Meta-Analysis and Program Outcome Evaluation

mark w. lipsey

Meta-analysis is a technique for statistically representing and analyzing the findings from a set of empirical research studies. In application to program evaluation research, it provides a means for systematically synthesizing knowledge about the characteristic and outcomes of effective programs. Six lessons learned from meta-analysis of evaluation research illustrate the application and findings of this approach: (1) many social programs are more effective than generally realized; (2) individual evaluations can easily produce erroneous results; (3) the methods used in an evaluation play a large role in the program effects found in the evaluation; (4) program effectiveness is a function of identifiable program characteristics; (5) there is much room for program improvement; and (6) the most credible evidence about program effects comes through integration of multiple evaluation studies.

Introduction

Evaluation provides an assessment of how a particular social program is performing in the context of its mission and the expectations of its stakeholders. However, when designing a new program or reforming an existing one, the responsibility of the evaluation field is to provide evidence about what program approaches have proven most effective in prior evaluation studies. To accomplish this task, evaluators must be able to learn from prior studies what kinds of interventions work for what purposes under what conditions. This, in turn, requires that at least some researchers in the evaluation field systematically gather and integrate the evaluation findings for a

Mark W. Lipsey is Professor of Public Policy at Vanderbilt University and he serves as Director of the Center for Evaluation Research and Methodology at the Vanderbilt Institute for Public Policy Studies.

wide range of programs and program variations. As a fringe benefit, such endeavors also provide opportunity to examine the methods evaluators use and how they relate to the results generated by those methods so that the evaluation field may learn how to improve its methodology.

The central issue raised here is one of generalization how to go from the particulars of individual program evaluations to a broader understanding of the differential effectiveness of different programs for different social problems (Cook, 2000). Valid generalization is the means by which we are able to derive evidence-based principles about what characterizes more and less effective programs. A well developed set of such principles in a given program area is a critical tool for designing, improving, and understanding effective interventions.

Unlike more academic social science fields, where research literature reviews and other forms of knowledge synthesis are commonplace, relatively little attention has been paid to systematic synthesis in the evaluation field. This is primarily due to the nature of evaluation research itself, not because such synthesis is useless. By their nature, evaluations tend to focus on the program under scrutiny and develop in ways that are tailored to the particulars of that program, the concerns of its stakeholders, and the specific purposes of the evaluation. When the findings are analyzed and reported, little or no effort is typically devoted to consideration of the generalizability of the results, how they might apply to other program situations, what has been learned that would be of interest to those who have not yet embarked on a program of that type, and so forth. As applied research, evaluation is organized around application to a specific program context and, correspondingly, evaluators, upon finishing one such project, generally move on to the next without much concern for extracting and reporting the broader lessons of the project for others in the field.

An especially interesting and important area in which the evaluation field would benefit greatly from systematic synthesis of the nature and findings of prior evaluation studies is with regard to outcome evaluation. For most programs, having the intended ameliorative effects on the target problem they address is of paramount political and practical concern. For purposes of program planning and improvement, however, it is of equal importance to know what kinds of programs have meaningful effects on such problems and, among those, which are most effective. More specifically, we might want to know which characteristics of the programs, the target populations, and the evaluation methods are associated with findings of larger and smaller program effects on major outcome variables.

What is Meta-Analysis?

Outcome evaluation is generally conducted using experimental or quasi-experimental research designs with quantitative outcome measures and results that are reported in statistical terms. For research of this type, the technique known as meta-analysis is especially well suited to the task of synthesizing the findings of multiple studies (Cooper, 1998; Cooper

& Hedges, 1994; Lipsey & Wilson, 2001). Meta-analysis revolves around a statistic called an effect size that represents the findings about the program effect on an outcome variable as estimated, for instance, from a comparison between outcomes for a sample of program participants and those for a control sample that does not receive services. The most commonly used effect size statistic for representing the results of intervention research is the standardized mean difference, defined as the difference between the mean value on an outcome variable for the treated group and that for the control group, divided by the pooled standard deviation of the two samples. Division by the standard deviation standardizes the effect size so that, no matter what the original units of the outcome measure, the effect size represents it in standard deviation units. An effect size of .50, for example, indicates that the outcome for the program group on a particular measure was one-half a standard deviation better than that for the control group, irrespective of the measurement scale actually used. Suppose one evaluation study measures depression outcomes on the Beck Depression Inventory and finds that the mean score for the treated group is .40 standard deviations lower (better) than that for the control group. Another study of similar treatment might measure the depression outcome on the Hamilton Depression Scale and find a difference equivalent to .25 standard deviations between the treatment and control group means. We could then compare these, noting that the first study showed a larger effect of treatment on depression. Also, if we wished, we could combine these effect sizes with similarly expressed depression outcomes from many more evaluations of treatment effects into a data set that would allow us to assess the distribution of outcomes, their overall mean, which types of interventions produced the largest effects on depression and which the smallest, and so forth. At this point, we are doing a meta-analysis.

Other types of effect sizes are also used in meta-analysis to represent the outcomes of different studies in a common metric. When the outcome variable is binary, e.g., sick or well, dead or alive, housed or homeless, and so forth, a useful effect size statistic is the odds ratio-- the odds of someone in the program group having the favorable outcome divided by the odds of someone in the control group having that outcome (Haddock, Rindskopf, & Shadish, 1998). Thus an odds ratio of 1.5 means that the odds of a good outcome in the sample receiving service were one and a half times as great as the odds of a good outcome in the control group. Odds ratios are widely used as effect size statistics for representing the outcomes of biomedical interventions and appear frequently in evaluations of medical treatments.

A synthesis of evaluation results using meta-analysis techniques involves computing an effect size for every outcome variable of interest for a collection of evaluations involving the same or similar interventions. These effect sizes are best referred to as the observed effects of the interventions, that is, the effects observed using the measures and methods applied in the evaluation research. Other information

about the nature and circumstances of the intervention, the characteristics of the persons receiving the interventions, the study methods and procedures, and the like are also usually coded for a meta-analysis. All of this information for all the studies included in a meta-analysis is then organized into a database that permits statistical analysis of the distribution of observed effects resulting from those studies.

The typical statistical analysis of a meta-analytic database would first sort the effect sizes according to the type of outcome variables they represent. For example, if the evaluation studies included in the meta-analysis assessed the effects of family therapy on such outcomes as marital satisfaction, quality of communication, and childrens problem behavior, the effect sizes for each of these outcome categories would be analyzed separately. Then, the mean effect size across all the studies would be calculated for each outcome, then the variation of the effect sizes around that mean would be assessed. If the variance of the effect sizes was no larger than expected from the sampling error associated with the samples of persons for whom outcomes were measured in the studies, the mean effect size would provide a good summary of the intervention effect. Because this effect size mean averages over whatever number of studies are included in the meta-analysis, it provides a more representative estimate of the effect of the particular type of intervention on the outcome represented in the effect sizes than estimates derived from any one outcome study.

Frequently, however, the effect sizes from different studies show more variation than likely to result from subject-level sampling error. In that situation, the task of the meta-analysis is to determine if there are systematic relationships between the characteristics of the different studies and the effect sizes they produce. The observed effects of a set of intervention studies can be viewed generally as a function of the efficacy of the treatment, the characteristics of the samples receiving treatment, the methods used to study the effects, and some amount of statistical noise. One useful way of summarizing the information generated by a meta-analysis is to depict the proportion of the variation in the observed effects that is associated with each of these different aspects of the evaluation situation. Further examination can than be made of the specific characteristics of the interventions, treatment recipients, and methods that are most closely associated with larger and smaller observed effects. The results of this process provide the evidence on which we can support useful generalizations about which treatments are most effective on which outcomes for which types of recipients.

Lessons from Meta-Analysis

Meta-analysis has been widely applied to outcome evaluation findings since the pioneering work of Smith and Glass (1977). Though in many ways still not fully developed, it has already generated important lessons about social programs and the methods evaluators use to study them. To illustrate the nature and results

of meta-analysis, and its potential for further enhancing the field of program evaluation, we will describe six lessons we have learned from meta-analysis. The findings that support these insights derive to greater or lesser extent from the work of many meta-analysts. We will not attempt to review the relevant meta-analysis literature here, however. Instead, we will simply summarize what we view as the significant conclusions to be drawn using examples conveniently at hand from our own work over the last decade.

1. Many Social Programs Are More Effective Than We Thought.

One of the troublesome facts of outcome evaluation is that it often finds no significant effects produced by the social programs assessed. It is not unusual for the results of outcome evaluation to be so weak that we cannot be confident the program has meaningful impact. What Rossi and Wright (1984) once called the parade of null results in evaluation can lead to the pessimistic conclusion that nothing works in the world of social programs. The usual basis for such conclusions is a body of outcome evaluations using experimental and quasi-experimental designs that show relatively few statistically significant positive effects on the outcome variables of greatest interest.

One of the distinctive characteristics of meta-analysis is that it focuses on the magnitude of the effects observed in each study, not their statistical significance. Moreover, by combining these magnitude estimates from numerous outcome evaluations, it can reveal the actual distribution of effect sizes that characterize a certain type of intervention. When this is done, it often becomes evident that many of the program effects observed in the original evaluation studies are larger and more consistently positive than they appeared when only those reaching statistical significance were counted. The reason for this, in brief, is that statistical significance is influenced by both the magnitude of an intervention effect and the size of the sample upon which it is measured (Cohen, 1988). Thus effects large enough to be of practical significance may, and in evaluation often do, fall short of statistical significance in an individual evaluation study because the research is conducted with small samples and correspondingly low statistical power.

It is relatively easy to demonstrate the different, and more positive, image of program effects that is revealed by meta-analysis in contrast to the vote-counting approach of assessing the proportion of effects that are statistically significant. Lipsey and Wilson (1993), for instance, assembled all the meta-analyses of the effects of psychological, educational, and behavioral interventions that could be located at the time, more than 300. Many of these were conducted in program areas marked by a history of controversy over whether the interventions produced any positive effects. However, when examined, the distribution of mean effect sizes across this wide range of interventions, and the hundreds of studies and thousands of participants included in the studies meta-analyzed, revealed that the vast majority

of outcome effects were positive and of nontrivial magnitude.

Figure 1.
Distribution of Mean Effect Sizes for 302 Meta-Analyses of the Effects of Psychological, Educational, and Behavioral Interventions

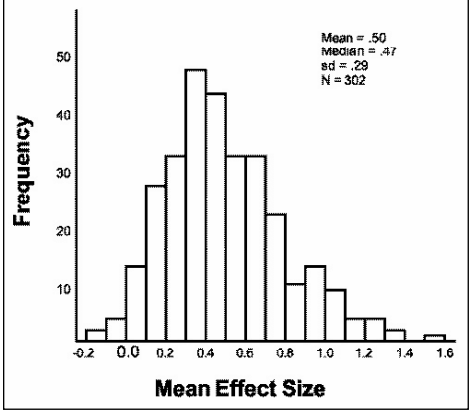


Figure 1 shows the summary distribution of mean effect sizes from all those meta-analyses. The vast majority of the meta-analyses found positive effects on the outcomes of interest (mean effect sizes greater than zero) and the average over these means was about .50. That is, on average across all the interventions represented, the outcomes for the individuals receiving program services were about a half standard deviation better on whatever scale was used for measurement than the outcomes for those in the control conditions who did not receive the program. To put this into perspective, suppose that, on their own, 50% of the individuals in the control group would end up with acceptable outcomes. An effect size of .50 means that, by comparison, nearly 70% of those in the program group would have acceptable outcomes. In many program areas, even smaller effects than this would be of great practical significance.

These meta-analysis results do not mean that all social interventions have positive effects, of course. Nevertheless, they do indicate that to reach any generalization about program effectiveness we should analyze the actual quantitative effect size estimates generated by the available outcome evaluations. The obvious wisdom of this approach, operationalized in meta-analysis techniques, reveals the full range of evaluation findings, and that often proves to represent a wider and more positive set of outcomes than otherwise evident.

2. Individual Outcome Evaluations Can Easily Produce Erroneous Results

The situation described above, in which many outcome evaluations show positive effects, and sometimes relatively large ones, that nonetheless fall short of conventional levels of statistical significance has sobering implications for the design of individual outcome evaluations. By examining the effect sizes over a number of evaluations, and thus in essence combining all their individual study samples, meta-analysis can focus directly on the distribution of observed effect sizes without much consideration of whether each is statistically significant. What we see when we do this, however, is that many of the individual evaluation studies do not show statistically significant effects, even when the meta-analysis reveals that the actual magnitude of the effects for that intervention are

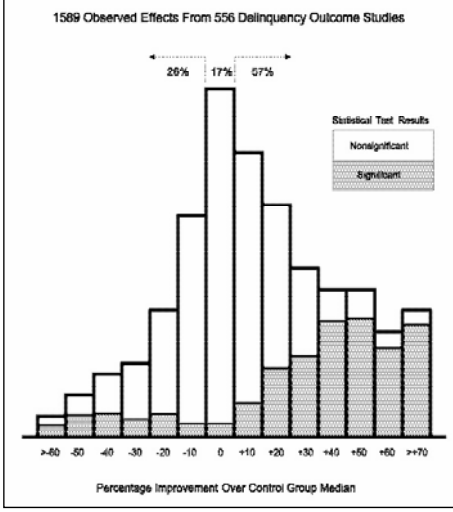
generally positive. In other words, the estimates of the effect sizes for key outcomes in individual studies yield positive values, but fall short of statistical significance and thus cannot be confidently identified as beneficial program impacts within the context of an individual evaluation study.

As noted earlier, this can easily happen when the sample size used in the evaluation design is too small to provide sufficient statistical power for attaining statistical significance even when the effect estimates are of meaningful size. Meta-analysis has revealed that insufficient statistical power is quite common in evaluation research (Lipsey, 2000). An underpowered evaluation design applied to an effective program will usually yield findings that fall short of statistical significance and thus commit what is called Type II error, failing to reject the null hypothesis (of no effect) when, in fact, it is false. From a scientific perspective, effects that fall short of statistical significance in an individual study for whatever reason have little credibility. By definition, they have an unacceptably high likelihood of being spurious, that is, representing statistical error rather than actual intervention effects.

Technically, failure to attain statistical significance in an underpowered outcome evaluation means only that the research has failed to reject the null hypothesis of no effects, not that it has confirmed the absence of effects. However, this is a subtle distinction easily lost on policy makers, program stakeholders, and many researchers. Statistically nonsignificant results are widely interpreted as indications that the program is not effective, with the associated political and practical implications. In this regard, the program is blamed for failing when it is the evaluation research that has failed to use a design with sufficient statistical power to find meaningful effects when they are there to be detected.

The relationship between observed effect sizes, as computed in a meta-analysis, and the statistical significance of those effect sizes found in the individual evaluation studies included in a meta-analysis is illustrated in Figure 2. That figure shows the distribution of effect sizes on all outcome variables reported in over 500 evaluation studies intervention programs for juvenile delinquents. For ease of interpretation, the effect sizes are represented in terms of the percentage improvement shown by the treatment group relative

Figure 2.
Reported Statistical Significance for Different Effect Sizes Observed in Evaluation Studies of Program for Juvenile Delinquents



to the control group median. Thus +30 means that, on whatever outcome variable was measured, the treated juveniles showed a 30% improvement compared to the control group. As can be seen, over half of the observed effect sizes are positive (greater than zero) and many are relatively large (e.g., representing 20% and greater improvement with treatment. Overall, there is little doubt that the interventions evaluated in these studies had positive effects on a majority of the outcomes measured (thought note that 17% of the outcomes were zero and about one-fourth were negative; that is, the control groups did better).

Within each effect size range, Figure 2 shows the proportion of effects found statistically significant in the original evaluation studies. Because the sample sizes in these evaluation studies tended to be modest (a median of about 60 each in the intervention and control groups), they do not have a great deal of statistical power. Figure 2 shows that the majority of the positive effects were not found statistically significant in the individual studies until they were out in the range where treated juveniles were showing improvements of 40% or more compared to those in the control groups. In practical terms, meaningful effects occur below this level, of course. Many programs would be pleased with a 10-20% improvement among the juveniles they served. Moreover, the many positive effects in that range are quite evident in the meta-analysis. But, as can be seen, the individual evaluation studies had a diminishing likelihood of detecting them at a statistically significant level as they got smaller.

It is interesting to note that a similar pattern appeared on the negative end of the continuum. Effects for treated juveniles had to be 50% worse than for control juveniles, or more, before the majority was statistically significant. The decreased proportions of statistically significant results in the negative direction compared with the positive direction that is evident in Figure 2 also represents a problem of small sample sizes. The samples on which negative effects were found tended to be especially small, raising the possibility that, even among those found significant, many may represent no more than sampling error.

The practical limitations imposed on outcome evaluation in field settings is such that it is often quite difficult to enroll samples large enough to ensure a high degree of statistical power. Given the substantial role of statistical noise in such research that has been demonstrated by meta-analysis, outcome evaluation on individual programs can easily fail to attain statistical significance for what are, nonetheless, meaningful program effects. It follows that the results of such evaluation, taken alone, may be misleading. One important contribution meta-analysis can make to this situation is to provide a context of results from other similar interventions within which to interpret the potentially ambiguous findings of an individual outcome evaluation. For example, effect sizes from an outcome evaluation could be compared to the distribution of effects found in a relevant meta-analysis. Their magnitude relative to those found in similar programs could then be assessed as a supplement to assessment by statistical significance testing.

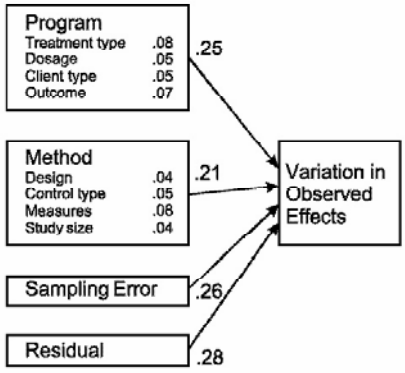
3. Method Matters

Ideally, the experimental and quasi-experimental research designs and procedures typically used for outcome evaluation would generate estimates of actual program effects that were relatively undistorted by the methods themselves. For example, we expect random assignment experiments to produce unbiased estimates of intervention effects, but it is not always possible to use such designs in practical outcome evaluations. It would be comforting to know that a range of more manageable nonrandomized designs would provide results reasonably similar to those from a randomized design. Similarly, when there are various reasonable ways to measure an outcome variable, it would be desirable for them to yield comparable results when applied to the same intervention.

One of the advantages of meta-analysis is that it can investigate the extent to which variation across studies in the methods and procedures of outcome evaluations are related to the effects those studies find. A simple approach is to assess the proportionate variation in observed effects that is associated with the methodological characteristics of the studies in contrast to that associated with such substantive aspects of the programs as the characteristics of the participants, the type of intervention, and the amount of treatment. If most of the effect size variation is associated with differences across studies in program-related characteristics, it is a good indication that the observed outcomes indeed mostly convey information about actual program effects. If, on the other hand, a very large portion of the effect size variation is associated with methodological differences among the studies, it tells us that the outcomes found in those studies may be heavily influenced by the manner in which the program was studied rather than the outcomes it actually produced.

When we have analyzed effect size variation this way with large meta-analytic data sets, we have been dismayed to find that about as much effect size variation is associated with methodological differences among studies as with program characteristics. Summarized over the 300 meta-analyses of psychological, educational, and behavioral interventions we mentioned earlier, for instance, we found the pattern of associations shown in Figure 3 (drawn from Lipsey & Wilson, in press). We have already commented on the large role of sampling error, reflecting typically small

Figure 3.
Sources of Between-Study Effect Size Variance Averaged Over 300 Meta-Analyses



Sources of Between-Study Effect Size Variance Averaged Over 300 Meta-Analyses

sample sizes. Comparing program-related and method-related sources of influence on effect sizes, however, Figure 3 shows that the variation in effect sizes associated with the methods used by the evaluators is larger than that associated with the characteristics of the interventions (21% vs 25%).

When different categories of methodological characteristics are broken out, there are additional surprises. Research design, representing mainly random vs. nonrandom assignment to intervention conditions, and, closely related, the type of control group (e.g., »no treatment« vs. placebo) are influential, as would be expected. There is a large methodological literature on the potential biases associated with design factors (e.g., Shadish, Cook, & Campbell, 2002). Aspects of the outcome measures, however, which have received much less attention in the literature on evaluation methods, also appear to have a substantial influence on the observed effect sizes. The measurement features represented in this category include the way in which the outcome constructs are operationalized (e.g., self-report measures, standardized test, official records) and the timing of measurement (e.g., immediately after intervention or lagged some time later).

Further exploration of evaluation studies with meta-analytic techniques should help determine which methods and procedures yield the most valid results and which create so much distortion that they are not appropriate to use in outcome evaluation. What meta-analysis has already demonstrated is that the neutrality of the typical range of methods for outcome evaluation cannot be taken for granted. What we observe as program effects may reflect as much influence from the methods with which the program was studied as the actual effects the program has on its intended beneficiaries.

4. Program Effectiveness is a Function of Identifiable Program Characteristics

Every social program is, in some regards, unique and the assessment of its impact must be tailored to its particular characteristics and situation. Nonetheless, there are commonalities among programs in a given intervention area that allow for generalizations across programs. It is useful for the evaluator to know what characteristics of an intervention tend to be associated with the most positive outcomes. Such information makes it easier to design an effective evaluation by highlighting the aspects of the program on which the evaluation should focus. In addition, for purposes of program design and improvement, identification of the characteristics of effective programs helps define the »best practices« in a particular intervention area that should be emulated.

Meta-analysis provides a probing way to analyze the characteristics of intervention programs that differentiate those which produce larger outcome effects from those producing smaller ones. Because of the relationship between the methods used in evaluation studies and the observed outcomes described above, however, it can be misleading to simply compare the effect sizes for programs with different characteristics. A potentially clearer picture

is provided by using meta-analysis techniques to statistically control for methodological differences between studies so that the program characteristics most closely associated with larger and smaller effects can be disentangled from methodological artifacts.

With such statistical controls, analyses like those shown in Figure 4 for the studies in our delinquency meta-analysis can be conducted. The details of this analysis are described elsewhere (Lipsey, 1992a,b, 1995), but the results demonstrate that there are consistent relationships between the type of program, how well the program services are delivered and implemented, and the outcomes. Figure 4, for instance, shows that different groups of delinquency intervention programs have quite different mean effects on the juveniles' reoffense recidivism. In particular, the more behaviorally oriented, skill-oriented, and multi-service programs tend to have larger effects.

Figure 4.
Mean Reoffense Recidivism Effect Sizes for Different Groups of Delinquency Intervention Programs with Different Levels of Treatment Implementation

Program Group	Integrity of Treatment Implementation: Low	Medium	High
Group I Skills training Cognitive-Behavioral Multi-service	.29	.45	.62
Group II Counseling Drug abuse Milieu therapy Community residential	.20	.37	.53
Group III Group counseling Challenge programs	.07	.24	.40
Group IV Employment-related Guided group Teaching family home	-.02	.14	.30

The largest effects, however, do not simply follow from using one of the more effective program models. Figure 4 also shows that the integrity of the treatment implementation has at least equal influence on the outcomes. Treatment implementation in this analysis encompasses the amount of treatment provided and the extent of the program efforts to guard against degradation or incomplete coverage in their services. Even programs in the generally most effective group do not have effects in the larger ranges if they are not implemented well. Conversely, programs of a generally less effective type can nonetheless have relatively large effects by implementing their services well. Our analysis has shown many other program characteristics that are also systematically related to their effects, but this example illustrates the general point. Program effectiveness depends upon particular combinations of program features that must be optimally configured to achieve the best outcomes. Moreover, the critical program features are not necessarily unique to any particular program but show general patterns across programs.

Generalizations about the characteristics of the most effective programs, and how they are best combined, cannot be identified in the evaluation of a single program. They are only evident when patterns across programs can be examined. Discovering such relationships, therefore, is a distinctive and important contribution of meta-analysis to the field of evaluation research.

5. There is Much Room for Program Improvement

The outcome evaluation research studies generally available for meta-analysis in any program area typically include a mix of ongoing »real world« programs for which an evaluation has been conducted and various demonstration programs or research-oriented tests of program concepts. One of the useful comparisons that can be made in meta-analysis is to contrast the magnitude of the effects for the best-designed and implemented programs with those of an everyday sort. Demonstration programs designed and implemented by researchers to test state-of-the-art intervention concepts would be expected to produce better outcomes than routine practical programs. Not only do they potentially use more effective intervention approaches, but they also generally have greater control over the consistency of their services and the nature of their clientele.

In this regard, demonstration programs explore the upper limit of program effectiveness attainable with available intervention techniques and thus show what practical programs might aspire to under optimal circumstances. A large gap between the effects of practical programs and those of demonstration programs in an intervention area suggests that the practical programs may be able to improve their effectiveness by modeling key features of the demonstration programs. Unfortunately, meta-analytic investigation of the effectiveness of demonstration programs in contrast to everyday practical programs has, to date, only been undertaken in a limited way. The early indications, however, show rather sizeable gaps in favor of the demonstration programs (e.g., Weisz, Weiss, & Donenberg, 1992, on childrens mental health programs).

The nature of the situation can be illustrated with data from the meta-analysis of programs for juvenile delinquents to which we have already made reference several times. We divided the programs into real world practical programs evaluated by a researcher who was not involved in designing the program or delivering the service and contrasted their outcomes with demonstration programs designed and implemented by the researcher. Simply comparing the overall effect sizes for reoffense recidivism outcomes revealed that the mean for the practical programs (.07) was only about half that for the demonstration programs (.13), though both were modest (but with much variation around them).

When the characteristics of the practical and demonstration programs were compared, a number of specific differences emerged. Among the most important and interesting were the following.

- Type of program: less likely to be one of the more effective types (skill-building, behavioral, multi-service) for practical than demonstration programs.
- Administered by juvenile justice personnel: more likely for practical than demonstration programs.
- Monitoring of the integrity of the service implementation: less likely for practical than demonstration programs.
- Difficulties in treatment implementa-

tion reported: more likely for practical than demonstration programs.

- Program duration: about 25 weeks for practical programs; about 38 weeks for demonstration programs.
- Intensity of treatment: rated lower for practical programs than for demonstration programs.

Although some of the advantageous characteristics of the demonstration programs may be difficult for practical programs to emulate (e.g., program types that require highly trained personnel), others are clearly feasible. The results of comparisons such as this, therefore, can be used to guide the improvement of practical programs in ways that should enhance the magnitude of their outcome effects. The validity of this perspective is supported by analysis of the considerable variation within the domain of practical programs themselves. Not surprisingly, practical programs have many of the favorable program features identified above while others have less favorable configurations. If we examine the mean outcome effects for the practical programs that are more favorable configured in these terms, we find that they are indeed more effective.

Figure 5 shows one such comparison for the juvenile delinquency programs that focuses on reoffense recidivism outcomes. The practical programs are categorized according to how many characteristics they have from the set found in the meta-analysis to be related to effect sizes. There is a clear trend for those with a greater number of favorable characteristics to produce greater mean reductions in recidivism among their juvenile clients relative to control cases. Indeed, those with none of the favorable characteristics actually show an increase in recidivism among the juveniles they treat.

Figure 5.
Improvement in Recidivism Rates Relative to the Control for 196 »Real World« Delinquency Programs with Different Numbers of Favorable Program Characteristics

Number of Favorable Characteristics*	Distribution of Programs	Percentage Reduction in Recidivism
0	7%	+12
1	50%	-2
2	27%	-10
3	15%	-20
4	2%	-24

*Favorable Program Characteristics:
Uses one of the more effective types of service, e.g., skill-oriented, multimodal
Juvenile justice administered program conducted in non-JJ facility
Good program implementation with relatively high amount of service
Works with juveniles with mean age >15 years or with mixed prior offenses.

Perhaps equally interesting is the distribution of the programs represented in the meta-analysis across the various categories shown in Figure 5. More than half of the programs evaluated had zero or one favorable characteristic and, correspondingly, minimal or counterproductive effects. On the other hand, only 2% of the practical programs had the full complement of favorable characteristics and achieved the higher levels of recidivism impact. Possibly the most favorably configured programs are not evaluated, or their evaluations not reported, so that they would be underrepresented in the research available for meta-analysis. It seems more likely, however, that most practical programs, in fact, are not configured for optimal impact and have considerable room for improvement.

6. There is Safety in Numbers

Perhaps the most significant lesson from meta-analysis is the one that encompasses all the others: Many factors influence the findings of an outcome evaluation and, even under the best of circumstances, the validity of those findings is uncertain. While there is, and will continue to be, an important role for outcome evaluation of individual programs, we must be very cautious in interpreting a single set of results, even from a well-designed evaluation study. Ultimately, the most credible evidence about effective programs will come through careful integration of evaluation results from many studies and programs. Correspondingly, one of the greatest challenges facing the evaluation profession is how to ensure that high quality, useful synthesis of evaluation studies are carried out and the results disseminated to relevant evaluators, practitioners, and policymakers.

An important recent initiative offers great promise for meeting this challenge. In 1999 an international group of evaluators, policymakers, and researchers met at University College in London and agreed to launch the Campbell Collaboration for developing and disseminating systematic synthesis of outcome evaluation findings for social programs. This endeavor is modeled on the Cochrane Collaboration, which organizes syntheses of medical health-related research, and was named in honor of the American psychologist and methodologist, Donald Campbell, a renowned advocate for rigorous program evaluation. The Campbell Collaboration (C2) has grown rapidly and currently has a membership drawn from 15 countries and coordinating groups in the areas of crime and justice, education, social welfare, synthesis methods, and dissemination. C2 aspires to sponsor and facilitate high-quality synthesis of outcome evaluations for social programs and make them readily available on the world wide web to interested parties (http://www.campbellcollaboration.org). Though still in its infancy, the Campbell Collaboration has numerous syntheses underway and holds great promise as a way to extract and share the lessons that can be learned from the thousands of studies conducted in the vigorous field of program evaluation.

References

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cook, T. D. (2000) Toward a practical theory of external validity. In L. Bickman (ed.), Validity & social experimentation: Donald Campbell's Legacy (vol. 1, pp. 3-43). Thousand Oaks, CA: Sage.

Cooper, H. M. (1998). Synthesizing research: A guide for literature reviews (3d ed.). Thousand Oaks, CA: Sage.

Cooper, H. M., & Hedges, L. V. (Eds.). (1994). The handbook of research synthesis. New York: Russell Sage.

Haddock, C. K., Rindskopf, D, & Shadish, W. R. (1998). Using odds ratios as effect sizes for

meta-analysis of dichotomous data: A primer on methods and issues. Psychological Methods, 3, 339-353.

Lipsey, M. W. (1995). What do we learn from 400 research studies on the effectiveness of treatment with juvenile delinquents? In J. McGuire (ed.). What works? Reducing reoffending (pp. 63-78). NY: John Wiley.

Lipsey, M. W. (1992a). The effect of treatment on juvenile delinquents: Results from meta-analysis. In F. Loesel, D. Bender, & T. Bliesener (eds.). Psychology and law: International perspectives (pp. 131-143). Berlin; NY: Walter de Gruyter.

Lipsey, M. W. (1992b). Juvenile delinquency treatment: A meta-analytic inquiry into the variability of effects. In T.D. Cook, H. Cooper, D.S. Cordray, H. Hartmann, L.V. Hedges, R.J. Light, T.A. Louis, & F. Mosteller (eds.). Meta-analysis for explanation: A casebook. NY: Russell Sage Foundation.

Lipsey, M. W. (2000). Statistical conclusion validity for intervention research: A ($p<.05$) problem. In L. Bickman (ed.), Validity and social experimentation: Donald Campbell's legacy (vol. I). Sage.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. American Psychologist, 48, 1181-1209.

Lipsey, M. W., & Wilson, D. B. (2001). Practical meta-analysis. Applied Social Research Methods Series, vol. 49. Thousand Oaks, CA: Sage.

Rossi, P. H., & Wright, J. D. (1984). Evaluation research: An assessment. Annual Review of Sociology, 10, 331-352.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. American Psychologist, 32, 752-760.

Weisz, J. R., Weiss, B. D., and Donenberg, G. R. (1992). The lab versus the clinic: Effects of child and adolescent psychotherapy. American Psychologist, 47, 1578-1585.

Wilson, D. B., & Lipsey, M. W. (in press). The role of method in treatment effectiveness research: Evidence from meta-analysis. Psychological Methods.