J. ENVIRONMENTAL SYSTEMS, Vol. 13(1), 1983-84

PROCEDURES FOR AGGREGATING THE RESULTS OF ENERGY CONSERVATION RESEARCH

BRIAN STECHER
Educational Testing Service

ABSTRACT

Various procedures for aggregating the results of research on energy consumption are analyzed and some important criticisms are addressed. The methods include the "voting" method, combined significance tests, measures of overall effect size, and cluster analysis. Methods for testing the influence of relevant mediating variables are also described. Effect size measures are recommended, but the energy evaluator is cautioned to select the procedure that is appropriate to the body of research under investigation.

The number of evaluation studies in the energy field has increased rapidly in the past decade, and there is a burgeoning literature in this area [1-3]. Even by conservative estimates, the amount of research and evaluation of energy conservation efforts is substantial. The purpose of this paper is to describe a set of procedures called, variously, "meta-evaluation," "meta-analysis," "outcome synthesis," or "research integration" that can be used to combine the results of multiple energy studies.

Meta-analysis (the term we will use most frequently—but not exclusively—to refer to those procedures) has evolved over the last two decades in response to the growing number of research and evaluation studies that address particular topics. While each study contributes some bit of information to an overall understanding of the phenomena under investigation, the accurate appraisal of the current level of understanding cannot be achieved without integrating the findings across all of the studies. This is the purpose for which meta-analysis procedures were developed. Even a cursory reading of the literature in the area of energy planning and conservation suggests a number of topics that have received substantial research attention and would be appropriate subjects for

doi: 10.2190/WPMG-5YNV-EPAC-JPMM
http://baywood.com

meta-analysis. These include: the effect of advertising on conservation attitudes, the impact of appliance efficiency rating information on customer purchasing behaviors, the relationship between peak hour pricing patterns and energy consumption, etc. Consequently, there should be little argument about the value of procedures for systematic aggregation of information in the energy field.

Meta-analysis can be valuable both as a guide to program planning and as a policy tool. An accurate synthesis of past evaluations can enhance new program development.

> To concentrate our energies on more primary research and evaluation studies without systematic integration of previous studies is scientifically and educationally wasteful. Systematic integration . . . permits us to assign degrees of confidence to conclusions, to estimate their scope of applicability, and to point to the most fruitful direction for subsequent empirical study [4].

In fact, meta-analysis can even be a useful tool for improving future evaluations of conservation programs [5].

Similarly, a concise and accurate summary of research and evaluation studies can provide useful information for policy formation.

> If research findings are to inform policy, they must be put into understandable form and provide some answers. Occasionally these are clear-cut, but often they are likely to be more complex, reflecting the real world relationships between policy variables and outcomes . . . Even when a policy question is complex, however, there is a strong need for summary information. Again, a narrative description of a hundred studies is frequently not enough [6].

Though much of the work on synthesizing outcomes has been done by educational psychologists, it seems apparent that meta-analysis will be used more frequently in the energy field. Those involved in energy-related program development and policy formation should become familiar with the strengths and weaknesses of these techniques.

STRATEGIES FOR META-EVALUATION

There are two general approaches to the task of summarizing the results of a collection of related evaluations. One approach is "to read through the various findings and reach a series of impressionistic conclusions. A second approach is to apply precise analytic procedures to the collection of studies." [6] The thrust of this article is to argue that more "precise, analytic procedures" are appropriate. Yet, it is important to understand the elements involved in a subjective synthesis in order to appreciate the advantages that can accrue from more systematic procedures.

Narrative/Impressionistic Approaches

The dominant style of research synthesis during the past two or three decades was impressionistic, and this was directly related to the amount of research and evaluation literature being examined [7]. When there were only ten or fifteen studies on a given topic, a knowledgeable reviewer was usually able to summarize the conclusions in a qualitative manner without relying on specific numerical comparisons or quantitative aggregation techniques. A careful perusal of the reports formed the basis for a personal judgement of their aggregated impact.

For example, a cursory glance at the evaluation studies reviewed by the Energy Conservation Group reveals about a dozen reports on the use of advertising as a conservation tool [8]. Suppose, for the moment, that there are twelve evaluation reports that examine the impact of advertising on attitudes toward conservation. It would be useful for utility companies and state regulatory agencies to understand the relationship between advertising and conservation behavior as thoroughly as possible. Each of the studies may shed some light on this topic, and it would be shortsighted to plan new programs or implement new regulations without considering the results of these investigations.

As a result, an independent researcher, who is knowledgeable in the area of energy conservation, might be commissioned to summarize the findings of these twelve studies. The simplest procedure that might be followed would be to read the studies and gradually accumulate impressions about the important variables that have been investigated. A competent researcher should be able to combine results and formulate an impression of the dominant themes of the various evaluations. This requires an ability to balance contradictory findings, to assess the strength of various conclusions, to synthesize and distinguish differences in approach, procedures and outcome measures, and to arrive at a cumulative judgement of the aggregated impact of these evaluations.

Although this might be a difficult task, it is not impossible. If the weight of the evidence is clearly in favor of one conclusion, it may be easy to state this fact and to discuss related questions that have not been directly addressed. However, if the findings of the various evaluations are contradictory, the task becomes more difficult. Yet, a clever researcher may still be able to develop a qualitative impression of the impact of advertising on conservation attitudes and present a general summary.

This brief example stressed the positive elements of impressionistic aggregation, but the weaknesses should be obvious. First, a narrative synthesis is subjective; two individuals working independently might not arrive at the same conclusions. In fact, as the number of studies grows and the amount of information that must be compared increases, the subjective aspects of the procedure take on ever increasing proportions.

> Faced with tens or even hundreds of studies on a single topic, a reviewer unarmed with systematic procedures is forced to utilize subjective criteria for deciding how to synthesize. He may choose several favorite studies, relatively well done from a classical experimental design standpoint. Or he may favor studies carried out by investigators he respects. In either case, his impressionistic conclusions will often differ from those of the next well-intentioned reviewer [6].

A second drawback of informal integration procedures is that they can yield an inaccurate assessment of the significance and size of the overall effect that is present. Without systematic, quantitative procedures for combining tests of significance and measures of impact, a reviewer is forced to aggregate these measures intuitively. Very few people have a refined mathematical intuition that would serve them well in such circumstances. In fact, it has been found that the estimated magnitude of the effect being reviewed increases when statistical procedures were used in contrast with impressionistic ones [9]. Cooper and Rosenthal concluded that, "traditional reviewers either neglect probabilities or combine them intuitively in an overly conservative fashion."

As a result of these and other problems, most researchers now agree that narrative, impressionistic approaches lack sufficient rigor and reliability for large scale meta-evaluation. In fact, these are precisely the concerns that led to the development of statistical techniques for research aggregation, which will be discussed in the next section.

Nevertheless, one should realize that impressionistic techniques are still appropriate under certain circumstances. One conclusion to be drawn at the end of this paper is that different meta-analysis procedures are appropriate under different circumstances, and it is the task of the evaluator to select the proper strategy depending upon the availabile data, the subject matter and the goals of the meta-analysis. This is true for narrative, rhetorical meta-analysis as well. In fact, impressionistic techniques may be the best way to summarize qualitative studies. Only a narrative review can maintain the contextual richness captured in good qualitative, naturalistic evaluation. Even here, of course, the number of studies that can be considered at any one time in an impressionistic meta-evaluation remains severely limited.

Quantitative/Statistical Approaches

The alternative to impressionistic aggregation is to apply the same kinds of quantitative statistical techniques to a collection of evaluation findings that would be applied to the data in each of the evaluation studies.

> The accumulative findings of dozens or even hundreds of studies should be regarded as complex data points, no more comprehensible without the full use of statistical analysis than hundreds of data points in a single study could be so casually understood (sic) [7].

Just as there are a number of statistics that can be used to describe a collection of data, there are a number of statistics that have been proposed for aggregating outcomes across studies. These will be described in subsequent sections. However, there is a preliminary question that must be addressed before any of these techniques can be applied. That is the determination of which studies to include in the analysis. This simply stated question is actually quite complex, and it has been the subject of heated debate. Thus, it deserves specific attention.

WHICH STUDIES SHOULD BE INCLUDED?

The most thoroughly debated issue in the field of meta-analysis has been the question of selection. Which studies are properly included in a quantitative aggregation of research findings? The debate has focused on three points:

1. Should inclusion in the analysis be a function of methodological rigor?
2. Is the analysis biased because unpublished, unreported results are excluded?
3. Does an aggregated measure derived from programs that differ in many ways have any meaning at all?

Each of these points deserves further examination.

Methodological Quality

There is strong disagreement about the role of evaluation quality in a meta-analysis [11-15]. At one extreme it is argued that a poorly conducted study provides no information whatsoever and should be excluded from any research synthesis. The argument is summarized neatly by the computing axiom "garbage in–garbage out." The alternative is first to assess the methodological rigor of the collection of evaluations that are being addressed. Those that could not meet some appropriate standard of research quality would be discarded. As Mansfield and Busse argue, "Poorly designed studies are likely to yield spurious findings precisely because these studies are poorly designed." [11]

Others prefer much more lenient standards, accepting any study with sufficient information to calculate an effect size [15]. (The notion of effect size will be discussed below.) They argue that a less-than-perfect study (and how often does one encounter a *perfect* study) may yield valid findings. More importantly, the relationship between evaluation quality and findings is an empirical question that can be subject to post hoc analysis. If studies lacking "good design" yield similar conclusions to those with superior methodology, then they should be included in the calculation of overall impact. If not, then this relationship deserves further investigation.

At this time there is no best resolution to the question of methodological quality. One set of minimum guidelines seems reasonable.

> It seems clear that all studies included in the synthesis should adhere to certain basic standards for research reporting. These include providing summary statistics, performing statistical analyses correctly, and adequately describing outcome measures [6].

Beyond that, procedures suggested by Glass and Smith seem most reasonable [15]. In fact, their approach generalizes beyond questions of research quality to examinations of other relevant study characteristics. Any of "the features of the research problem and setting which might mediate results must be measured or otherwise expressed in quantitative terms" and the relationship between these characteristics and the measures of effect should be examined [7].

Unreported Studies

Another criticism of meta-analysis focuses on the "file drawer" problem. Some have argued that there is a significant number of unavailable studies, languishing in file drawers, which might change the overall conclusion of the meta-evaluation if they could be included. In fact, the argument continues, it is reasonable to assume that those unpublished dissertations and unreported studies had null conclusions. Studies which are published are the ones in which significant results were obtained, while equally valid studies with no significant differences are more likely to be unpublished and unavailable. This is true for evaluation as well as research.

> The evaluations available for review may constitute a biased sample. This is especially likely in evaluation research, where many final reports are never published and are only available from agency files to which access is not always easy [16].

However, statistical examinations of the "file drawer" problem found it to be less of a concern than had been anticipated. Rosenthal found that the number of "missing reports" with null hypotheses which would have to be added to a research synthesis to counteract otherwise significant findings would be enormous [17]. He concluded "when the number of studies available grows large or the main (overall effect) . . . grows large, the file drawer hypothesis as a plausible rival hypothesis can be safely ruled out." However, this may not be the final word. Smith, who looked at measures of effect size, rather than simply tests of significance, found that "the average experimental effect from studies published in journals was larger than the corresponding effect estimated from theses and dissertations." [18] Thus, while Rosenthal's examination suggests that the failure to include "file drawer" studies is unlikely to affect the overall significance of a meta-evaluation (if a large number of studies have been included), Smith's analysis suggests that it may affect the estimate of the effect size.

Yet, this should not be cause for alarm. The presence of publication bias can be examined as an empirical question. By comparing the published studies and

unpublished theses that are available, it is possible to estimate the size of this effect and include this information in the meta-analysis report. This discussion points up a useful guideline for meta-analysis: Include as diverse a sample of research as possible—one which represents all of the characteristics which might potentially affect the overall results.

Program Differences

Another concern for those who would conduct meta-analysis is the problem of "comparing apples and oranges." Simply stated, this criticism says that it is meaningless to derive a combined effect measure from studies of programs that are not the same.

In the conservation context, the argument might be that a program in which salesmen tell customers about the efficiency ratings of various appliances is not the same as one in which these ratings are posted inconspicuously on the appliances and are never specifically mentioned by the salespersons. Since these are very different treatments, the argument continues, it is meaningless to combine them in a single analysis. Only when two studies are essentially the same—the same dependent measures, the same treatments—does it make sense to aggregate the results.

Those who believe in meta-analysis counter this argument by noting that no two studies are *ever* completely the same. If they were, the results would, of necessity, be identical, and there would be no need for synthesis. The goal of a research synthesis is to combine evidence from investigations that are *similar*, not identical. The "topic" that is addressed is a general issue of concern that encompasses innumerable, small, real-world variations. The bounds for the meta-evaluation are set when the topic area is defined, and this can be as broad or as narrow as one chooses. The appropriateness of an individual evaluation depends upon this definition.

In fact, not only is the "topic area" an undefined term, but the definition of what constitutes a "study" and what constitutes a "finding" are equally vague.

> The basic units on which a meta-analysis is carried out are essentially undefined terms. One must trust a relatively widely shared understanding of the words *study* and *findings* [17];

A similar statement could be made about the topic of meta-evaluation. The problem of combining apples and oranges is resolved by noting that they are both pieces of fruit.

Though there has been a great deal of debate about the question of what should be included in a meta-analysis, we are comfortable with the resolution that has been proposed above. It may not be the final answer, but it contains a practical working basis for anyone who would attempt to do a meta-analysis in the area of energy conservation or related research.

META-EVALUATION PROCEDURES

We now turn our attention to the question of how to combine research results and derive an overall aggregated assessment of the findings in a particular area. Four procedures have been proposed, varying in complexity and sensitivity. In the following sections we will describe each of the procedures and then address the question of how to supplement these analyses to test for significant interactions. Finally, we will discuss how to choose which procedure is appropriate in a given situation.

We will use the following hypothetical data to illustrate each of the aggregation procedures. Assume that there are six evaluations of the effect of electricity-related conservation education on home electricity use. Although the evaluations differ in a number of ways, we will suppose they are similar in certain respects. The dependent variable involved the presence or absence of a conservation program directed towards electricity consumption. Another similar element in all evaluations was the presence of a control group whose electric consumption was compared with the treatment group to assess the effect of the conservation program.[1] Finally, all evaluations reported basic univariate statistics on electricity usage for the two groups. These data are summarized in Table 1.

Beyond this, the studies differed in a number of ways, including sample size, methodological quality, and the specific form of the education program. In addition, the results were reported in different forms: One study tested the difference between treatment and control groups using a t-test, another reported an F-statistic, while a third reported consumption levels but did not conduct a significance test on the observed differences.

Methods Based on Significance Tests

There are two meta-analysis procedures that involve tests of statistical significance. One, called the voting method, counts the significant differences found in individual studies. The other involves calculation of a combined significance test for the whole collection of evaluations.

The *voting method* is one of the simplest procedures for combining the results of a number of evaluations [19]. One simply tallies the number of evaluations in which there was a significant difference in favor of the treatment group, the number in which there was a significant difference in favor of the control group and the number in which no significant differences were found. Whichever category has the largest number of "votes" is deemed to represent the combined impact of the collection of studies that was examined.

[1] In order to use most meta-analysis techniques, the studies being analyzed all must report either comparisons between treatment and control groups or correlations among similar variables. That is, there must be some basic measure of the direction of the effect or its magnitude.

Table 1
Electric Usage in Six Hypothetical Conservation Programs

Name	*N*	*Treatment Mean* ($\bar{X}_T$)	*Treatment S.D.* (S_T)	*Control Mean* ($\bar{X}_C$)	*Control S.D.* (S_C)
Study A	50	25	5	27	5
Study B	450	24	2	20	3
Study C	53	19	4	22	5
Study D	1,275	27	2	33	2
Study E	120	23	3	17	2
Study F	25	31	3	35	4

In our hypothetical example not all evaluations reported significance tests, but from the information given it was possible to compute this for all six evaluations. (Any evaluations for which such tests cannot be computed or estimated must be excluded from the meta-evaluation.) The voting method ignores the size of the differences and uses only the direction of the effect. Thus, in Table 2, a plus signifies a significant finding in favor of the experimental group, a minus signifies a significant difference in favor of the control group and a blank means that there was no significant difference. According to the voting method, the weight of evidence from these six evaluations is that conservation education has no discernible impact on the home use of electricity (three to two to one).

While this approach can be praised for its simplicity, its weaknesses are apparent. First, it ignores sample size. In this particular case, it might be appropriate to give more credence to the results of Study D, which involved

Table 2
Significance Tests in Six Hypothetical Conservation Evaluations

Name	$\bar{X}_T$	$\bar{X}_C$	*Significant Difference?*
Study A	25	27	None
Study B	24	20	–
Study C	19	22	None
Study D	27	33	+
Study E	23	17	–
Study F	31	35	None

more participants than all the other evaluations combined. Study D concluded that conservation programs were effective—a result at odds with the overall conclusion drawn by the voting method. In response to this concern, some have suggested that a weighting scheme based on sample size would be a more appropriate way to "tally" the votes. However, there are complex statistical issues involved in determining what the appropriate weighting should be and there is no consensus at the present time about how such a procedure should be done.

An alternative procedure has been suggested which incorporates sample size into the calculations. This involves integrating significance tests across studies into a single *combined significance test* of the difference between experimental and control groups. The major advantage of such procedures is that they increase the power of the overall test, by increasing the sample size. Moreover, the calculations are not difficult, requiring only that each study report the sample size and the value of the significance test, expressed either as a probability, a t-statistic or an F-statistic.

The potential power of this procedure can be seen if we restrict our attention to studies A, C, and F. Because of the small sample size, each of these tests of difference was not significant, though all three studies favored the treatment group. By combining information from all three studies and tripling the sample size, a significant difference might well be found.

The simple example illustrates one of the strengths of combined significance tests. However, there are numerous disadvantages. First, the procedure assumes that the individual studies are independent. This may be an untenable assumption in real life. Second, as the number of studies increases and the sample size grows, the likelihood of finding statistically significant differences increases. Though such small differences may be statistically significant, they may have little, if any, *practical* value.

This points up a common drawback of combined significance tests and the voting method. Significance tests, themselves, may not provide the most useful information about measurable differences. Notice that the differences that were observed in the six studies in Table 1 were not of the same magnitude. As Glass pointed out, "tallies of statistical significance or insignificance tell little about the strength or importance of a relationship." [7] It would seem appropriate to try to incorporate some measure of the magnitude of the observed differences when aggregating the findings. Referring to Table 1, the positive differences that were observed in Study D far outweigh the negative and null differences detected in the other studies. Yet, this fact is entirely ignored when using methods based on significance tests.

Average Effect Size

The meta-analysis procedure which is most widely used at the current time is based upon the computation of an average effect size across studies. The effect

size of a study is defined as the difference between the experimental and control group means expressed in terms of the control group standard deviation. The formula for effect size is:

$$d = \frac{\bar{x}_t - \bar{x}_c}{s_c}$$

This difference is represented pictorially in Figure 1. The figure shows the distribution of treatment group means and control group means across all evaluations. The average effect is represented as the difference between the average of the treatment means and the average of the control means.

For example, in a classic meta-analysis, Smith and Glass examined a collection of 375 studies of the effects of psychotherapy and counseling [13]. They found that, on the average, therapy had a .69 standard deviation positive effect over control. In Table 3, the effect sizes have been calculated for each of the six hypothetical conservation evaluations. Averaging yields an overall measure of the effect of conservation education on consumption of electric power.

While computing average effect size enjoys the greatest current popularity among meta-analysis techniques, it too has drawbacks. As with the voting technique, no attention is paid directly to sample size. (Though this can be examined in post hoc analyses.) In addition, calculation of actual effect size

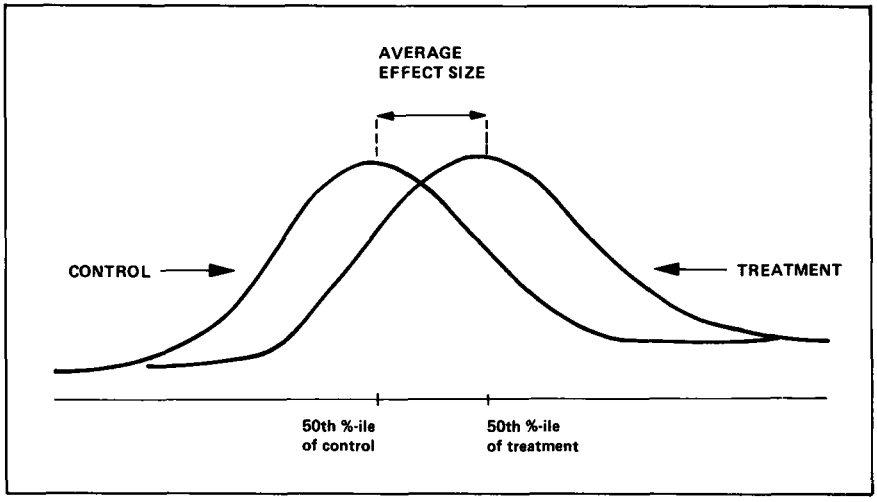

Figure 1. Average effect of treatment across studies. (Adapted from Glass [10]).

Table 3
Effect Sizes for Hypothetical Conservation Programs

Name	$\bar{X}_T$	$\bar{S}_T$	$\bar{X}_C$	$\bar{S}_C$	*d*
Study A	25	5	27	5	- .4
Study B	24	2	20	3	1.3
Study C	19	4	22	5	- .6
Study D	27	2	33	2	-3
Study E	23	3	17	2	3
Study F	31	3	35	4	-1

Note: A negative effect means that treatment group has a lower level of consumption, and, therefore, the conservation education program was successful.

can be quite difficult. Many evaluations do not report the necessary figures for computing effect size. Though a number of transformations can be made (if certain reasonable assumptions are met) to derive measures of effect size from other commonly reported statistics, not all computational questions have been resolved [20]. For example, while there are good arguments for using the control group standard deviation as the denominator in the measure of effect size, difficulties arise when two treatment conditions are compared. Similarly, problems in estimating effect size are exacerbated when the variances within groups are heterogeneous.

Many of these issues are addressed in more recent work on meta-evaluation. For example, the statistical properties of Δ, the probability distribution of the effect size estimator, have been determined [21]. This allows an overall test of the significance of the average effect size derived in a meta-evaluation. Hedges also found that the sample statistic, d, is not an unbiased estimator of the actual effect size in the population and derived a correction factor that can be used to provide an unbiased estimate of the population parameter [22].

Effect size has been the most widely used quantitative technique for research aggregation [23]. It is also the procedure that has been subject to the most careful scrutiny. While there remain unsolved problems in estimating effect sizes in certain situations, and the procedure has certain general weaknesses, it is the most sensitive of the synthesis techniques discussed so far. It appears to be the approach that best combines sensitivity and practicality.

Cluster Approach

Perhaps the most rigorous approach that has been suggested for meta-analysis is called cluster analysis [19]. The procedure is impractical, however, because it requires access to the raw data from which the evaluations were made. It is

called the cluster approach because it involves an examination of individual clusters of data prior to combining them in a broader pool for analyses. A series of analytic steps are taken to determine if it is appropriate to combine treatment group data from one study with treatment group data from another.

For example, the cluster approach suggests that it might be inappropriate to combine the information on conservation education from evaluations D and E because subjects evidenced such dramatically different response patterns. Further analysis should be done to determine whether there were substantial differences in the treatment conditions before it would be appropriate to combine these data in a single assessment.

Though impractical for most real situations, the cluster approach does emphasize the importance of examining potentially relevant sources of variation. It is a highly rigorous approach, which, though praised in theory, has seldom been used in practice.

EXAMINING INTERACTIONS

No description of meta-analysis procedures would be complete without discussing how to ascertain the effects of relevant program and study characteristics on meta-analysis results. Glass, McGaw and Smith argued that the presence of interactions between study characteristics and meta-analysis findings is an empirical question which can best be addressed through subsequent analyses [20]. This prescription is widely accepted, and in this section such a posteriori procedures will be described in greater detail.

As has been mentioned frequently in preceding discussions, it is not enough merely to report the average effect size or the results of a vote analysis. The competent researcher must undertake a thorough analysis of other relevant variables that might affect the results of the meta-analysis. Significant characteristics that might influence the final results—demographics characteristics, variations in treatment, etc.—should be coded and subjected to post-hoc analyses.

For example, if there is concern about the possible differences between "high quality" evaluations and less rigorous ones, all of the studies in the analysis can be classified according to this criterion and an empirical determination of effect size differences can be made. Moreover, it behooves the meta-evaluator to examine specifically any characteristic that might logically mitigate the observed relationships. These include subject demographic characteristics, variations in treatment and variations in outcome measures.

This necessitates a large enough sample of evaluation studies so that any characteristic of interest will be represented frequently. Consequently, meta-evaluators should obtain as many relevant studies as possible. No one has yet defined what a sufficient number is, but most meta-analyses that are conducted in the area of educational psychology include fifty or more individual studies.

Depending on the diversity of the findings and the complexity of the issue, the number of evaluations that will be required to arrive at a consistent conclusion will differ. Unfortunately there is no magic number. Interaction analyses can use any number of different analytic techniques. For example, regression has been used to test for the effects of potential mediating variables [13]. It is also possible to examine contingency tables and cross tabulations. When average effect size procedures are used, correlation coefficients can be computed to test for the presence of important interaction effects.

WHICH TECHNIQUE SHOULD BE USED?

In the preceding section, four different approaches to meta-analysis were described. The obvious question is, which approach is best? Pillemer and Light argue that "no procedure is always best, but rather the different kinds of questions require different ways of combining outcomes." [6] It is important to understand that the meta-evaluator has a choice and that no single method is appropriate under all circumstances.

However, while there are circumstances under which each of the techniques has an advantage over the others, the bulk of the current literature on meta-analysis focuses on effect size. Glass, McGaw and Smith [20], who have authored the only comprehensive textbook on meta-analysis to date, argued strongly in favor of using effect sizes whenever such measures can be calculated (and similar mathematical combination of correlation coefficients when results are reported in that form). Our personal preference for effect size measures, conditioned upon other potentially mediating characteristics, should be obvious from the previous discussion. The voting procedure may be appropriate when the evaluations under consideration do not report enough data to calculate effect sizes. The combined significance test is less attractive. While it may be more sensitive than the voting procedure, conditions that allow calculation of a combined significance test may also present enough information to estimate effect size, and this would be preferable.

The critical factor in the choice of method must be the reviewer's substantive understanding of the issue being evaluated. None of these techniques will substitute for a basic knowledge of the subject matter. The choice of technique must be based upon the nature of the question being addressed, the purpose of the review, the number and type of evaluations that are available, and the kind of information that is reported in the studies [4]. The meta-evaluator selects the best technique, or develops a new one, after carefully reflecting on all of these considerations.

SUMMARY

Interest in the aggregation of evaluation findings is not new, nor are efforts to conduct such syntheses. What is new, however, is the emergence of systematic, quantitative procedures for producing meta-analyses. As a result, there is more

active interest in meta-analysis, and its potential impact on evaluation and policy has increased.

This paper examined the use of meta-analysis to summarize the findings of multiple energy-related evaluation studies. As conservation efforts grow and the amount of related research and evaluation increases, the potential benefits of meta-evaluation will increase. Such research synthesis will help advance our understanding of the factors that affect conservation and the strategies that might be employed to increase the efficiency of energy use. This kind of information will be relevant both to utility companies, who have to develop conservation programs, and to regulatory agencies, who are charged with the responsibility of establishing a conservation policy. Consequently, anyone concerned with the evaluation of conservation programs should be familiar with the procedures employed in meta-analysis and the benefits that can accrue from this process.

The techniques that have been developed for conducting meta-analyses are not complex. In fact, anyone with the knowledge to conduct a competent, quantitative evaluation, should be able to master the techniques that are employed in meta-analysis. The procedures require an understanding of the subject matter, a commitment to the value of the data aggregation, and attention to detail. Such qualities should not be difficult to find among the evaluation personnel currently involved in research on energy conservation.

REFERENCES

1. J.D. Claxton, C.D. Anderson, J.R.D. Ritchie, and G.H.G. McDougall, (eds.), *Consumers and Energy Conservation*, Prager, New York, 1981.
2. A. Baum and J.E. Singer, (eds.), *Advances in Environmental Psychology*, Vol. 3, (Energy Conservation: Psychological Perspectives), Lawrence Erlbaum Associates, Hillsdale, NJ, 1981.
3. C. Seligman and L.J. Becker, (eds.), *Journal of Social Issues, 37*:2, Spring, 1981.
4. H.J. Walburg and E.H. Haertell, Research Integration: An Introduction and Overview, *Evaluation in Education, 4*, pp. 5–10, 1980.
5. J. Sonderstrom, L. Barry, and E. Hirst, The Use of Meta-Evaluation to Plan Evaluations of Conservation Programs, *Evaluation and Program Planning, 4*, pp. 113–122, 1981.
6. D.B. Pillemer and R.J. Light, Synthesizing Outcomes: How to Use Research for Many Studies, *Harvard Educational Review, 50*, pp. 175–195, 1980.
7. G.V Glass, Integrating Findings: The Meta-Analysis of Research, in *Review of Research in Education,* L. Shulman (ed.), Vol. 5, F.E. Peacock, Itasca, Illinois, 1977.
8. Energy Conservation Group. *The Energy Conservation Programs and Research of California's Major Energy Utility Companies,* 1977–1980. University of California, Santa Cruz, 1982.

9. H.M. Cooper and R. Rosenthal, A Comparison of Statistical and Traditional Procedures for Summarizing Research, *Evaluation in Education, 4,* pp. 33-36, 1980.
10. G. V Glass, Primary, Secondary, and Meta-Analysis of Research, *Educational Researcher*, November, 1976.
11. R. S. Mansfield and T. D. Busse, Meta-Analysis of Research: A Rejoinder to Glass, *Educational Researcher*, p. 3, October, 1977.
12. G. V Glass, Reply to Mansfield and Busse, *Educational Researcher*, p. 3, January, 1978.
13. M. L. Smith and G. V Glass, Meta-Analysis of Psychotherapy Outcome Studies, *American Psychologist, 32*, pp. 752-760, 1977.
14. H. J. Eysenck, An Exercise in Mega-Silliness, *American Psychologist, 33*, p. 17, 1978.
15. G. V Glass and M. L. Smith, Reply to Eysenck, *American Psychologist, 33*, pp. 517-519, 1978.
16. T. D. Cook and C. L. Gruder, Meta-Evaluation Research, *Evaluation Quarterly, 2:*1, pp. 5-51, February, 1978.
17. R. Rosenthal, Combining Probabilities and the File-Drawer Problem, *Evaluation in Education, 4*, pp. 18-21, 1980.
18. M. L. Smith, Publication Bias in Meta-Analysis, *Evaluation in Education, 4*, pp. 22-24, 1980.
19. R. J. Light and P. V. Smith, Accumulating Evidence: Procedures for Resolving Contradictions among Different Studies, *Harvard Educational Review, 41*, pp. 429-471, 1971.
20. G. V Glass, R. McGaw and M. L. Smith, *Meta-Analysis and Social Research*, Sage Publications, Beverly Hills, California, 1981.
21. L. V. Hedges, Distribution Theory for Glass's Estimator of Effect Size and Related Estimators, *Journal of Educational Statistics, 6:*2, pp. 107-128, Summer, 1981.
22. _____, Unbiased Estimation of Effect Size, *Evaluation in Education, 4,* pp. 25-27, 1980.
23. W. A. Stock, M. A. Okun, M. J. Haring, W. Miller, C. Kinney and R. W. Cuervorst, Rigor in Data Synthesis: A Case Study of Reliability in Meta-Analysis, *Educational Researcher, 11:*6, pp. 10-14, June-July, 1982.

Direct reprint requests to:

Brian Stecher
815 Colorado Boulevard
Suite # 606
Los Angeles, CA 90041