Improving Meta-Analysis for Policy Purposes

Larry V. Hedges

Many empirical sciences have developed formal methods of combining information across independent research studies, an enterprise with a long history that was named "meta-analysis" (Hedges 1992; National Research Council 1992). When the question to be addressed is a narrow scientific one, the standard methods of meta-analysis provide adequate tools for combining the evidence. These are discussed in Cooper (1984), Hedges and Olkin (1985), Hunter and Schmidt (1990), Light and Pillemer (1984), Rosenthal (1984), or the new "Handbook of Research Synthesis" (Cooper and Hedges 1994) which includes contributions by all the authors previously mentioned.

Society is entering an era in which systematic research syntheses reasonably can be expected to contribute to the formation of public policy. In the area of health care research, this is already happening. In 1989, an act of Congress created the Office of the Forum for Quality and Effectiveness in Health Care within the Agency for Health Care Policy and Research. The forum was created to develop guidelines for clinical practice. A novel aspect of this effort to develop medical practice guidelines is that forum guidelines are required to be based on research evidence whenever possible (National Institute of Medicine 1990; Woolf 1991). Expert opinion or clinical judgment is substituted only when research evidence is not available to support some aspect of a guideline. The forum has already issued guidelines in a number of areas including the management of pain, depression, urinary incontinence, pressure ulcers, and cataracts, and other panels are currently developing guidelines on other issues. It is important to recognize that these clinical practice guidelines are practice policies and thus their development is an act of policymaking (Eddy 1990*a*, 1990*b*, 1990*c*; Woolf 1991, 1992.)

It is a matter of some concern, then, whether systematic syntheses of research can provide reliable evidence from which to gauge the likely effects of policies that might be adopted. The record of meta-analysis in providing valid syntheses of scientific research for purely scientific purposes is unassailable both from an analytic (deductive) standpoint and from an empirical standpoint. In medicine, meta-analytic conclusions have been repeatedly validated by larger clinical trials (see Chalmers et al. 1987) and in the physical sciences by subsequent experiments of higher accuracy (Rosenfeld 1975). However, the record of meta-analysis is not nearly so compelling in the arena of providing reliable evidence for policy purposes. Two examples illustrate the point.

In the health care field, the General Accounting Office's cross-design synthesis project (Silberman et al. 1992) explored a notable lack of correspondence between estimates of the effectiveness of experimental treatments derived from clinical trials and data derived from population surveys after those experimental treatments became the standard of practice. The clinical trials found that the experimental treatments could drastically reduce death rates among those treated for a particular disease. Consequently, one would expect to see the death rates from the disease drop as the new treatments became standard. However, the population survey data failed to validate the clinical trials estimates of the likely treatment effect when implemented as a practice policy.

A second example comes from educational research, particularly from syntheses of research on classroom learning. A series of such syntheses produced singularly unconvincing recommendations for policy, even though the research foundation is rather sound (Wang et al. 1993). Celebrated examples from this tradition include mastery learning methods; their efficacy and practicality are supported by an enviable body of research, but the practical applications have been disappointing.

The purpose of this chapter is not to provide a comprehensive review of previous work in meta-analysis, but to question its applicability for the purposes of drawing inferences for policy. It is argued that conventional approaches to meta-analysis are ill suited to inform many policy questions—not because they are technically flawed, but because they answer the wrong questions. Thus the failure is one of articulating the problem precisely and insuring that the methods are well suited to address the problem. Because these tasks (particularly stating the problem in a way that is useful for ensuring relevant statistical analysis) are very difficult, some researchers may have fallen into a trap that Tukey (1994) identified as a perennial problem in applied statistics: having a good answer to the wrong question. To avoid this trap, Tukey suggests researchers think carefully about the question and try to get an answer (even a poor answer) to the question that they really care about.

The genesis of the problem is that scientific research literatures consist of studies (experiments) whose designs are selected according to practical and scientific criteria. The criteria used in selecting the study context and variations of the treatment to be studied may change as research on a treatment progresses. Early in a research program, intensive variants of the treatment are likely to be studied in contexts or with subjects believed to be susceptible to the treatment; such studies may continue throughout the research program,. After the treatment efficacy is established under such highly favorable conditions, scientific interest may shift to the efficacy of less intense variants of the treatment under less favorable conditions. These less favorable conditions almost always correspond to the conditions under which treatments will be applied in practice, and hence are more relevant to policy questions. This is well known in evaluation research; for an interesting example in another context, see Feinstein (1985).

This chapter proposes a model or framework for thinking about the problems of drawing inferences from research literatures for policy purposes and suggests how this model may be used in research syntheses. It is argued that use of the model will reveal the nature of the research evidence available, identify knowledge gaps when evidence is unavailable, and better summarize the available evidence for policy decisions. By estimating components of variability, this model will also help quantify the likely generalizability of research findings.

Related Conceptualizations

The model proposed is in the same spirit as other models for inference from collections of studies. Cordray's policy space (Cordray and Fischer 1993) incorporates the idea of classifying studies according to treatment type (intensity) and context (subject type). Rubin's response surface model (Rubin 1990, 1992) incorporates the idea of classifying studies by study design and treatment type. Cronbach's model of construct generalization (Cronbach 1982) incorporates the idea that the relevant population (universe) about which generalizations are desired is multifaceted, including facets for context. Becker's (in press) model of generalizability of study results extends Cronbach's formulation to research synthesis.

The approach of this chapter is informal and generally nontechnical (although some of the content is by its nature technical), but the statements are precise. For example, the term "uncertainty" is used to refer to the variance of certain quantities without fully qualifying the random variables involved or whether it is to be taken as a subjective distribution or a classical sampling distribution (in most cases it could be made precise as either).

INFERENCE MODELS IN META-ANALYSIS

It is convenient to summarize inference models in meta-analysis within three categories: fixed-, random-, and mixed-effects models. These distinctions have been made in other contexts and have been applied before to meta-analysis (Cooper and Hedges 1994; Hedges 1992*;* Hedges and Olkin 1985). To understand these models, assume that researchers are interested in summarizing a collection of independent research studies, each of which can be described by a numerical index (such as a proportion, a correlation coefficient, a mean difference, or a rate ratio). In research synthesis, such indexes are generically known as indexes of effect size because they provide a quantification of the degree of relationship between variables. In any particular meta-analysis, it is usually desirable to work with the same type of index of effect size from all studies.

All three inference models distinguish the concepts of a population effect size or effect size parameter from that of a sample effect size or effect size estimate. When necessary, the effect size estimates from k independent studies are denoted by Roman letters subscripted by the study identification number and the corresponding effect size parameters by Greek letters. Thus $T_1 ... T_k$ might be the effect size estimates from k studies, $_1$,..., $_k$ are the corresponding effect size parameters, and T_i differs from $_i$ by an amount $G_i = T_1 -$ $_i$, which is usually referred to as a "sampling error." Except for biases that arise in some estimation conditions, sampling errors are due to variations across the samples of individuals that might be used to compute effect size estimates. Sampling errors arise because researchers estimate effect size in any individual study from a sample of finite (often quite small) size. If a study had a sample of infinite size available, there would be no sampling error.

The uncertainty of T_1 as an estimate of $_i$ is usually quantified by the standard error (the square root of the sampling error variance), which is denoted &$_i$. Indexes of effect size used in meta-analysis have a property that permits the sampling error variance to be analytically derived as a function of the effect size itself and the sample sizes; consequently, sampling error variances can be treated as "known" quantities and not as quantities that have to be estimated from replications in the data.

Fixed-Effects Models

Fixed-effects models are both the simplest and the most widely used statistical models in meta-analysis. They treat the effect size parameters as if they were fixed quantities. The parameters may differ across studies, but such differences are not thought of as a consequence of chance processes. The simplest fixed-effect model, and the model most often used in meta-analysis, treats all studies as having the same effect size parameter = $_1$ = . . . = $_k$. More complex fixed-effects models posit that the effect size parameters $_1$, . . ., $_k$ are a simple (usually linear) function of study characteristics. For example, the effect size might be taken as a function of duration or intensity of treatment, and fixed-effects models might be used to test whether studies with short duration or low intensity have smaller effect sizes than studies of long duration or high intensity.

Note that fixed-effects models make rather strong assumptions about the data. One is that between-study variations in effect size parameters are not the consequence of random processes, and thus do not add to uncertainty of summaries such as the average effect. However, various tests of model specification have been developed to determine if sample effect sizes are consistent with fixed-effects models (Hedges and Olkin 1985), and there is a considerable body of evidence that these models are often reasonably consistent with meta-analytic data.

Random-Effects Models

Random-effects models differ from fixed-effects models in that they treat the effect size parameters $_1$, . . ., $_k$ as if they were sampled from a universe (hyperpopulation) of possible effect size parameters. The conceptual model usually considers the observed studies as a (random) sample from a universe of studies that might have been observed. Since the studies are selected at random, their effect size parameters are a sample from a universe of effect size parameters. The object of the analysis is to estimate the (hyper-) parameters that describe this (hyper-) population of effect size parameters, usually the mean and the variance (which is often called the between-studies variance component).

Although random-effects analysis is superficially similar to fixed-effects analysis, yielding for example an estimate of the mean effect size and its uncertainty (in the form of a standard error), the meaning of these quantities is subtly different. The estimated mean in the random-effects model is the mean of a population of effect size parameters, and it is possible for the average effect size parameter to be positive but for some (perhaps a large proportion) of the effect size parameters to be negative. The characteristics of the distribution of the random effects (in particular, its variance) help determine how likely this is to occur.

Mixed-Effects Models

Mixed-effects models incorporate some of the characteristics of both fixed- and random-effects models. In mixed-effects models, the effect size parameters are partly determined by knowable characteristics of the studies (fixed effects) and partly the result of random processes. The models are typically employed by using a specified set of study characteristics as fixed-effects predictors of the effect size parameters, and defining any remaining between-study variation as random. One can think of this as defining a universe of studies that have precisely the same set of characteristics (the same values of the fixed effects) and treating the observed studies with that set of characteristics as a sample from that universe.

PROBLEMS WITH APPLICATIONS OF META-ANALYTIC MODELS FOR POLICY

Any of the meta-analytic models are quite capable of providing valid answers to the questions they are designed to answer. Unfortunately for policy purposes, they are usually used to address the wrong question. The models are frequently used to summarize studies that have been done; a simple summary of studies is rarely the answer to a question of real interest to policymakers.

PROBLEMS WITH THE RANDOM-EFFECTS CONCEPTUALIZATION FOR DRAWING POLICY RELEVANT INFERENCES

Random-effects models (as conventionally used) make the assumption that the sample of studies is a simple random (or at least, representative) sample of the universe of studies to which generalization is desired. This is an astonishingly naive assumption. Even if it happened to be true in any one case, the fact that multiple perspectives on the same issue would often prescribe different universes makes it impossible for the sample of studies actually conducted to represent all the policy relevant universes to which generalization is desired.

An Illustrative Example. Consider a simplified example of estimating the rate of drug use (or the effect of a drug use prevention program). Suppose that the rate (or effect) depends on the age and ethnicity of the target population, and assume for simplicity that there are two age groups (young and old) and two ethnic groups (African American and European American). Now consider a collection of equally valid research studies, each of which provides data on one of the age or ethnicity categories. How should one combine the information across studies to make an inference about the rate of drug use (or the effect of the prevention program)? It depends on the precise question one wants to answer.

In determining the rate for the entire population, all studies are relevant. But if one wants to know about the rate among young people, only some studies are directly relevant. Moreover, the average rate (or the average effect of the treatment) is highly unlikely to be the relevant summary. For example, if one is interested in young people in general and there are equal numbers of studies of African Americans and European Americans, then (since there are more European Americans than African Americans in the general population) the simple average will overweight the results of studies of African Americans.

Not only is the simple average unlikely to be the relevant summary, but its uncertainty is unlikely to be the relevant estimate of uncertainty for two reasons. First, the variance of the combined estimate depends on the variance of the estimates that go into it. If the uncertainty of the estimated effects within ethnic groups is not the same, misweighting the groups in the combined estimate also misweights data for the purposes of computing uncertainty. Second, even if the within-group uncertainty is the same for each race, misweighting the ethnic groups will lead to misweighting the between-group component of the uncertainty of the combined estimate.

A CONCEPTUAL FRAMEWORK FOR POLICY RELEVANT INFERENCE

The first step in a policy relevant synthesis is to classify the relevant variables that have a systematic effect on study results. It is reasonable to assume that effect parameters depend on three general categories of study characteristics: treatment type, study design, and study context. Treatment characteristics include all of the ways that the nominal treatment may vary systematically across studies including the duration, intensity, and mode of treatment administration. When the treatment itself is diffusely defined, the particular variety of treatment is a relevant variable here. Note that unplanned variations in treatment implementation would not be included as variables here because they are not controlled and add to unsystematic variation.

Study design characteristics include all of the systematic aspects of the research design and procedure except those that are part of study context. These characteristics include procedures used to ensure internal validity (the conventional meaning of experimental design) as well as characteris-tics of the outcome measures used.

Study context characteristics include aspects of the target populations and the settings in which the research study was conducted: all of the usual demographic characteristics of the subject population, and the character-istics of treatment settings as well (e.g., a school-based, community-based, or individualized program). Obviously there will be some ambiguity among these categories of variables, and some treatment types can occur only in some contexts, but in any given policy question, decisions (albeit arbitrary ones) can be made to classify a variable in that way for the purposes of the analysis.

Treatment Type and Context Define the Estimand

The technical development of valid statistical inference depends on unambiguous statement of the quantity to be estimated. Researchers often forget this point when working in areas where the statistical procedures and underlying conceptual models are so well understood as to be conventional. However, problems are often complicated by conceptual ambiguity about the quantity to be estimated. The purpose of this lengthy theoretical development is to provide a framework for achieving clarity on what should be estimated to help meta-analytic summaries better inform policymaking.

In order to define the inference problem precisely, it is necessary to define the treatment type and the study context variables. To be clear about what treatment effect to estimate, researchers must know which treatment variations to count as implemented, with which subject population, in which settings.

The treatment type variables serve to define the treatment itself. Researchers may wish to draw inferences about the likely effect of any particular subtypes of treatment or about a mix of them. If the treatment mix is of interest, it is important to recognize that changing the propor-tions of each subtype in the mix may change both the estimate of the overall effect and its uncertainty.

A more technical way to put the above argument is that, to specify the population to which researchers wish to generalize, the distribution of contexts and the treatment types must be specified. The most convenient way to do that is to define a context stratification system and specify the population weights given to each cross-classified context stratum.

Study design characteristics reflect the standard of evidence of internal validity to be applied in drawing inferences. In principle, these characteristics could be considered technical parameters. The policy-maker is unlikely to be interested in the relations between these characteristics and effect size, in and of themselves, although scientists studying quasi-experimental design would find them substantively interesting. The policymaker wants to know what the effect is, not what design features lead to biases (unless this information helps interpret evidence). An optimist might consider the study design characteristics as a way to categorize the departures of existing studies from hypothetically perfect studies (i.e., studies that provide an unbiased estimate of a conceptual treatment effect). In fact, Rubin (1990, 1992) has suggested that researchers estimate a "response surface," precisely characterizing the relation between study design characteristics and effect size in order to estimate the effect size of such an ideal study.

Estimation and Inference

After an estimand has been precisely defined by specifying the relevant distribution of contexts and treatment types, the problem becomes one of estimating the mean and uncertainty (variance) of the treatment effects as a well-defined statistical problem. It is most natural to carry out the estimation in the context of a random- or mixed-effects model, although the analysis requires some modification of existing methods to accommodate the weighting used to define the relevant distribution of contexts. This would involve a reasonably straightforward adaptation of methods that are already well developed in the analysis of stratified samples from survey designs (Cochran 1977). In principle, these methods would involve stratifying the sample of studies and then carrying out a meta-analysis within each of the strata using standard methods. These summary statistics from the meta-analyses would then be combined using a weighted combination procedure similar to those used in the analysis of stratified sample surveys. In practice, a few difficulties will arise.

Methods for Weighted Combination of Meta-Analyses Do Not Exist

Even with a modest set of context and treatment type strata, some (and perhaps many) of the strata will have no studies. That is, the stratified sample will have missing data. Note that this is a limitation of the data, not a limitation of the synthesis method. This limitation is a strength of the method—it forces researchers to confront the fact that the available data are not adequate to provide an empirically based estimate of the relevant treatment effects.

When faced with missing data, there are three choices: get more data, substitute assumptions for data, or change the question. The first option is typically the best, but least immediately feasible. However, it is important to note that the identification of missing data in a synthesis is equivalent to identifying studies that need to be done and whose results would reduce uncertainty in policy relevant inferences.

One practice in dealing with missing data is imputation of missing data (or more sophisticated model-based inference under models that include missing data in their specification). In this case, the assumptions that substitute for the data are embodied in the imputation (or missing data) model (Little and Rubin 1987; Rubin 1987). Here the assumptions concern the relation between the observed data and the missing data, so that empirical evidence plays some role in values substituted for the missing data.

A different way of adjusting for missing data is to go entirely outside the data set and use expert opinion. Estimates derived via expert opinion could be used in place of empirical research results in strata where data are missing. There are many methods of gathering such information, including a considerable literature on how to elicit prior information for Bayesian statistical analyses (Kadane et al. 1980; Winkler 1967). One particular advantage of the sampling frame for contexts is that it narrows the domain about which expert opinion is elicited. Expertise, by definition, is a consequence of substantial experience and is necessarily context bound. The use of relatively narrow contexts helps make it possible to ensure that the experts provide information within their domains of expertise. Within those domains, it is quite likely that expert opinion can be satisfactorily substituted for empirical research results. Indeed, the adequacy of expert opinion could be monitored by eliciting expert opinion in domains where satisfactory empirical evidence already exists.

Knowledge of Population Weights

In order to specify the population of interest, it is necessary to specify the population weights for each stratum. It is probably possible to specify strata for which these weights would be unknown. However, it seems obvious that these weights are of critical interest; knowing the composition of the target population and settings to affect is critical to formulating wise policy. Perhaps one should be wary of any tools that purport to yield targeted evidence on policy that do not also require knowledge of who is to be affected and in what contexts.

It may not be critical that the weights be known exactly. If effects do not vary profoundly across adjacent strata, then modest variations in the weights will produce only small variations in the overall effect. (If effects do vary profoundly across strata, one should be cautious about averages because they may obscure real variation.) Examining alternate possible values of the weights will permit bracketing of effects and sensitivity analysis. In fact, uncertainty in the weights could (and probably should) be incorporated into the overall estimates and their uncertainty.

Ambiguity in Classification of Studies Into Strata

It is clear that some studies will be difficult to classify into strata. Some may overlap stratum boundaries. In other cases parts of a study may fall into different strata. Such problems are common in meta-analyses and there is little reason to believe that they would be insurmountable in this context.

CONCLUSION

The model of synthesis proposed here defines questions more sharply in a fashion more relevant to policy concerns—what might happen if a policy were implemented in a relevant range of contexts. It is a more difficult approach, but one that is not impossible to carry out. The model will reveal gaps in evidence and make explicit precisely how assumptions have been substituted for empirical evidence to make inferences when some of the necessary empirical evidence was unavailable. This new model could produce estimates of treatment effects that are similar to those produced by more traditional meta-analytic methods. For example, if all studies gave the same estimate of treatment effect regardless of context or treatment type, the overall estimates from a simple meta-analysis and the more complex variety described here would coincide. Most likely they would not. In that case, the model proposed here provides more valid answers to questions of interest to policymakers.

REFERENCES

Becker, B.J. The generalizability of empirical research results. In: Benbow, D., and Lubinski, D., eds. *From Psychometrics to Giftedness: Papers in Honor of Julian C. Stanley*. Baltimore: Johns Hopkins University Press, in press.

Chalmers, T.C.; Levin, H.; Sacks, H.S.; Reitman, D.; Berrier, J.; and Nagalingam, R. Meta-analysis of clinical trials as a scientific discipline, I: Control of bias and comparison with large co-operative trials. *Stat Med* 6:315-315, 1987.

Cochran, W.G. *Sampling Techniques*. 3d ed. New York: Wiley, 1977.

Cooper, H.M. *The Integrative Research Review*. Beverly Hills, CA: Sage Publications, 1984.

Cooper, H.M., and Hedges, L.V., eds. *The Handbook of Research Synthesis*. New York: The Russell Sage Foundation, 1994.

Cordray, D.S., and Fischer, R.L. Practical aspects of evaluation synthesis and its variations. In: Wholey, J.S.; Harty, H.P.; and Newcomer, K.E., eds. *Handbook of Practical Program Evaluation*. San Francisco: Jossey-Bass, 1993.

Cronbach, L.J. *Designing Evaluations of Educational and Social Programs*. San Francisco: Jossey-Bass, 1982.

Eddy, D.M. Practice policies—What are they? *JAMA* 263:877-880, 1990*a*.

Eddy, D.M. Practice policies: Where do they come from? *JAMA* 263:1265-1275, 1990*b*.

Eddy, D.M. Practice policies—Guidelines for methods. *JAMA* 263:1839-1841, 1990*c*.

Feinstein, A. *Clinical Epidemiology*. Philadelphia: W.B. Saunders, 1985.

Hedges, L.V. Meta-analysis. *J Educ Stat* 17:279-296, 1992.

Hedges, L.V., and Olkin, I. *Statistical Methods for Meta-Analysis*. New York: Academic Press, 1985.

Hunter, S.E., and Schmidt, F.L. *Methods of Meta-analysis: Correcting Error and Bias in Research Findings*. Beverly Hills, CA: Sage Publications, 1990.

Kadane, J.B.; Dickey, J.M.; Winkler, R.L.; Smith, W.S.; and Peters, S.C. Interactive elicitation of opinion for a normal linear model. *J Am Stat Assoc* 75:845-854, 1980.

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

Little, R.J.A., and Rubin, D.B. *Statistical Analysis of Missing Data*. New York: Wiley, 1987.

National Institute of Medicine. *Clinical Practice Guidelines.* Washington, DC: National Academy Press, 1990.

National Research Council. *Combining Information: Statistical Issues and Research Opportunities*. Washington, DC: National Academy Press, 1992. [Reprinted as Draper, D.; Gaver, D; Goel, P.; Greenhouse, J.; Hedges, L.; Morris, C.; and Waternoux, C. *Combining Information: Statistical Issues and Research Opportunities*. Washington, DC: American Statistical Association, 1993.]

Rosenfeld, A. The particle data group: Its nature and operation. *Ann Rev Nuclear Sci* 555-599, 1975.

Rosenthal, R. *Meta-analytic Procedures for Social Research*. Beverly Hills, CA: Sage Publications, 1984.

Rubin, D.B. *Multiple Imputation for Missing Data in Sample Surveys.* New York: Wiley, 1987.

Rubin, D.B. A new perspective on meta-analysis. In: Wachter, K.W., and Straf, M.L., eds. *The Future of Meta-Analysis.* New York: The Russell Sage Foundation, 1990.

Rubin, D.B. Meta-analysis: Literature synthesis or effect-size surface estimation? *J Educ Stat* 17:363-374, 1992.

Silberman, G.; Droitcour, J.A.; and Scullin, E.W. *Cross Design Synthesis: A New Strategy for Medical Effectiveness Research.* Report No. GAO/PEMD-92-18. Washington, DC: U.S. General Accounting Office, 1992.

Tukey, J.W. Methodology and the statistician’s responsibility for BOTH accuracy AND relevance. *J Am Stat Assoc* 74:786-793, 1994.

Wang, M.C.; Haertel, G.D.; and Walberg, H.J. Toward a knowledge base for school learning. *Rev Educ Res* 63:249-299, 1993.

Winkler, R.L. The assessment of prior distributions in Bayesian analysis. *J Am Stat Assoc* 62:1105-1120, 1967.

Woolf, S.H. *AHCPR Interim Manual for Clinical Practice Development*. AHCPR Pub. No. 91-0018. Washington, DC: Agency for Health Care Policy, 1991.

Woolf, S.H. Practice guidelines, a new reality in medicine II: Methods of developing guidelines. *Arch Intern Med* 152:946-952, 1992.

AUTHOR

Larry V. Hedges, Ph.D.
Professor
The University of Chicago
5835 South Kimbark Avenue
Chicago, IL 60637

Click here to go to page 216