC mothers were volunteers who were motivated to maintain the health of themselves and their children, the effects of WIC are probably overstated. To the extent that WIC tends to target the highest-risk mothers, the effects are probably understated. WIC participants do tend to be more disadvantaged than eligible nonparticipants in comparison groups.

Since WIC is a national program, there are no replications of it or anything remotely comparable to it in terms of scale. It is so unique that it does not fit into any distribution of programs. It has essentially been almost fully scaled up already. Although the program has not been replicated, Buescher and others (1993) applied similar methods to North Carolina data and generated results comparable to Devaney’s. There is no information on the relationship between effect size and implementation fidelity or duration. There is some information in other studies on time of participation, but it is confounded by the child’s age at intervention, so its relationship to effect size cannot be distinguished (Rush and others 1988).

A multisite analysis could shed light on these issues. Devaney does this to a limited extent by conducting separate analyses for each of five states. She finds a positive effect on each of five outcomes in five states, with twenty-three of the twenty-five results statistically significant and one other nearly so. This consistency lends credence to the overall findings. However, if selection biases the effects upward, it could very well affect every outcome in every state.

Because WIC is an ongoing national program, it will be hard to design an evaluation that is free of selection bias. Nevertheless, further research using quasi-experimental methodologies is warranted, because the program is so large. If Devaney’s results are valid, there is a least one large-scale federal social program that really works. And this would be true, even though the cost-benefit analysis completely omitted any possible medium- and long-term medical and social gains. Taking these into account, it is quite possible that the benefit-cost ratio of WIC could be much higher than Devaney’s upper bound of 3:1.

Even many people who believe that social programs can work if they reach children early, like WIC does, are all too ready to write off older children and adolescents. This is particularly true with respect to delinquency and drug abuse. Yet there is hope that these terrible problems are not necessarily intractable, even if we get a late start.
Two drug prevention programs offer sound evidence that they reduce rates of substance use and abuse among adolescents. Phyllis Ellickson describes one of these programs, Project ALERT, in chapter 8. Project ALERT is a module of eleven classes offered to seventh graders and a booster set of three sessions to eighth graders. The curriculum is designed to increase students’ motivation to abstain from drug use and to give them the social skills needed to resist peer pressure.

Relative to students in randomly assigned control schools, Project ALERT lowered the proportion of students who started using marijuana in junior high school by about 30 percent. It also substantially reduced the frequency of both marijuana and cigarette smoking among individuals who had already been using the drugs. Effect sizes were roughly the same at schools with high percentages of minorities as they were at largely white schools. The evaluation design was based on random assignment of schools, but not of individuals. Significance tests were adjusted for within-school correlations.

There are a number of reasons to be cautious about the results. There was one anomalous finding. The most committed smokers in the treatment schools actually smoked more than their counterparts in control schools. This weakens our confidence in the positive results obtained. The evaluator was also the developer of the program. The results have not yet been replicated, and the program has not been upscaled. Thus there is no information on the relationship between effect size and implementation fidelity or site experience. There are also no data on the association between effects and time of participation. And perhaps most important, all of the effects faded away during high school.

There are also reasons to be hopeful. Project ALERT is by no means unique. A number of similar programs have also reduced various kinds of drug use in the short run (Botvin, chapter 9 in this volume; Murray and others 1984; Perry and others 1980). And the fade-out of effects does not make the program a failure, by any means, because the costs of treatment are trivial. In chapter 8, Ellickson estimates that the entire cost of the program is $1.50 per student, assuming that teaching time is drawn from other drug prevention uses. Even if we impute the value of the teaching time, the total cost of the program is probably no more than thirty-five dollars per student. The fact that such a low-intensity, inexpensive program may have reduced several types of drug use by fairly large amounts for two years is quite extraordinary. It is very likely that even the short-run medical and social benefits of the reduced drug
use amount to more than thirty-five dollars per person. More impor-
tant, the amazing “bang for the buck” that the program yields suggests
the possibility that permanent results might be achieved either by in-
creasing its intensity or by offering additional boosters in later years.
The Life Skills Training (LST) Program, presented by Gilbert Botvin in
chapter 9, offers a chance to test that hypothesis.

LST is similar to Project ALERT, but it provides twice as many ses-
sions and an additional year of boosters. It consists of fifteen classes
of forty-five minutes each in the seventh grade and boosters of ten
classes in the eighth grade and five in the ninth grade. Its curriculum
covers drug information, self-management skills, general social skills,
and social resistance skills.

LST has been replicated many times over and implemented on a
medium scale. There have been numerous evaluations of it, involving
several thousand students. These evaluations consistently demon-
strate large, and in some cases huge, short-term reductions in both
overall use and heavy use of cigarettes, alcohol, marijuana, and hard
drugs. Scaling the program up reduced the effect size a bit, but the im-
pacts were still large. There was some tendency for effects to fade out
over the long term, but at least some effects remained three years after
the last boosters. At the end of high school, the program reduced cig-
arette smoking by more than 20 percent. It lowered the incidence of
drunkenness by about 15 percent. It reduced weekly marijuana use
by 33 percent, but the effect was not significant. There were small re-
ductions in casual use of alcohol and marijuana, but they were in-
consistent and not significant. Every combination of multiple use of
these three drugs was lower in each of two treatment groups, but only
about half of these effects were significant. Implementation fidelity
was systematically measured in a random sample of classes. Effects
were consistently larger for students who received treatments that
were more faithful to the program design. Of the thirty-four long-term
effects measured for the high-fidelity sample, twenty-nine were sig-
nificant, and the other five were all positive (Botvin et al. 1995).

There are some problems with the LST evaluations. All of them
were done by Botvin, the program’s developer. They were based on
random assignment, but with schools as the unit of assignment and
students as the unit of analysis. Because no adjustments were made
for within-school correlations, the tests of significance are biased.
Also, long-term results are not yet available for samples with high per-
centages of minorities.
However, there are factors that lend credence to the results. Project ALERT and similar programs have generated similar results, so LST is not an outlier. LST is distinct from Project ALERT in offering more sessions over more years. The fact that it produces larger effects suggests that there may be a relationship between program intensity and effect size. There is no analysis of the relationship between site experience and effect size.

By far the most impressive aspect of the program is the extraordinary relationship between cost and long-term effects. Including the value of teachers’ time, the entire program probably costs less than seventy-five dollars per student. Yet it generates substantial long-term reductions in various kinds of drug use. Although no formal analysis of the value of such reductions has been performed, the benefit-cost ratio is probably very large.

Some work remains to be done. First, the tests of significance in the studies need to be adjusted for within-school correlations. Second, it would be nice to have at least one evaluation in which students were the unit of assignment as well as the unit of analysis. Third, we need long-term results from at least one study in a population with a substantial proportion of minorities. And fourth, a formal cost-benefit analysis should be done.

If all of those analyses corroborate the findings presented here, ratcheting up the scale of LST would be justified, with the goal of turning it into a national program over the course of several years. Given the low cost of the program, the risks are small relative to the huge potential benefits.

While conducting further analysis of the basic LST program, we should also be developing and piloting new, more intensive versions. Versions that offer three or five or even ten times as many sessions over several years would still be relatively cheap. Given the record of the current version, these “super-LSTs” could potentially yield huge long-term reductions in drug use. Of course, there are no guarantees that increasing the intensity would enlarge the impact. But the findings to date offer ample justification for investing in a planned variation study that would compare several versions of different intensity levels to each other and to a “no-treatment” control group.

The success of the drug programs with heavy users offers hope that social programs can do some good even if intervention begins after people have developed serious social problems. The Treatment Foster Care (TFC) Program, described by Patricia Chamberlain and
Kevin Moore in chapter 10, suggests that even chronic juvenile delinquents can be helped. TFC placed eighty delinquent boys between the ages of twelve and seventeen into foster families recruited from the community. The boys had been ordered into residential treatment by a juvenile court and randomly assigned to either a group home or TFC. The average age of the boys was 14.3 years, and their mean number of arrests prior to beginning treatment was thirteen. The program trained the foster parents in the use of behavior management skills and monitored them closely throughout the program. It emphasized the importance of preventing contact between the boy and any delinquent peers. Only one boy was placed with each particular TFC family. The average treatment period was seven months in both TFC and group care.

TFC had very large effects both during the treatment period and one year after it ended. Six months into treatment, arrest rates were more than two-thirds lower for the TFC group. There was some fade-out of the effect during the post-treatment year. But at the one-year follow-up, TFC arrest rates were still 52 percent lower. Most of this gain was attributable to the relatively large proportion of TFC boys who stayed completely “clean”: 41 percent of the TFC boys had no arrests by the follow-up compared with only 7 percent of the group-care teens, an extraordinary ratio of almost 6:1. TFC tripled the proportion of boys who returned to their original families and decreased the number of days of incarceration by 59 percent (from 129 to 53) during the follow-up period.

The average cost of TFC was $18,620 for a full course of treatment. This was 29 to 47 percent lower than the cost of group care. So, with substantially lower costs and much larger benefits, the program is clearly superior to group care. Because TFC was not compared to a control group that received no treatment at all, we do not know if TFC is cost-effective relative to doing nothing for the boys. But given its large advantages over group care and the high costs of arrest, trial, and incarceration, TFC is an extremely promising program.

The version of TFC evaluated here was itself a replication. The first two pilots were also successful, which lends credence to the results. But all three trials were small, so we do not know if quality would suffer if the program were enlarged. The approach of the program is fairly unique, so it is clearly not a random outlier in a large distribution. Of course, this means that there are no corroborating results from similar programs. We do not have any information on the relationship be-
tween site experience or fidelity of implementation and effect size. We also do not know if longer or more intense treatment would improve the results. Overall, the design of the evaluation was quite good, although the evaluators were also the program developers.

A good deal of additional research on the program needs to be done. A formal cost-benefit analysis should be carried out. The boys from the original program should be followed to determine how long the effects last. A long-term study could also determine whether any other benefits, such as gains in education or employment, emerge. And, of course, the program should be replicated at other sites with larger samples and different populations. Some of these evaluations should use a “no-treatment” control group (which would probably necessitate applying a selection process other than court mandate). Clearly, it is too early to recommend that this program be implemented on a large scale. Nevertheless, TFC is a tremendously exciting program that offers hope for teenagers who currently seem out of reach.

If early intervention programs occupy one end of the spectrum of social programs, work and training programs occupy the other. They attempt to intervene late in the game, in adulthood or late adolescence. There is a long history of experiments with and evaluations of work and training programs. These programs have generated a set of results that are consistent enough to have created a fairly widespread consensus among scholars with regard to their potential.

On the one hand, the best programs can be cost-effective. On the other hand, their effects are small. By investing a small amount of money per participant, it is possible to increase the amount that the individuals work, to reduce their welfare dependence, and to raise their earnings by enough to justify the cost of the program. However, the earnings gains typically amount to just several hundred dollars a year. Increasing the intensity (and cost) of the program beyond a certain point does not increase the size of those gains (Friedlander and Gueron 1992). No matter how much is invested, these programs cannot come close to making a typical welfare recipient financially self-sufficient. The consensus is that work and training programs cannot possibly move large numbers of people off welfare and into jobs in the private labor market.

Lawrence Mead challenges this consensus in chapter 11. Chapter 11 is unique in this volume in that it does not present an evaluation or a meta-analysis of a single program. Instead it reinterprets the body of evidence generated from all research on one type of program. Mead
contends that a combination of good training programs and stringent work requirements in welfare programs can reduce welfare dependence. He supports this contention in three ways. First, he argues that the results from work and training evaluations understate the potential of the approach, for technical reasons. Second, he cites results from recent programs suggesting that we are learning how to generate larger gains. And third, he presents data suggesting that the combination of strict work requirements and good programs has reduced welfare caseloads by substantial amounts in various localities.

Some of Mead’s technical arguments have merit, but even if he is right on every count, the overall picture remains the same. It is unlikely that all of these factors together could have led to underestimates of potential income gains from work and training programs by more than five hundred dollars per year and certainly not by more than a thousand dollars. Even at the upper end of that range, the programs would not make most welfare recipients financially self-sufficient.

The recent successes that Mead cites clearly offer hope. He argues that there was a clear pattern of learning and improvement over time in California. The Saturation Work Initiative Model in San Diego improved on an earlier generation of programs by enforcing participation more aggressively. Then the Greater Avenues for Independence (GAIN) in Riverside got even better results by emphasizing job search. Indeed, Riverside GAIN is probably the most successful work and training program ever done. It increased earnings by 40 percent. Two other sites using the same basic “labor force attachment” strategy also achieved unusually good results. So there is reason to hope that the unique features of that strategy make it genuinely superior. If so, work and training programs should start producing better results by using this approach. Nevertheless, even Riverside GAIN, which is the most successful program on record, increased earnings by just $1,010 a year. So if the goal is to make families self-sufficient, we clearly have a long way to go up the learning curve.

Mead also presents evidence that a combination of well-run programs and strict work requirements can reduce welfare caseloads by substantial amounts. In particular, he cites data from Wisconsin, which established programs with strict work requirements and saw its caseload fall by half between 1987 and 1996. Given the effect sizes of work and training programs, the larger declines could only have been the result of a diversion effect, that is, people getting thrown off
welfare, leaving it voluntarily, or choosing not to go on it in the first place because of the work requirements.

Unfortunately, there is no way to be sure whether the caseload declines were actually the result of the programs and the requirements, for a number of reasons. No experimental evaluation has been done. Much of the decrease in some of the counties occurred before the particular programs or requirements were instituted. And Wisconsin's economy was growing strongly during most of the period, while welfare benefits were being cut. But Mead's statistical analyses suggest that these factors cannot account for the large declines in caseloads (Mead 1997).

We need to do more rigorous evaluations of the diversion effect of work requirements. Mead's evidence suggests that the effect may well be substantial. But this evidence is by no means strong enough, in and of itself, to justify major policy decisions.

Even if all of the reduction in Wisconsin's caseloads could be attributed to the programs and work requirements, we still could not use them as stand-alone alternatives to a guaranteed benefit program, without decreasing incomes of the poor by a large amount. Unfortunately, this is essentially what our new welfare system, enacted by the Personal Responsibility and Work Opportunity Reconciliation Act of 1996 (PRWORA), does. Thus it seems very likely that this legislation will lead to large losses of income among the poorest Americans and perhaps a catastrophic rise in abject poverty.

The reason for this is that the numbers just do not add up. The 1996 legislation has "ended welfare as we know it" in the simplest possible way—by ending it, without offering any viable alternative. Although this legislation has a number of components, some good and some bad, ultimately they all pale in comparison to the provision that places a five-year lifetime limit on welfare benefits. This provision is implicitly predicated on the assumption that given strong incentives and enough time, we can prepare 100 percent of welfare recipients to become self-sufficient in the private labor market. But every bit of research ever done on welfare and training programs clearly suggests that nothing could be further from the truth. A large proportion of welfare recipients, probably somewhere in the range of 25 to 50 percent, are not capable of sustaining regular employment in a competitive labor market. There is absolutely no evidence that any amount of training can change that fact. And although the most optimistic interpretation of Mead's evidence suggests that work require-
ments may divert large numbers of recipients away from welfare, it
does nothing to offer hope for the substantial portions that would be
left without any alternative sources of income.

Mead suggests that friends and relatives and other programs can
pick up the slack. With respect to friends and relatives, only time will
tell. With respect to programs, current alternatives are clearly inade-
quate. State-funded general assistance programs are the main alter-
native to federal welfare programs. None of them offers benefits even
remotely close to what is necessary to support a family. Traditionally,
individuals who have depended on these programs for support have
been single men who are either homeless or living in single-room-
occupancy hotels. Recent cuts in these programs have created sub-
stantial hardship among that population (Danziger and Kossoudji
1994/95).

PRWORA does allow states to exempt 20 percent of its welfare pop-
ulation from the five-year time limit. This will no doubt help to pre-
vent some suffering. But not all states will use this exemption, and
even the ones that do will not be able to ensure that every one of those
spots will go to people who really need them.

Mead's analysis in chapter 11 offers some hope of a viable compro-
mise between the new system and the old. This compromise involves
the resurrection of an old reform proposal—workfare. In workfare,
welfare recipients have to work to receive their benefits. This is radic-
ally different from the five-year time limit, because it guarantees all
recipients a job.

PRWORA will almost inevitably appear to be a great success on the
surface. There is no doubt that caseloads and welfare costs will fall
dramatically. It is easy to cut caseloads and costs simply by kicking
people out of the program. Another result will be that work will in-
crease among the poor, which will probably be the greatest legitimate
benefit of the new system. At the same time, however, PRWORA will
almost certainly reduce the income of the poor substantially. And it
may consign large numbers of American families to abject poverty
without any choice at all or, in some cases, perhaps with a choice to
remain in abusive domestic situations. In evaluating the success of the
new system, it will be vital to weigh the lost income and other hard-
ships suffered by the poor against the governmental savings and the
increases in work.

Workfare would offer most of the upside of PRWORA, without
most of the downside. It would provide a reasonable alternative to
every adult welfare recipient, while dealing with the one complaint that the overwhelming majority of Americans have with the old system—that people can receive money without working. Critics have argued that workfare is too expensive. Administering work programs certainly costs more per participant than simply writing benefit checks. It may well be that most Americans are willing to pay that price in order to balance their concern for the poor with their resentment of idleness. The most important contribution of chapter 11 is its suggestion that we might not have to pay that price. If work requirements have a strong diversionary effect, then workfare might not be more expensive overall than welfare, because it would cut caseloads substantially. It would steer the more capable people into the private labor market, while guaranteeing a public job and a living wage to those who need that guarantee.

SUMMARY AND POLICY IMPLICATIONS

Table 1.1 encapsulates the analysis presented here. It summarizes how each of the programs evaluated in this volume stack up when judged according to the thirteen criteria.

Only a small percentage of social programs offer convincing evidence that they generate both substantial effects and benefits that exceed costs (Crane, in preparation). This conclusion has important implications for social policy. It suggests that a decentralized approach to program development, which is increasingly our de facto national policy as the federal government continues to hand off responsibilities to the states, is sensible in the early stages of program development. If only a small fraction of the approaches that are tried work, many different approaches need to be tried to yield even a single success. However, it also suggests that once a successful model has been developed, it is probably necessary to recentralize control over the implementation process. If the overall success rate is low, using social programs as an instrument of social policy will never be cost-effective unless the very best programs are replicated on a large scale.

Title I illustrates this principle. Although federally funded, it is a highly decentralized program. Every school chooses what programs to use to help disadvantaged students. Because only a small percentage of the programs used actually work, the overall impact of Title I is virtually zero. As long as we had no alternatives, this approach served a valuable purpose. It enabled us to try lots of different methods and, ultimately, provided funds that contributed to the develop-
Table 1.1 Performance of Programs Judged According to Thirteen Criteria

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Success for All</th>
<th>Reading One-to-One</th>
<th>Chicago Child-Parent Centers</th>
<th>High/Scope Perry Preschool</th>
<th>Abecedarian Project</th>
<th>WIC&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Project Alert</th>
<th>Life Skills Training</th>
<th>Treatment Foster Care</th>
</tr>
</thead>
<tbody>
<tr>
<td>Key outcomes are significant&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Effect sizes for key outcomes&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Substantial</td>
<td>Substantial</td>
<td>Substantial</td>
<td>Substantial</td>
<td>Substantial</td>
<td>Substantial</td>
<td>Substantial</td>
<td>Substantial</td>
<td>Substantial</td>
</tr>
<tr>
<td>Cost-benefit relationship&lt;sup&gt;d&lt;/sup&gt;</td>
<td>Uncertain</td>
<td>Promising</td>
<td>Uncertain</td>
<td>Excellent</td>
<td>Unpromising</td>
<td>Good</td>
<td>Promising</td>
<td>Promising</td>
<td>Promising</td>
</tr>
<tr>
<td>Long-term effects&lt;sup&gt;e&lt;/sup&gt;</td>
<td>2 years +</td>
<td>—</td>
<td>5 years +</td>
<td>22 years +</td>
<td>7 years +</td>
<td>—</td>
<td>No</td>
<td>3 years +</td>
<td>—</td>
</tr>
<tr>
<td>Evaluation design&lt;sup&gt;f&lt;/sup&gt;</td>
<td>Good</td>
<td>Good</td>
<td>Good</td>
<td>Excellent</td>
<td>Excellent</td>
<td>Good</td>
<td>Good</td>
<td>Good</td>
<td>Excellent</td>
</tr>
<tr>
<td>Independent evaluators&lt;sup&gt;g&lt;/sup&gt;</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Successful replications&lt;sup&gt;h&lt;/sup&gt;</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>In progress</td>
<td>Yes</td>
</tr>
<tr>
<td>Success of similar programs&lt;sup&gt;i&lt;/sup&gt;</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Unique qualities explain success&lt;sup&gt;j&lt;/sup&gt;</td>
<td>—</td>
<td>Probably</td>
<td>Probably not</td>
<td>Probably not</td>
<td>Probably not</td>
<td>Maybe</td>
<td>—</td>
<td>—</td>
<td>Probably</td>
</tr>
<tr>
<td>More of program is better</td>
<td>—</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>Maybe</td>
<td>—</td>
</tr>
<tr>
<td>--------------------------</td>
<td>---</td>
<td>-----</td>
<td>-----</td>
<td>----</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>------</td>
<td>---</td>
</tr>
<tr>
<td>Implementation fidelity helps</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>Yes</td>
<td>—</td>
</tr>
<tr>
<td>Site experience helps</td>
<td>Yes</td>
<td>Yes</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Program has been upscaled</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

— Not available in all cases.

* Special Supplemental Nutrition Program for Women and Infant Children.

* Yes was the minimum standard for inclusion in this volume.

* Effect sizes were given as a rating or a range, and the minimum standard for inclusion in the volume was substantial (other categories were large and huge).

* If a formal cost-benefit analysis was done, the categories were good or excellent (none was worse than good); if no formal analysis was done, the categories were promising, uncertain, or unpromising.

* The number of years indicates the period after the last treatment for which effects remain significant, and a plus sign indicates that the effects were significant at the time of follow-up.

* The categories were excellent or good (none was worse than good); this does not reflect the quality of the work of the evaluators so much as the practical opportunities for developing a control or comparison group.

* Yes means that the evaluators were not involved in program development or implementation; no means that they were.

* Yes means that at least one replication was implemented and an evaluation of it, as an independent experiment, successfully generated results similar to those of the original program; in progress means that such a replication is in progress; no means that no successful replications have been documented.

* Yes means that most similar programs have yielded similar results; no means that they have not.

* Applicable if the program yields at least one result that is different from those of similar programs; the categories are probably (if the unique results seem clearly linked to a unique feature of the program), probably not (if there does not appear to be a link between the unique results and a feature of the program), and maybe (if there is a possible link).

* Yes means that the time of participation or program intensity was significantly correlated with effect size; no means that the time of participation and program intensity were not significantly correlated with effect size; maybe, in the case of Life Skills Training, means that although no formal analysis was done, the fact that the program provided more sessions than Project ALERT may explain why its effects lasted longer.

* Yes means that implementation fidelity was significantly correlated with effect size; no means that it was not.

* Yes means that size experience was significantly correlated with effect size; no means that it was not.

* Yes means that the program increased in size from one to at least ten sites; no means that it did not.
ment of good programs like Success for All and Reading One-to-One. As we become more and more convinced that at least one of these programs can work on a large scale, continuing local control over program selection in Title I will make less and less sense.

Of the nine programs evaluated here, WIC is the only national program, and it does appear to work. None of the others offers “plug and play” solutions that can be applied immediately on a national scale, although two seem to have that potential. Three others are also promising, but they need substantially more confirmation of their effectiveness. Three programs have less potential to be scaled up dramatically, mainly because of their high costs. They are important nonetheless, because they demonstrate that certain outcomes appear to be attainable. In each of these cases, further evidence of long-term benefits could change the prognosis.

The two programs that are closest to being ready for large-scale application are Reading One-to-One and the Life Skills Training drug prevention program. Both of them have had large effects in evaluations with quasi-experimental designs. Both of them are extremely cheap in terms of cost per participant per year. Both have been implemented on a medium scale, having served thousands of students. And in both cases, similar programs, like Success for All and Project ALERT, have proven successful using similar strategies.

However, neither one of them has been evaluated in a study using an experimental design in which individuals are the unit of assignment. If either one of these proves successful in such a study, and if some of the smaller issues are adequately addressed, it should be turned into a national program, serving millions of people. That growth process should be carried out in steps, and it should only be continued as long as evaluations show that the program remains successful at each step.

The three programs that are also promising but further away are Project ALERT, Treatment Foster Care, and the Perry Preschool. Project ALERT is similar to LST, but at an earlier stage of development. We would not need national versions of both. The new version of Project Alert, with a booster session in high school, bears watching and could prove even more effective than LST.

Treatment Foster Care appears to have the potential to reduce crime and delinquency by large amounts, but it is too early in the testing process to be fully confident of its impressive results. We need to determine how long its effects last and to conduct a cost-benefit analysis. If the analysis yields positive results, a multisite replication
should be done. Only if that proves successful could we begin to think about developing it into a large-scale program.

The Perry Preschool is a conundrum. It has as much or more potential than any social program anywhere. If its long-term effects on crime, employment, earnings, and welfare receipt could be replicated on a national scale, some of the nation's worst social problems would be dramatically reduced. However, we simply do not know if such replication is possible. Thus, it is absolutely vital that we do as many long-term follow-up studies of other preschool programs as possible. This opportunity is far too important to waste.

Success for All, the Chicago Child-Parent Centers, and the Abecedarian Project may be too expensive to be turned into national programs, at least given our current state of knowledge about the benefits. Nonetheless, each makes an important contribution to the state of knowledge in the field, which could lead to the development of programs that have national potential. Success for All has demonstrated that at least one technique (one-on-one tutoring) increases the amount that children learn. If Reading One-to-One is successful on a national scale, it will be because it found a cheaper way to do what Success for All did first. This does not mean that we should abandon Success for All. Long-term follow-up studies may reveal that it yields greater benefits than its cheaper cousin, particularly for the poorest readers. These benefits may ultimately justify the high cost. Success for All has also given birth to Roots and Wings, a tremendously promising model of school reform.

The Chicago Child-Parent Centers and the Abecedarian Project have demonstrated the possibility of generating substantial long-term academic gains if the intervention lasts long enough, starts early enough, or is intensive enough. Unfortunately, the costs of these programs are too high to justify expanding them, given our current knowledge about their benefits. However, this could change, if the benefits increase substantially over the long term.

With respect to the work and training programs discussed in chapter 11, current research clearly indicates that we should not invest the money needed to run them on a national scale. Unfortunately, the new welfare legislation specifically mandates that we do just that. Mead's analysis suggests that work requirements may divert large numbers of people away from the welfare system. If that is true, the cost of guaranteeing people public jobs, rather than just kicking them off the rolls, might not be prohibitive.
In sum, we, as a nation, can benefit from social programs if we use rigorous evaluation methods to identify the very best programs and then replicate them (and only them) on a large scale. Critics of social programs often argue that such large-scale replication is impossible, typically predicking this argument on one or both of two premises. One premise is that social experiments are usually done in university settings by extraordinarily skilled people using unusually skilled labor. The other is that a large, centralized federal bureaucracy inevitably screws up anything that it attempts to do.

There is no systematic evidence to suggest that either of these premises is true. Critics often cite the failure of specific federal social programs as evidence, but this argument is not germane. There are no instances in which a specific federal program has been developed by faithfully replicating a small successful program on a large scale. Some might contend that Head Start essentially replicates the Perry Preschool model, but that is simply not the case. No attempt was made in Head Start to copy the Perry Preschool or any other model in a uniform way. Moreover, Head Start’s funding levels are substantially lower than Perry Preschool’s on a per child basis, so it could not offer all of the same services regardless of its design.

The dearth of evidence to support these claims does not mean that they are wrong. There is enough anecdotal evidence of bureaucratic inefficiency to justify skepticism, but not enough to warrant giving up without conducting a reasonable trial on programs that may have the potential to ameliorate serious social problems.

The little evidence that does exist on the question of scale is in fact promising. WIC is a national program that appears to work quite well. And a number of good smaller programs, including Reading One-to-One, Success for All, Reading Recovery, the Child-Parent Centers, and Life Skills Training, have been successfully scaled up from a single site to numerous sites serving thousands of individuals. Clearly there are differences between the scale of thousands and the scale of millions. But successful replications on a medium scale tend to disprove the premise that a hothouse environment is necessary to produce good results.

The pattern of success and failure of replications also bears on the second premise. One factor that tends to be positively correlated with the outcome of a replication is implementation fidelity. When a successful program is duplicated, the closer the copies are to the original, the more successful they tend to be.
This suggests that, contrary to the conventional wisdom, tight centralized control may actually be a necessary condition for success rather than a barrier to it. When control is decentralized, there is a tendency for changes to occur in implementation. People who implement programs like to throw in their pet ideas or adapt the program to the unique circumstances of a particular site. Schorr (1988) even cites such flexibility as an important ingredient of success. Also, funders and legislators may try to expand participation without increasing the budget or try to cut the budget without decreasing participation. Either modification yields a less intense version of the successful model.

Every change increases the probability of failure. The best way to turn a successful small program into a successful large one is to ensure that each replication is as faithful as possible to the original design. Accomplishing this may require the establishment of some centralized authority that oversees the expansion process, acting as the guardian of the model. Ultimately, we will never know whether an effective program can be implemented on a large scale until we identify a genuinely successful program and attempt to scale it up with an absolute commitment to be faithful to the original design.

Because we have not tried it, there is clearly an element of risk in the process. Therefore, any attempt to scale up a successful model should be done in steps, evaluating outcomes at each level. This would enable us to identify and fix any problems that crop up along the way or to scrap the project if it seems destined to failure. There is no formula for the optimal size of each step, but an order of magnitude seems reasonable. A program that has worked with hundreds of participants should be tried with thousands, then tens of thousands, then hundreds of thousands, then millions, and if necessary, then tens of millions.

All evaluations should be done using pure experimental designs. All funding of programs, whether private or public, should be made conditional on this. At first glance that requirement might seem unrealistic. What legislator would pay attention to esoteric issues of social science methodology? Yet this was precisely what was done in the 1988 round of welfare reform legislation, the Family Support Act of 1988. In ordering evaluations of the Job Opportunities and Basic Skills Training Program, Congress mandated that “a demonstration project conducted under this subparagraph shall use experimental and control groups that are composed of a random sample of partici-
pants in the program” (Public Law 100-485 [October 13, 1988], sec. 203, 102 statute 2380).

We have invested at least thirty-five years and billions of dollars in developing good social programs. Most of them have failed, but some of them have worked. If there is a chance that the few successes can be applied on a large scale, it makes no sense to continue developing hundreds of new programs, knowing that the vast majority will fail.

We stand poised at a watershed in the history of social policy. The most important lesson of this volume is that the time has come to focus substantial energy and resources on attempting to turn small successes into large ones. Although there have been many disappointments in the past thirty-five years, we have learned a tremendous amount. If we let this knowledge go to waste, then the money spent on programs over the years will have gone for naught. If, instead, we capitalize on the handful of successes that we have worked so hard to develop, then all the money, the failures, and the disappointments will have been worthwhile after all.

NOTES

1. These programs were originally selected for a conference. The evaluators of eleven programs were invited to present papers at the conference, and those papers indicated that ten of the programs met at least one of the standards described here. The evaluators for each of those ten were invited to contribute chapters to this volume. One group, the evaluators of the Elmira home-visiting program (Olds and others 1988), declined, leaving the nine programs presented here.

2. Even if experimental results are valid for the participants in the study, they may not be applicable to the population at-large.

3. It is possible to enlarge a program at the original site without replicating it, but this is not how programs typically grow.

4. This estimate and those for Reading One-to-One are averages for grades one through three.

5. The average cost for a year of school is assumed to be $6,000, so twelve years of tutoring would increase the overall cost of a child’s primary and secondary education from $72,000 (that is, 12 × $6,000) to $78,000, or 8.3 percent.

6. Many features of the Abecedarian Project were applied in the multisite Infant Health and Development Project. Early results from that program are consistent with those of the Abecedarian Project, but the long-term effects will not be known for several years.

7. This is not to say that WIC is the only federal poverty program that works. There is some evidence that Head Start raises test scores (Cur-
rie and Thomas 1993), and both Aid to Families with Dependent Children and Food Stamps clearly reduce material deprivation. But there is more controversy about the overall cost-effectiveness of those programs, as well as the size and value of the benefits they confer.

8. This was not the result of a design error on the part of the evaluator but rather was necessary because WIC eligibility is legally mandated.

9. If a teacher earns forty thousand dollars a year in salary and benefits, averages five classes of twenty-five students a day for one hundred and eighty days a year, then the personnel cost is $24.89 per student for the fourteen-session program. Training, materials, and administrative costs add a little more.

10. This figure is generated using the same assumptions used for Project ALERT.

11. Because boys were placed in TFC by court order, it would be difficult to create an experimental control group that received no treatment at all.

REFERENCES


