META-ANALYSIS

The Systematic Synthesis of Research and Evaluation

OED Program Evaluation Seminar

April 1996

Gregg Jackson
202-882-0375

Ray Rist
202-998-3306

I. INTRODUCTION

A. Reviews of a set of evaluations on a given topic can help guide policy development by:

1. Simplifying: Inferring generalizations about factors that are associated with project outcomes: project planning, project components and processes, beneficiary characteristic, and contextual factors.
2. Complexifying: Examining additional interaction effects not testable within any one evaluation--such as when characteristics of beneficiaries and contexts are homogenous within each evaluation, but vary among the evaluations.

B. Meta-analysis is a conceptual framework and a growing bag of methodological tools that can be helpful when conducting reviews of a set of evaluations.

1. Most of the procedures in wide use today are applicable when you have large numbers of evaluations--usually 30 or more, but I will also discuss some procedures that are applicable when you have only four or five evaluations.

C. It was developed in the 1970's, though we now know that several famous statisticians such as Fisher and Cochran did some preliminary conceptual and statistical work on the issues in the 1930s.

D. Meta-analysis is most applicable to your Annual Review of Evaluation Results and your "Evaluation Studies," but derivative small sample procedures are applicable to some of your cluster PARs and cluster Impact Evaluations.

1. In addition, meta-analysis has implications for the individual PIFs because all are used in the Annual Reviews, and for individual PARs and impact evaluations that are expected to be used later in Evaluation Studies.
2. Lastly, meta-analysis is now frequently used in reviews of social science research, so you are likely to encounter them when reading such reviews.

E. Meta-analysis is particularly helpful when a set of evaluations appear to have disparate results, because it seeks to explain the variance in such results.

F. Types of traditional reviews:

1. When challenged about a pre-formulated impression of the truth on some matter, the evaluator seeks out two or three evaluations that can be cited in support of the impression.

    EXAMPLE: Whenever in a draft paper we are asked to "substantiate" a statement that we thought was obvious or widely recognized as true.

2. When curious about a matter, the evaluator examines a few easily available evaluations on the subject and concludes that whatever is common in their findings is true.

3. When seriously interested in a matter, the evaluator seeks out all readily-available evaluations on a given topic, and when the results differ (and they usually will), uses one of the following strategies to determine which results are correct:

    a. Scrutinize the methodology of the evaluations having what you "think" is the erroneous result, sighting all the methodological shortcomings, and then concluded the evaluations with the other result are the valid ones--without scrutinizing them for possible methodological shortcomings.

    b. Identify the one evaluation that you consider to be best methodologically, and rely on its results;

    c. Conclude that unless more than half of the results are statistically significant in one direction, the evidence is inconclusive and thus more and better evaluation is needed.

        EXAMPLE: 8 sig(+) and 17 others

            BUT: if were 8 sig(+), 12 nonsig(+), and 4 nonsig(-), and 1 sig(-) the evidence would be overwhelming in favor of a positive effect.

    d. Scrutinize the methodology of all the evaluations, as well as the nature of the interventions and the contexts in which they took place, and then make judgments about which results are likely to be true under various contexts.

II. META-ANALYSIS OVERVIEW

A. Churchill observed that "true genius resides in the capacity for the evaluation of uncertain, hazardous, and conflicting information."

1. Meta-analysis is tool intended to help those of us without true genius to cope effectively with the same challenge

B. DEFINITION: Meta-analysis is a systematic, quantitative, and replicable approach to integrating or synthesizing results from a set of research studies or evaluations on a given topic.

1. The term "meta" is used not in the manner of "metaphysics" which means "fake" physics, but rather in the manner of "meta-theory" of philosophy which means an over-arching theory that subsumes several existing ones.
2. There is a term "meta-evaluation" which is generally used to mean an evaluation of a single evaluation; others refer to this an "evaluation audit" much as OED does in its "PARs."

C. Meta-analysis begins with the presumption that any set of evaluations on a given topic will have some variation in the results because of random sampling error and random measurement error, and additional variation in the results due to variations in the characteristics of the planning, project components and processes, contexts, characteristics of intended beneficiaries, outcome measures, and the evaluation methodology.

1. Meta-analysis assumes that most of the evaluations on a topic are not perfect replicates

D. Meta-analysis is usually applied to <u>all</u> evaluations meeting broadly specified criteria, but narrower criteria can be used when appropriate, and so can a sample from all evaluations meeting specified criteria.

E. Each evaluation's results are expressed with a common metric; various characteristics of the planning, projects, contexts, beneficiaries, and evaluation methodology are also coded with common metrics.

F. Statistical analyses are used:

1. To summarize characteristics of the project components and processes, beneficiaries, contextual factors, and outcomes over the full set of evaluations and over various subsets of the evaluations.
2. To explore associations among characteristics of planning, projects, contexts, beneficiaries, and the outcomes.
3. To explore the extent to which various methodological shortcomings in the evaluations are associated with the magnitude and direction of the outcomes.

G. For those of you who have worked on the <u>Annual Reviews</u> during the last few years, the procedures of meta-analysis is not entirely new, because the first few chapters of the Annual Reviews have used several components of meta-analysis.

H. Meta-analysis is an **approach** to doing integrative reviews, not a **single methodology**.

1. The fundamental underlying principle of meta-analysis is that a review of prior evaluations is itself an evaluation, and should be conducted according to the guidelines for good evaluations. Because of that, much of what I will say about doing good reviews will be a reiteration of what Ray and I have previously said about doing good program evaluations.

I. Just as there are various methodologies that can be used for individual evaluations, there are a variety of methodologies that have been used for meta-analysis, and more are sure to be developed in the coming years.

J. The use of meta-analysis in the conduct of reviews of behavioral and social science has spread dramatically in the past decade. For instance in the leading journal of psychology reviews, (Psychological Bulletin) it was used in only 7 percent of the 1981-83 articles, but between 1986-89 it was used in 23 percent of the articles. Meta-analysis is also now being used fairly regularly in reviews of the medical research that are published in JAMA, The New England Journal of Medicine, and Lancet.

III. BASIC TASKS OF META-ANALYSIS

A. There are five basic tasks for meta-analysis:

1. Formulate the problem
   (Cooper & Hedges, Chaps. 2-3).
2. Search the literature
   (Cooper & Hedges, Chaps. 4-7).
3. Examine and code the primary evaluations
   (Cooper & Hedges, Chaps. 8-17)
4. Analyze the coded data
   (Cooper & Hedges, Chaps. 14-15, 18-26).
5. Report the results
   (Cooper & Hedges, Chaps. 27-29).

IV. FORMULATE THE PROBLEM

A. It is helpful when formulating the problem to examine prior reviews of research and prior Evaluation Studies on the same or similar topics, as well as some individual evaluations on the topic. They usually will provide useful ideas on:

1. Identifying theoretical issues that should be considered when planning the coding and analysis of the individual evaluations.

2. Identifying important mediators and contextual variables that you might want to code and analyze.
3. Identifying important differences in the methodology of the evaluations that you might want to code and analyze.
EXAMPLE: Study Brief: Review of Gender Issues in Bank Lending, by Donaldson???, pg. 14: Mentions contacts with other development agencies, some of which are undertaking reviews of their own experience with gender issues.

B. Seek input of all major stakeholders
1. Seek input in respect to:
a. questions to be addressed in the review
b. kinds of evidence to be examined
c. types of analyses to be conducted

2. Seek input from:
a. Bank board members
b. Bank operations staff
c. country officials: tell them you will be reviewing a substantial number of operations in a given area to learn more about how to maximize the effectiveness of those efforts, and ask them what issues they think should be explored.
d. implementing agency officials
e. NGOs
f. OED colleagues
NOTE: OED staff do seek some input when planning studies--they talk with operations staff, prepare a study design that is reviewed by a few people, and prepare an Approach Paper which is reviewed by the JAC (often after the project has been started). But input is limited, partly because of fear of loosing independence, and partly perhaps just by tradition.

C. Frame and select the questions
1. Comprehensively map out possibilities
2. Nest: over-arching, primary, secondary
3. Operationalize the primary and secondary questions
4. Investigate availability of data in the evaluation reports for answering the primary and secondary questions
a. Even if the data is not in the PARs, it might be in other sources such as the PIFs, national statistics, etc.
5. Select questions to be addressed in review, on basis of:
a. Importance
b. Ability to answer with the available data
- For the first few chapters of the Annual Reviews you are pretty much limited to the data on the PIFs

c. Try to select the questions to anticipate and address the challenges critics are likely to raise

- That isn't easy because who will be the critic depends on the results of the evaluation and you won't know for sure what they will be--
- But the range of results is importantly determined by the foci of the evaluation.

V. SEARCH THE LITERATURE

A. The literature search for OED reviews normally is a fairly straight forward process of identifying all projects with a given focus for which a ICR has been submitted, or a PAR completed.

1. Complications arise, however, when, over a period of years, sectors or regions have been redefined
2. Other complications?

B. When preparing for reviews of research, the task is much more complicated. Computerized indexes provide the most efficient search for journal articles, doctoral dissertations, and books. But in some fields there may not be one index that covers all three types of publications, and two or more indexes must be used. Unpublished literature is rarely indexed, and searching for it requires checking bibliographies of published research, and contacting researchers who are known to have been working on the topic.

C. OPRIS has indexes to other Development organization's reports and the research literature. Charles Schueller will conduct a separate session to introduce you to what OPRIS can do for OED staff and I encourage you to attend, because when Schueller gave an introduction to OPRIS during our first program evaluation seminar this spring, participants found it very helpful.

D. In some cases there may be hundreds of individual evaluations on the topic.

1. If the analysis is to be limited to just data already coded on the PIFs, little time is saved by analyzing a sample of the full set.
2. But if the ICRs are to be further coded, or the PARs coded, a sample may be necessary to complete the review within allotted resources. In those cases, simple random samples or stratified random samples will

usually be preferable,

a. though there may be good reason to also draw a purposeful sample of extreme cases or critical cases.

b. when the review is being conducted to predict the effectiveness of a proposed, but not yet implemented project, a similar cases sample should used, purposefully drawing the evaluations of prior projects that are most similar to the proposed ones.

EXAMPLE: The Bank's Experience in the Telecommunications Sector did a statistical overview of all 80 projects subject to a ICR by a given date, a detailed analysis of a "representative sample of 21 projects" (with no explanation of how those projects were select--as typical in OED reports), and an in-depth examination of the full telecommunications portfolios of seven countries (with some explanation of why they were selected.) Very Nice Zooming!

(1) Additional coding of some or all of the projects was done, but as is typical in OED reports and undesirable, the form is not appended nor is its availability referenced.

VI. EXAMINE AND CODE THE EVALUATIONS

A. The reviews done for the first chapters of the Annual Review are based on data already coded on the PIFs.

1. The results of the meta-analyses will depend not only on what characteristics have been coded in the PIF, but also partly on the reliability and validity of that coding.

B. Thematic reviews in the Annual Review are based on PARs and Impact Evaluations and they address issues of planning, project components, beneficiaries, and contexts beyond those coded on the PIFs. 1. Decisions have to be made about what characteristics should be coded. Ideally we would code everything possible, but with constraints on time and resources, there are inevitably trade-offs.

2. Questions that might be asked when deciding what to code include:

a. What is important for the review to be able to answer the high priority questions?

b. What does theory, research, popular thinking, prior evaluations, or insightful conjecture suggest is important to code?

c. Is there more than one indicator of an important variable that

should be coded?

d. Is there likely to be at least moderate variation across evaluations in each characteristic that you are considering coding?

e. Will most of the evaluation reports provide sufficient information for reliable coding of each characteristic?

f. How much incremental time will it take to code a given characteristic?

C. Characteristics that may be desirable to code in addition to those on the PIF form include:

1. Characteristics of the project, such as
   a. specific components
   b. intensity
   c. duration
   NOTE: the PIF covers project objectives and some very general judgments about the design (how demanding, innovative), but asks very little about specific project components and project processes
2. Characteristics of beneficiaries, such as:
   a. mean age
   b. percent who are legal residents
   c. average land holdings
   d. average educational achievement
   NOTE: the PIF does not ask for information on the initial characteristics of the intended beneficiaries.
3. Characteristics of the contextual factors
   a. degree of cultural traditionalism
   b. land tenure laws
   NOTE: the PIF has two pages of ratings of "Factors affecting achievement of major outcomes, but they are necessarily stated generally to be applicable to a wide range of operations
4. Characteristics of the outcomes, such as:
   a. number of new teachers trained and employed
   b. number and portion of children enrolled in school
   c. achievement gains of children in school
   d. any likely side effects
   NOTE: the PIF focuses mostly on degree of achievement of various outcomes, but again the ratings are in respected to broadly stated matters rather than specifics such as those just mentioned
5. Characteristics of the evaluation methodology
   a. quality of baseline and post-project data on the outcomes of interest
   b. number of project sites and beneficiaries

c. breadth and intensity of data collection on project processes and contextual factors that may have influenced the outcomes

d. inclusion of comparison sites in the evaluation

NOTE: the PIF calls for two ratings of the M & E, but no description of the actual methodology

6. The conclusions you infer about what factors appear to influence the project outcomes will be importantly determined by what characteristics are coded and analyzed. Just as is the case with mis-specified regression equations and econometric modeling, if you don't include the variables that really influence the outcome of interest, you may reach false conclusions.

a. You can get a high multiple R without including the actual causal variables if you include other variables that co-vary with the causal ones.

D. Coding problems in meta analysis

1. Books on meta-analysis are often filled with formulas for coding the outcomes of evaluations. The formulas aim at generating a comparable outcome metric over a set of evaluations that differ in the measures used and the statistics reported. The most common metrics are Glass's Effect Size (ES), the Pearson product-moment correlation, and Hedges' g). Numerous equations are needed to transform data or estimate equivalencies.

a. For instance, adult literacy programs' impacts on participants' reading skills might be measured on three different reading tests, and some studies might test program effects with a two-sample "t" test, some with a three- or more sample "F" test, and others with analysis of covariance tests.

b. OED is spared the considerable hassle of producing comparable metrics in its Annual Review of Evaluation Results, because the outcomes analyzed are OED staff judgments recorded on the PIF form. In the April 1994 PIF these are a rating of overall project performance, a rating of the extent of institutional development, a re-estimated ERR, and judgments as to whether the project is outstanding and whether sustainability is likely.

a. All the evaluations are to be judged in respected to the same criteria and same metric. Whether that actually occurs depends on the reliability of the coding of the PIFs, which a March 1996 test by OEDD2 averages only about 55 percent.

c. In the evaluation studies (reviews), cluster PARs, cluster Impact Evaluations, additional coding generally will be needed, but it too would mostly involve staff judgments which can be made in reference to given criteria and metric.

2. Another coding problem that plagues meta-analysts is that important characteristics of interventions, beneficiaries, and contexts often are inadequately reported in the research and evaluation reports that they have assembled.

a. This is the same problem that bedevils OED staff when reviewing ICRs and filling in PIFs, but OED can potentially control the adequacy of reporting in the PARs and impact evaluations that feed into its evaluation studies.

(1) For instance, advance planning would permit OED to specify the characteristics to be examined and the manner in which they are reported in the PARs and Impact Evaluations.

(2) Indeed, forms to be filled in might be prepared for PARs and Impact Evaluations that are expected to be used in subsequent evaluation studies, to assure that the information anticipated as needed for the evaluation study will be collected and reported in a consistent manner when preparing the PARs and Impact Evaluations.

(a) A few Cluster PARs and cluster Impact Evaluations have recently been prepared in OED with advance planning to meet the needs of subsequent Evaluation Studies.

3. The reliability and validity of coding can be a serious problem in meta-analyses.

a. The coders should be trained thoroughly;

b. Their coding accuracy should be monitored at random;

c. Retraining should be provided as needed.

d. Checking reliability

(1) After appropriate training, over short periods of time, it is usually sufficient to compute inter-coder reliability.

(2) Over two or more months, there is a need to check coding against fixed benchmarks, because all coders may drift.

(3) Even with these precautions, for the highest reliability, it is desireable to have each evaluation coded independently by two different persons, and then for a senior staff person to meet with both to resolve differences.

4. Meta-analysts have noted that the unreliability of independent and dependent variables will attenuate measures that indicate the magnitude of outcome effects. So will range restriction.

a. Hunter and Schmidt have developed procedures to correct for these downward biases in measures of effects. Corrections can be applied to each primary evaluation if the needed information is available, or applied to the average effect computed for the

analyzed set of evaluations (Cooper and Hedges, Chap. 21).

VII. ANALYZE THE DATA

Someone observed that "Many people use statistics as drunken men use lamp posts--for support rather than illumination."

A. Meta-analysis leaves plenty of room for intuition, wisdom, & insight, but requires that each be verified empirically--to the extent possible.

B. Meta-analyses usually summarize results and also search for correlates of the variation in the results. The correlates are usually examined univariately and also by multiple regression or related multivariate techniques.

C. Common analyses in meta-analysis include:

1. Calculation of the average performance on each outcome.
2. Recalculation of the average performance on the outcomes for subsamples of interest.
3. Assessment of the extent to which variations in methodology of the evaluations are associated with variations in the outcomes. a) Examination of how characteristics of the planning, projects, beneficiaries, and contexts are associated with variations in the outcomes
    a. examining BOTH the univariate and multivariate relationships.
    EXAMPLE: The 1991 Annual Review used univariate analyses almost exclusively, but the 1992 version did report several multivariate ones.
4. More than once I found statements in Study Designs similar to "**the analysis would identify main characteristics of projects where sustainability is likely.**" The more useful analyses would be of the characteristics that distinguish between projects judged sustainable and non-sustainable.
    a. We have, of course, seen comparative analyses in many OED reports. But, the above statement does not seem to be merely a casual use of the language, but rather to reflect some tendency in OED to avoid comparative analyses when they would be appropriate.
        - For instance, in both the 1991 and 1992 Annual Reviews there is a description of characteristics common to projects judged outstanding, but no comparison with the frequency of those characteristics in projects judged

unsatisfactory and frequency in those judged satisfactory but not outstanding.]]

b. Remember, for you to establish empirically a causal relationship between project components and outcomes, there must be a association, temporal ordering, and isolation; for there to be association between the project components and outcomes, there must be variations among the components and among the outcomes of the sample of evaluations.

D. Numerous supporting or auxiliary analyses are also possible, including:

1. If substantial associations are found between shortcomings in the methodology of the primary evaluations and the results, the analyses can be conducted with statistical controls for the methodological shortcomings or limited to the subset of evaluations that is free of those shortcomings. It should be remembered that methodological flaws in a evaluation constitute "threats" to the validity of the evaluation but do not always bias the results.

2. If characteristics of beneficiaries or contexts are not represented in proportion to their existence in the universe, the evaluations should be treated as coming from a stratified random sample and then averages should be calculated in a weighted manner.

3. Sensitivity analyses can be conducted to test the effects of any decisions about the methods of the meta-analysis that might be questionable.

5. Outlier analyses can be conducted

E. OED's Annual Review of Evaluation Results generally treats the set of reviewed evaluations as a census--all evaluations completed within a given year.

1. Consistent with that view, it generally reports only descriptive statistics, not inferential ones that test statistical significance.

F. Similarly, each OED study generally treats the set of reviewed evaluations as a census of all evaluations on a given intervention.

1. Consistent with that view, the studies generally report only descriptive statistics, not inferential ones testing statistical significance.

G. Like any multivariate data base, meta-analysis data bases provide the opportunity for hundreds of analyses. Whenever a large number of statistical tests are conducted, there is the risk of capitalizing on chance. To minimize this risk it is possible to:

1. Limit the statistical tests to a small number of hypotheses that are judged to be of high priority.

2. Do statistical tests for all hypotheses that might prove interesting,

but:

a. Use a very low alpha level, usually .01, (but this increases the chances of failing to reject null hypotheses when they are false).

b. Disregard the statistically significant results that have low magnitudes or are just barely statistically significant.

c. Disregard the statistically significant results that do not comprise part of a congruent pattern of results.

H. Multiple regression and other multivariate statistics are wonderful tools, but two precautions should be kept in mind.

1. Most fit equations using linear procedures; if the univariate relation between a predictor and outcome is strong but non-linear, it will appear to be a weak relationship. CHECK FIRST with scatter diagrams!

2. They determine multivariate associations, but those associations are not, in themselves, proof of causation.

a. To prove causation you must establish that all likely causal variables were included in the equations, and you have to establish that the predictors preceded the outcome (temporal ordering).

VIII. REPORT THE RESULTS

A. In research work, the scientific standards call for reporting of the sampling, interventions, measurement, and analyses in sufficient detail that another competent professional can replicate the investigation. The same should apply to reporting meta-analyses, and when that would make the report too bulky, the report should reference publicly available documents that provide these details.

B. The report should indicate the reliability achieved in the coding of the evaluations.

C. The report generally should include a table showing the salient characteristics of each individual evaluation, including characteristics of the planning, sample, project features, beneficiaries, contexts, and outcomes.

1. This is just an expanded form of the sampling matrix that I suggested you include whenever sampling from less than 100 studies.

D. The report generally should include one or more tables showing the key results for whatever subset analyses have been conducted and whatever multivariate analyses have been conducted.

E. Several kinds of conclusions can be drawn from meta- analyses, including:

1. Testing old theory or inferring new theory.
2. Testing the effectiveness of established policies or inferring new ones.
3. Testing the effectiveness of established practices or inferring new ones.
4. Testing the external validity (generalizability) of the evaluation findings over different planning, subjects, forms of the project, and contextual features.

IX. THREATS TO THE VALIDITY OF META-ANALYSES

A. The threats to the validity of meta-analyses include all threats to the validity of the individual evaluations as well as additional threats that occur in the meta-analysis procedures.

B. Threats to the validity of the individual evaluations that also threaten the validity of meta-analyses include:

1. Internal validity--(these threats can cause spurious increases or decreases in observed effects)
    a. intervention corruption
    b. history
    c. maturation
    d. measurement decay
    e. selection bias
    f. mortality
2. External validity--generalizability
    a. poor sampling
    b. Hawthorne effect
    c. disruption effect
    d. John Henry effect
    e. history-intervention interaction
    f. selection-intervention interaction
    g. context-intervention interaction

C. Threats to validity from the meta-analyses procedures include:

1. Inadequate information for correctly coding the various characteristics of the evaluation.
2. Unreliable coding when the needed information is available.

3. Biases due to the lack of independence among multiple outcomes from a given evaluation.
4. Failure to weight each evaluation by its likely precision (by sample size, or the inverse of the standard error).
5. Lack of statistical power when conducting the meta-analyses due to a small number of evaluations or a small number having the subset characteristics that are analyzed.
6. Capitalizing on chance by conducting a large number of statistical tests during the meta-analysis without adequate control for the accumulation of alpha errors.

X. CRITICISMS OF META-ANALYSIS

A. Some critics argue that the reductionist coding of evaluations and statistical analysis of the resulting data precludes a thorough and holistic understanding of the evaluations.

1.This is a legitimate concern, but there are trade-offs between traditional reviews and meta-analysis, much as with the advantages and disadvantages of qualitative and quantitative research approaches.

a. Traditional reviews are not mutually exclusive with meta-analysis. Meta-analysis can be incorporated into them.

B. Some critics claim that the meta-analysts' preference to examine all evaluations on the topic means that they often draw inferences from methodologically inadequate evaluations.

1.Meta-analysts disagree, saying that many methodological weaknesses in evaluations do not necessarily bias results. Rather they create threats to validity, and it is an empirical question whether the biases actually occur.

a. If preliminary analyses show that the methodologically flawed evaluations yield results similar to the methodologically strong evaluations, then to discard the former evaluations is seen as an unjustified waste of good information.

b. On the other hand, if the preliminary analyses show the weak evaluations do yield different results, and the bias is in the direction suggested by their weakness, the weak evaluations can be discarded from the analyses or their weaknesses can be statistically controlled.

2.Meta-analysts point out that traditional reviewers often are unsystematic in their selection of evaluations that they allege are exemplary, and consequently their selection process can easily be biased towards evaluations with results that are congruent with the reviewer's personal views of the topic.

C. Some critics claim that because meta-analysts seek a large sets of evaluations for multivariate analyses, they are driven to define their topics broadly, and consequently tend to "mix apples and oranges."

1. Meta-analysts respond that it is an empirical question whether apples and oranges differ importantly in respect to a given question.
2. In addition, they argue that when you wish to make reliable inferences about "fruit" you should analyze not only apples and oranges, but also melons and berries.

D. Some critics claim that meta-analysis substitutes statistical precision for conceptual clarity. By reducing the review task to coding and statistical analyses, it can be done mechanically and mindlessly.

1.Anyone who has paged through The Handbook of Research Syntheses would realize that proponents of state-of-the-art meta-analysis do not treat it as a number crunching process, but rather as intricate analyses of complex phenomena.
2. On the other hand, some practitioners have conducted meta-analyses on numerous topics, with limited background knowledge, and have not even read all the evaluations, but instead have trained research assistants to code them. Under these circumstances there is good reason to fear that statistical precision has been substituted for conceptual clarity.

XI. ALTERNATIVES TO META-ANALYSIS

A. Predominantly narrative reviews are still appropriate when there are only a few evaluations on a topic.

1. And there is nothing that prohibits marrying the holistic orientation, dialectical approach, and conceptual clarity of strong narrative reviews with the analytical precision possible with meta-analysis.

B. Secondary data analysis pools data from several quantitative evaluations, and then applies standard statistical analyses. Guidelines for sophisticated secondary analyses go back at least to the early 1970s (Light and Smith, 1971; Hyman, 1972).

1. This approach is often precluded by lost data sets, inadequately documented data files, and time and financial constraints.

C. There are several forums for weighing evidence on a topic. These include:

1. The National Academy of Sciences which uses experts to examine published and other evidence on a topic. The experts meet several

times over a period of nine to twenty-four months, and then issue a report of their findings. Usually the findings represent a consensus, but sometimes dissents remain and are reported.

2. Nearly 100 consensus development conferences have been conducted by the National Institutes of Health since 1977. A chairperson and a few NIH staff members spend several months collecting, synthesizing, and presenting evidence on a given issue. The materials are then sent to a panel of experts for their review, and a few weeks later the panel convenes for three days to weigh the evidence and forge consensus statements (Kanouse, 1989; Institute of Medicine, 1990).

3. Quasi-judicial proceedings have opponents on a controversial issue present the best cases they can make for their respective views. The presentations are in a face-to-face setting, subject to procedural rules that include the opportunity for cross examination. Quasi-judicial proceedings can use a judge or juries, permit lengthy adjournments to prepare additional arguments and evidence, and accompany the "verdict" with a written opinion explaining the rationale for it. These proceedings were tried in the 1970s, but have not be widely used. (Arnstein, 1975; Jackson, 1977; Popham and Carlson, 1977; Wright & Sachse, 1977).

4. If OED has not already done so, it perhaps should experiment with using expert panels or quasi-judicial proceedings to examine its available evidence on important and controversial issues.

5. But no data or analysis, by whatever method or from whatever forum, will be accepted universally.

   a. As John Kenneth Galbraith observed: "faced with a choice between changing one's mind and proving that there is no need to do so, almost everyone gets busy on the proof."

ADDITIONAL READINGS IN THE OED LIBRARY

Cooper, H. 1989. Integrating Research: A Guide for Literature Reviews. Thousand Oaks, CA: Sage.

This is the best brief introduction to mainstream meta-analysis as it is used by researchers.

Cooper, H. and Hedges, L. 1994 The Handbook of Research Synthesis. New York: Russel Sage Foundation.

This is the most comprehensive, authoritative, and up-to-date book on meta-analysis. Despite being called a "handbook," and having chapters written by different authors, it is structured and written in the form of a textbook. Half of the book is filled with complicated statistical formulas, but most of those formulas are for computing comparable outcome measures across evaluations--a task that usually is unnecessary when meta-analyzing OED's evaluations.

Light, R. and Pillemer, D. 1984. Summing Up: The Science of Reviewing Research. Cambridge: Harvard University Press.

This is mostly a conceptual discussion of issues involved in integrating research and evaluation studies on a given topic. Chapter 3 discusses applications of exploratory data analysis for reviews. Chapter 4 provides a good discussion of using both qualitative and quantitative approaches in reviews.

U.S. General Accounting Office. Undated. The Evaluation Synthesis. Washington, DC: author (GAO/PEMD-10.1.2).

This is a brief but illuminating discussion of the use of meta-analysis in integrating evaluations. It repeatedly refers to the real-world problems that arise when doing such integrations, and suggests ways of dealing with the problems.

BIBLIOGRAPHY

Arnstein, G. E., (1975). Trial by jury: A new method of evaluation: The outcome. Phi Delta Kappan, 57, 188-190.

Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments. Journal of the Royal Statistical Society (Supplement), 4, 102-118.

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

Fisher, R. A. (1932). Statistical methods for research workers (4th ed.). London: Oliver & Boyd.

Glass, G. V. (1976). Primary, secondary, and meta-analysis. Educational Researcher, 5, 3-8.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). Meta-analysis in social research. Beverly Hills, CA: Sage.

Hyman, H. H. (1972). Secondary analysis of sample surveys: Principles, procedures, and potentialities. New York: John Wiley.

Hunter, J. E. (1980). Validity generalization for 12,000 jobs: An application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATB). Washington, DC: U.S. Employment Service, U.S. Department of Labor.

Hunter, J. E., & Schmidt, F. L. (1990). Methods of meta-analysis: Correcting error and bias in research findings. Newbury Park, CA: Sage.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). Meta-analysis : Cumulating research findings across studies. Beverly Hills, CA: Sage.

Institute of Medicine (1990). Consensus development at the NIH: Improving the program. Washington DC: author.

Jackson, G. B. (1977). Adversary evaluation--Sentenced to death without a fair trial. Educational Researcher, 6, 2 & 18.

Jackson, G. B. (1978). Methods for reviewing and integrating research in the social sciences. (Final technical report to the National Science Foundation No. PB-283-747). Springfield, VA: National Technical Information Service.

Jackson, G. B. (1980). Methods for integrative reviews. Review of Educational Research, 50, 438-460.

Kanouse, D. E., & et al. (1989). Changing medical practice through technology assessment. Ann Arbor, MI: Health Administration Press.

Light, R. J., & Smith, P. V. (1971). Accumulating evidence: Procedures for resolving contradictions among different research studies. Harvard Educational Review, 41, 429-471.

Ogawa, R. T., & Malen, B. (1991). Towards rigor in reviews of multivocal literatures: Applying the exploratory case study method. Review of Educational Research, 61, 265-286. [Also see the commentaries in the same issue.]

Popham, W. J., & Carlson, D. (1977). Deep dark deficits of the adversary evaluation model. Educational Researcher, 6, 3-6.

Rosenthal, R. (1978). Combining results of independent studies. Psychological Bulletin, 85, 185-193.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. Psychological Bulletin, 86, 638-641.

Rosenthal, R. (1984). Meta-analytic procedures for social research. Beverly Hills, CA: Sage.

Rosenthal, R. (1991). Meta-analytic procedures for social research (rev. ed.). Newbury Park, CA: Sage.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.

Wright, W. J., & Sachse, T. P. (1977). Payoffs of adversary evaluation. Paper presented at the annual meeting of the American Educational Research Association, New York.

Yates, F., & Cochran, W. G. (1938). The analysis of groups of experiments. Journal of Agricultural Science, 28, 556-580.

(\wp\wb\sem6met1)