0196-206X/93/1401-0045$03.00/0
DEVELOPMENTAL AND BEHAVIORAL PEDIATRICS
Copyright © 1993 by Williams & Wilkins
Vol. 14, No. 1, February 1993
Printed in U.S.A.

Commentary

Children and Hospitalization: Putting the New Reviews in Methodological Context

The two preceding papers[1,2] can be viewed as manifestations of a movement that has slowly transformed how behavioral and medical scientists carry out integrative research reviews. Meta-analytic procedures, the term used to describe statistical techniques for synthesizing the results of studies,[3] have existed since the turn of the century.[4] However, only in the past 20 years has the swelling research base provided impetus for the widespread use of these techniques.[5] Today, it is not unusual to find topics, such as the two considered here, that have undergone dozens, if not hundreds, of empirical evaluations.

These large literatures place in bold relief the inadequacies of traditional reviewing procedures. Traditional narrative reviews often are characterized by haphazard literature searches, unrevealed criteria for the inclusion and exclusion of studies, and the use of a private cognitive algebra for "calculating" the combined findings of the series of acceptable studies.

The underlying premise of the new techniques for synthesizing research is that "locating and integrating separate research projects involves inferences as central to the validity of knowledge as the inferences involved in primary data interpretation."[6] Because of the amount of behavioral and medical research conducted today and because of the haphazard system for its dissemination, the validity of research syntheses cannot be taken for granted. A scientist performing an integrative research review makes decisions about problem formulation, study retrieval and evaluation, data analysis and interpretation that bear directly on the credibility of resulting conclusions. Therefore, if knowledge transmitted through research reviews is to be trustworthy, research synthesists must be required to meet the same standards of rigorous methodology required of primary researchers.

ELEMENTS OF A RESEARCH SYNTHESIS

A generic meta-analysis can contain up to four separate sets of statistics, in addition to descriptive information about the database of studies[7] including: (1) combinations of the *p* levels of independent inference tests, (2) a frequency analysis of positive, negative, and null results, (3) estimates of the average effect size with confidence intervals, and (4) homogeneity analyses examining study features that might influence study outcomes. The need for combined probability tests diminishes as the body of literature grows or if the synthesist provides confidence intervals around effect size estimates.

The growing interest in quantitative procedures for research integration served to focus attention on other issues concerning the systematicity and objectivity of empirical syntheses. To clarify and organize these issues, Cooper[6] conceptualized integrative research reviews as data-gathering exercises and applied to them the same conceptual scheme used to organize and evaluate primary research. Table 1 presents this conceptualization.

Similar to primary research, a research synthesis involves problem formulation, data collection, data evaluation, analysis and interpretation, and public presentation. During the problem formulation phase, the most important decisions concern how to define the constructs of interest, both conceptually and operationally. Another important decision concerns how much of the methodological and conceptual details of studies should be retrieved for later testing as possible moderators of treatment effects (or other relations).

Data collection requires choosing procedures to gather the relevant literature. Synthesists can differ in both the searching strategies they employ[5] and in how they interpret search outcomes.[8] The literature search is the activity that most distinguishes integrative research reviews from primary research. How a synthesist chooses to enter the scientific communication system to retrieve studies can have profound effects on the conclusions that eventually are drawn about a research topic.

Data evaluation requires making decisions about the value of individual studies. Studies are typically evaluated according to their research designs, samples, measurements, and age, among other criteria. Synthesists examining the same literature but using different evaluation criteria can easily end up with reviews that differ markedly in the studies they cover.

Data analysis involves deciding what procedures will be used to integrate the findings of the separate studies. When meta-analysis is used, these procedures and their underlying assumptions are made explicit. When narrative procedures are used to summarize findings of empirical research, the procedures and their assumptions are rarely revealed; even the synthesists themselves may be unable to articulate what the combining rules were.

Finally, public presentation concerns report writing and editing. Synthesists, such as primary researchers, must decide which aspects of their research process will help readers

TABLE 1. The Integrative Research Review Conceptualized as a Research Project

Stage Characteristics	Stage of Research				
	Problem Formulation	Data Collection	Data Evaluation	Analysis and Interpretation	Public Presentation
Research question asked	What evidence should be included in the review?	What procedures should be used to find relevant evidence?	What retrieved evidence should be included in the review?	What procedures should be used to make inferences about the literature as a whole?	What information should be included in the review report?
Primary function in review	Constructing definitions that distinguish relevant from irrelevant studies	Determining which sources of potentially relevant studies to examine	Applying criteria to separate "valid" from "invalid" studies	Synthesizing valid retrieved studies	Applying editorial criteria to separate important from unimportant information
Procedural differences that create variation in review conclusions	1. Differences in included operational definitions 2. Differences in operational detail	Differences in the research contained in sources of information	1. Differences in quality criteria 2. Differences in the influence of nonquality criteria	Differences in rules of inference	Differences in guidelines for editorial judgment
Sources of potential invalidity in review conclusions	1. Narrow concepts might make review conclusions less definitive and robust 2. Superficial operational detail might obscure interacting variables	1. Accessed studies might be qualitatively different from the target population of studies 2. People sampled in accessible studies might be different from target population of people	1. Nonquality factors might cause improper weighting of study information 2. Omissions in study reports might make conclusions unreliable	1. Rules for distinguishing patterns from noise might be inappropriate 2. Review-based evidence might be used to infer causality	1. Omission of review procedures might make conclusions irreproducible 2. Omission of review findings and study procedures might make conclusions obsolete

Source: Cooper HM: Scientific guidelines for conducting integrative research reviews. Rev Educ Res 52:291–302, 1982

evaluate the validity of their efforts and replicate them if desired.

PROBLEMS IN THE PRACTICE OF RESEARCH SYNTHESIS

Primary researchers encounter many problems that seem to repeatedly bedevil their work. These include failures in random assignment, errorful measurements, differences between target and accessible populations, to name a few. Similarly, research synthesists encounter problems endemic to their enterprise. The two meta-analyses on hospitalized children provide good examples of some of the more vexing issues that often face research synthesists. I will refer to the Thompson and Vernon paper[1] as the *response* meta-analysis and to the Vernon and Thompson paper[2] as the *treatment* meta-analysis.

Publication Bias

Every research synthesist confronts the problem of publication bias.[9] That is, not all tests of a particular hypothesis have an equal chance of being published. Therefore, different studies have different likelihoods of coming to the attention of the synthesist. This would not be a problem if publication criteria were random or related strictly to the methodological quality of the study. However, we know that decisions about whether to submit and publish studies also are related to the research results: studies that confirm previously reported findings and/or that reach statistical significance are more likely to find their way into print.[10]

Both the response and treatment syntheses contained some steps to overcome publication bias. These included attempts to locate unpublished research through personal communication with investigators who were known to be active in the field of research. Several other techniques for locating "fugitive literature" are available.

The treatment synthesis also reported an attempt to assess the possible impact of publication bias on the research base by comparing the mean effect sizes revealed by published and unpublished studies. No difference was found. As consumers of the synthesis we should be (somewhat) reassured.

The response meta-analysis contained no similar analysis of effect sizes. This is troubling because an examination of Table 1 reveals a noticeable lack of "smaller" effect sizes, hovering between .00 and +.40. These are exactly the values we would expect to be excluded from the accessible sample of studies if a bias favoring the retrieval of statistically significant studies existed.

Missing Information

Another critical problem faced by research synthesists concerns the fact that primary researchers often do not report important information about their studies.[11] In particular, the present meta-analysts bemoaned the lack of information

reported on characteristics of the patient samples.

In addition, both meta-analyses excluded a significant number of studies because their reports did not include enough data for the calculation of effect sizes. Some meta-analysts have attempted to retrieve the missing data by contacting the primary authors. This strategy often meets with limited success. Other meta-analysts have compared studies described by complete and incomplete reports on other characteristics (e.g., research design, treatment type) to see whether "incompleteness" is related to other study variations. Sometimes such an analysis can shed light on whether missing information on effect sizes may be systematically related to the effect size estimates themselves, for example, a preponderance of incomplete reports may deal with a treatment that complete reports suggest is ineffectual.

Parenthetically, the response meta-analysis includes an example of an ingenious strategy to retrieve information from research reports and put it to a use unintended by the original researchers. That is, the synthesists gathered data from only the untreated groups of treatment-vs-control studies and used it to assess the existence and magnitude of posthospitalization behavior problems. The original researchers may never have examined their data in this light, focusing only on the difference between treated and untreated patients.

Coder Reliability

Extracting data from research reports can also be a source of concern for synthesists. Just as some information is not contained in reports, other information can be missed or misrecorded. To guard against this, the treatment meta-analysis employed two people to code three treatment characteristics the synthesists thought might involve difficult distinctions. The authors were correct to assume that most other coded study characteristics (e.g., year of publication, sample size) usually are retrieved with a high degree of accuracy. However, for other characteristics not double coded here, especially effect size estimates, the assumed reliability of a single coder is harder to accept.[12] It is advisable (and becoming more frequent) for meta-analysts to have at least a portion of all variables double coded and to report both coder agreement rates as well as other indexes of reliability, such as Cohen's kappa.

Multiple Dependent Effect Sizes

Another problem for meta-analysts occurs when more than one effect size estimate can be gleaned from a single study. Such effect size estimates are not statistically independent and therefore violate an important assumption underlying quantitative analysis. Dependent effect sizes can happen because (1) a single study contains multiple samples, (2) multiple treatment conditions are compared with one another or with the same control, (3) the same outcome measure is administered more than once, and/or (4) multiple, related outcome measures are administered. The first three types of dependence were problems in these meta-analyses, the last was not. (It could be argued that dependence goes even further, that different studies conducted in the same laboratory are not strictly independent because they share investigators, settings, research assistants, etc.)

Both meta-analyses addressed the "dependence" issue in a number of ways: (1) if separate samples of subjects were used they were treated as though they came from separate studies (in the treatment meta-analysis); (2) if multiple treatments appeared, all but one comparison was discarded (with different discard criteria used in each synthesis); (3) if outcomes were measured at more than one time interval in the primary studies, the multiple estimates were averaged. In addition to these strategies, the statistical modeling of the actual degree of dependence has been suggested, but this procedure is quite difficult to apply.[13]

The two meta-analysts also disaggregated (or reintroduced) multiple effect sizes from a single study when the analysis so dictated, for example, when types of treatments were being compared or when the temporal effect of treatments was at issue. This is a generally accepted procedure, if used judiciously.

Correlated Moderators

Another problem faced by research synthesists involves correlation between the moderator variables tested for their influence on effect sizes. Quite often, synthesists will discover that characteristics of studies are confounded with one another, and this complicates their interpretation of which moderator might be the true causal agent (if any).[14] For example, in the treatment meta-analysis a confound was discovered between two significant predictors of effect size: mother-present interventions were most likely to occur with younger patients whereas preparation interventions were most likely with older children. In this case, the confound was not problematic because it was explainable developmentally and because the meta-analysts could convincingly argue that the causal relation lies in the interaction of the two variables.

A more troubling example is alluded to in the response meta-analysis. Here, the authors found a moderator effect associated with the year in which a study was published, with the newest studies producing a small, negative effect size (−.04). Should we conclude that children today are no longer troubled by hospitalization? Not necessarily. It was also found that the absolute questionnaire format led to smaller effect sizes than did the comparative format. Furthermore, inasmuch as the absolute format is newer, we might surmise that more recent studies are more likely to use this format. Thus, what appears to be a diminishing response to hospitalization in recent years actually may be attributable to a change in research methodology.

Interpreting Effect Sizes

The two meta-analyses contain good discussions of some of the problems associated with interpreting effect sizes. The authors point out that when the outcome measure is not easily interpretable (e.g., dollars spent, mortality rates, lengths of hospital stays) it is often difficult to translate a finding into meaningful terms.

A sometimes useful device is to transform the standardized mean difference into a measure of distribution overlap, dubbed U_3, by Cohen.[15] This measure tells the percentage of people in one sample (typically the untreated one) who were surpassed by the average person in the other (treated)

sample. For example, the treatment meta-analysis revealed a weighted standardized mean difference of +.44 (the weighted effect size is more precise and should be used in nearly all applications). This difference is associated with a U_3 equal to 67%, meaning that the average child (the child at the 50th percentile of the distribution) in the untreated groups was reported to have more behavior problems than did 67% of children in the treated groups.

Although overlap of distribution measures make effect sizes a bit more accessible, their interpretation is still ambiguous. Other questions need to be asked to understand what effect sizes mean, questions such as how the treatment effect under study compares with other treatments and how the methodology of the underlying experiments (sample restrictions, treatment strength and fidelity, sensitivity of measures) might have influenced the study outcomes.[16] These questions also need to be asked of more intuitively appealing effect size metrics as well.

Other interesting problems frequently confronted by research synthesists are more clearly tied to the particular topic domain under consideration. Four of these will be briefly described, again using the present meta-analyses as examples.

Theory-Derived Mediational Models

In both the response and treatment meta-analyses, it seems the synthesists used intuition and clinical observation to guide their choices of moderators of effect size (along, of course, with the practical constraint of which moderators were reported by primary researchers). The response meta-analysis makes no mention of existing theories addressing children's responses to hospitalization. The treatment meta-analysis mentions theory only in the context of why particular treatments were thought to be effective by primary researchers. Obviously, if research synthesists can ground their choice of moderators in current theory, the value to their work increases. Research syntheses can be powerful devices for the testing of theory and the comparison of competing theories.[17]

Mono-Method Bias

Both meta-analyses focused on a single outcome measure of hospitalization, the Posthospital Behavior Questionnaire. The validity of the questionnaire is defended in the discussion of the treatment meta-analysis. Still, all single methods of measurement contain forms of bias (e.g., distortions in recall, socially desirable responding). These biases cannot be proved irrelevant unless the treatment effect is demonstrated to hold across multiple measures of the same construct that do not share method variance.[18] Although it is often too expensive for primary researchers to collect multiple outcome measures, research synthesists can define their constructs broadly so as to encompass the multiplicity of operationalizations often found in a literature. They can then test for the generality of relations or treatment effects.

Categorization of Continuous Variables

All too often, researchers, both primary and secondary, convert continuous variables into categorical ones. In these two meta-analyses, this occurs for the variables age of patients, length of hospitalization, measurement lag, and year of report publication. Categorization can deplete the power of tests of effect size moderators.[19] It is best to leave these variables continuous (unless they exhibit decidedly nonnormal distributions) and to test directly for linear and curvilinear relations.

Multiple Degree of Freedom Tests

All specific hypotheses are related to single degree-of-freedom tests; multiple degree-of-freedom tests answer diffuse questions. Sometimes however, multiple degree-of-freedom tests are mistakenly used as evidence to support specific comparisons (in primary research as well as in meta-analysis).[20] For example, the meta-analysis of treatments found significant differences between three preparation strategies (dramatic, multicomponent, stress point) and three preparation times (before admission, after admission, at intervals). Because both significant findings were based on multiple degree of freedom tests ($df = 2$) the authors conclude that the variation in treatment effects was greater than chance. However, they *do not* identify significant differences between specific treatment variations. Thus, although these meta-analysts likely were correct to assert (given the observed magnitude of effect size differences) that stress point preparation was the most effective treatment and that treatments administered before admission were least effective, it is also the case that these specific assertions never were actually given a formal statistical test.

The list of problems outlined above may seem long and formidable, but the reader should bear two things in mind. First, the problems, in fact, do not arise from a scientist's decision to attempt a rigorous research synthesis or to employ meta-analysis. Instead, *the problems are intrinsic to research literatures, regardless of the rules for integration adopted by the synthesist.* It was not until large literatures pushed meta-analysis to the fore and until syntheses were examined in light of scientific standards that the problems were revealed. A potential synthesist should not believe that the decision to employ narrative techniques, as opposed to meta-analysis, is a way to overcome these problems. To the contrary, by ignoring them, traditional review procedures only heighten their threat to the validity of what we call knowledge.

Second, the potential meta-analyst and consumers of meta-analytic research should not be put off by the fact that only partial solutions to problems in research synthesis have been developed. As in primary research, the perfect research synthesis will always elude us. Still, we continue with our imperfect efforts and through the application of scientific standards knowledge claims come to be weighted differentially.

HARRIS COOPER, PH.D.
Department of Psychology
University of Missouri
Columbia, Missouri

REFERENCES

1. Thompson RH, Vernon TA: Research on children's behavior after hospitalization: A review and synthesis. J Dev Behav Pediatr 14:28–35, 1993
2. Vernon TA, Thompson RH: Research on the effects of experimental interventions on children's behavior after hospitalization: A review and synthesis. J Dev Behav Pediatr 14:36–44, 1993
3. Glass GV: Primary, secondary and meta-analysis of research. Educ Res 5:3–8, 1976
4. Olkin I: History and goals. In KW Wachter, ML Straf (eds.) The future of meta-analysis. New York, Russell Sage, 1992
5. Cooper HM: Literature searching strategies of integrative research reviewers. Knowledge 8:372–383, 1986
6. Cooper HM: Integrating research: A guide for literature reviews. Newbury Park, CA, Sage, 1989, p 12
7. Hedges LV, Olkin I: Statistical methods for meta-analysis. Orlando, FL, Academic Press
8. Cooper HM, Ribble RG: Influences on the accuracy of literature searches for research synthesis. Knowledge 10:179–201, 1989
9. Dickersin K: The existence of publication bias and risk factors for its occurrence. JAMA 263:1385–1389, 1990
10. Greenwald A: Consequences of prejudice against the null hypothesis. Psychol Bull 82:1–20, 1975
11. Orwin RG, Cordray DS: The effects of deficient reporting on meta-analysis: A conceptual framework. Psychol Bull 97:134–147, 1985
12. Stock WA, Okun MA, Haring MJ, et al: Rigor in data synthesis: A case study of reliability in meta-analysis. Educ Researcher 11:10–20, 1982
13. Raudenbush S, Becker BJ, Kalain H: Modeling multivariate effect sizes. Psychol Bull 103:111–120, 1988
14. Cooper H, Lemke KM: On the role of meta-analysis in personality and social psychology. Pers Soc Psychol Bull 17:245–251, 1991
15. Cohen J: Statistical power analysis for the behavioral sciences. Hillsdale, NJ, Erlbaum Associates, 1988
16. Cooper H: On the significance of effects and the effects of significance. J Pers Soc Psychol 41:1013–1018, 1981
17. Cook TD, Cooper H, Cordray DS, et al: Meta-analysis for explanation: A casebook. New York, Russell Sage, 1992
18. Cook TD, Campbell DT: Quasi-experimentation. Chicago, IL, Rand McNally, 1979
19. Cohen J, Cohen PL Applied multiple regression/correlation analysis for the behavioral sciences, 2nd edition. Hillsdale, NJ, Erlbaum Associates, 1983
20. Rosenthal R, Rosnow RL: Contrast analysis: Focused comparisons in the analysis of variance. New York, Cambridge University Press, 1985