DOCUMENT RESUME

ED 262 100	TM 850 588
AUTHOR	Bryant, Fred B.
TITLE	Improving the Quality of Research Synthesis in Program Evaluation.
PUB DATE	Oct 84
NOTE	22p.; Portions of this paper were presented at the Joint Meeting of the Evaluation Network and the Evaluation Research Society (San Francisco, CA, October 11-13, 1984).
PUB TYPE	Speeches/Conference Papers (150) -- Information Analyses (070) -- Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Educational Research; Effect Size; Evaluation Methods; Intervention; *Measurement Objectives; Meta Analysis; *Program Effectiveness; *Program Evaluation; *Research Methodology; Research Problems; Sampling; Statistical Dias; *Synthesis; *Validity
IDENTIFIERS	Evaluation Problems

ABSTRACT

Because research synthesis enables one to determine either the overall effectiveness of a particular treatment or the relative effectiveness of different types of treatments, it is becoming increasingly popular as a tool in program evaluation. Numerous methodological problems arise, however, when research synthesis is applied to studies conducted in field settings. The present paper categorizes and discusses these problems as being threats to either the (1) internal validity (whether one can draw conclusions about cause and effect), (2) statistical conclusion validity (whether one's inferential statistics are capable of detecting cause-and-effect relationships), (3) construct validity (whether one's treatments and outcome measures are valid operationalizations of the independent and dependent variables of interest), or (4) external validity (whether one can generalize results to particular populations, settings, or time periods) of research synthesis (see Cook and Campbell, 1979). Specific recommendations are made for minimizing these threats to validity, in order to improve the quality of research synthesis in program evaluation. (Author)

Improving the Quality of Research Synthesis in Program Evaluation*

Fred B. Bryant

Loyola University of Chicago

Address:
Department of Psychology
6525 North Sheridan Road
Chicago, IL 60626

* The author wishes to thank Thomas D. Cook, Linda S. Perloff, and Emil Posavac for helpful comments on an earlier draft of this manuscript. Portions of this paper were presented at the 1984 Joint Meeting of the Evaluation Research Society and the Evaluation Network in San Francisco, California.

Abstract

Because research synthesis enables one to determine either the overall effectiveness of a particular treatment or the relative effectiveness of different types of treatments, it is becoming increasingly popular as a tool in program evaluation. Numerous methodological problems arise, however, when research synthesis is applied to studies conducted in field settings. The present paper categorizes and discusses these problems as being threats to either the (1) internal validity (whether one can draw conclusions about cause and effect),(2) statistical conclusion validity (whether one's inferential statistics are capable of detecting cause-and-effect relationships),(3) construct validity (whether one's treatments and outcome measures are valid operationalizations of the independent ard dependent variables of interest), or (4) external validity (whether one can generalize results to particular populations, settings, or time periods) of research synthesis (see Cook & Campbell, 1979). Specific recommendations are made for minimizing these threats to validity, in order to improve the quality of research synthesis in program evaluation.

General Overview

This paper addresses strategies for improving the quality and utility of research synthesis in program evaluation. First, I will describe the advantages of research synthesis over other integrative techniques and will argue that these capabilities make it particularly useful for evaluating questions about program impact. I will suggest that one way to promote more excellent program evaluation is to improve the quality of research synthesis. Just as we can use validity criteria to improve the quality of primary research, I will argue that we may likewise improve the quality of research synthesis by controlling for threats to its validity. Finally, I will consider some threats to the validity of research synthesis and will suggest specific means of avoiding these pitfalls.

The Strengths of Research Synthesis

In this paper, I will use the term "research synthesis" to denote a set of integrative techniques for combining the results from independent empirical studies on a particular topic or issue. Other writers have used a variety of terms for research synthesis, including meta-analysis (Glass, 1976; Hunter, Schmidt, & Jackson, 1982), quantitative review (Cooper & Arkin, 1981), statistical review (Arkin, Cooper, & Kolditz, 1980), integrative review (Oliver & Spokane, 1983; Walberg & Haertel, 1980), empirical cumulation (Taveggia, 1974), data synthesis (Stock, Okun, Haring, Miller, Kinney, & Ceurvorst, 1982), and evaluation synthesis (Morra, 1983). Although the terminology varies, all of these integrative techniques emphasize a similar quantitative approach to reviewing primary research. This generally involves extracting from the original research reports the posttest means and standard deviations of treatment and control groups. These statistics are then combined to obtain a standardized "effect size," by subtracting the control group's mean from the treatment group's mean and dividing the result by some estimate of the population standard deviation. This effect-size statistic quantifies the magnitude and direction of treatment effects in a "common metric" of standard deviations, so that effect sizes can be pooled and compared across studies. By keeping track of contextual variables within primary studies, such as characteristics of the sample, the setting, the treatment, the outcome measures, and the research design, one can also search for variables that moderate treatment effects.

This quantitative approach has distinct advantages over traditional methods of literature review. For example, the traditional qualitative review is largely subjective and provides little or no statistical information about the strength of observed effects. Furthermore, other methods of quantitative review, such as a simple "vote count" that categorizes studies' outcomes as positive, negative, or zero effects, can produce misleading "no difference" conclusions, or Type II errors, because of low statistical power (Hedges & Olkin, 1980; Light & Smith, 1971; Light & Pillemer, 1984). Research synthesis allows a more systematic investigation of the mean and variance of effect sizes. Thus, the main strength of research synthesis is that it provides a quantitative index of treatment effects expressed in a metric that is comparable across studies.

Research Synthesis in Impact Evaluation

These capabilities make research synthesis a particularly useful tool for impact evaluation. Impact evaluation essentially provides information about a program's effectiveness. This may involve either (1) questions about a program's overall impact (e.g., Does the program work? Is it having its desired effect? Are there any unanticipated side-effects?) or (2) questions about a program's relative impact (e.g., What form of program is most effective and most cost-efficient? How should the program be implemented to maximize its effectiveness? For whom and in what settings does the program work best?).

In addressing questions about a program's overall impact, research synthesis enables one to "boil down" a set of primary studies into a single index of treatment effects. This facilitates more effective cost-benefit analysis by quantifying benefits for program recipients in a standard unit that can be meaningfully related to program expenditures. Synthesizing the literature on school desegregation and black achievement, for example, Wortman and Bryant (1985) found an overall average effect-size of +.20. This outcome represents a gain for edesegregated students (relative to segregated students) of roughly two months of growth in academic achievement on a standardized test. Because this expresses the magnitude of intellectual growth in a unit that is more meaningful than the number of points on an achievement test, the policy maker can better gauge the benefits of desegregation programs relative to their financial costs.

Although research synthesis is helpful in determining a program's overall impact, it is perhaps most useful in addressing questions about relative impact. Because effect sizes express each study's results in a common metric, one can use research synthesis to identify variables associated with stronger impact. Furthermore, the evaluator can use research synthesis to determine (a) the relative impact of a particular type of program on different outcome measures or (b) the relative impact of different types of programs on a particular outcome measure. As an illustration of how to determine a program's relative impact on different outcome measures, Messick and Jungeblut (1981) synthesized research on the effects of coaching for the Scholastic Aptitude Test and found that increases in verbal scores required more coaching time than equivalent increases in math scores. As an illustration of how to determine the relative impact of different programs, Shadish (1982) synthesized research on preventive child health care and found that specific interventions for specific problems were more effective than broad-scale interventions. Again, research synthesis improves cost-benefit analyses in these cases by making comparisons of relative cost efficiency more meaningful.

Improving the Validity of Research Synthesis

Given that research synthesis is a useful tool for evaluating program impact, then one way of promoting more excellent impact evaluation is to improve research synthesis. Accordingly, the remainder of this paper addresses strategies for improving the validity of research synthesis. Cook and Campbell (1979) have distinguished among four types of validity in primary research--internal, statistical conclusion, construct, and external validity. Just as Cook and Campbell (1979) have urged researchers to use these validity criteria to improve primary research, I am proposing that we also use these same criteria to improve research synthesis.

Internal validity. Internal validity concerns the degree of confidence that one has in drawing conclusions about cause and effect (Campbell & Stanley, 1966; Cook & Campbell, 1979). As with all forms of research, the conclusions drawn from research synthesis are only as good as the evidence on which they are based. If all the studies included in the synthesis are methodologically flawed, then the conclusions drawn from the synthesis will lack internal validity. For this reason, it is important to keep track of threats to the internal validity of each of the primary studies that are included in the research synthesis. By coding studies for specific threats to their internal validity, such as selection, maturation, history, and instrumentation, one may systematically examine how these threats influence effect size. For example, Wortman and Bryant's (1985) synthesis of research on school desegregation revealed that studies which were judged a priori as having problems with selection bias had significantly greater effect sizes than those without selection problems. If all the studies included in the synthesis suffer the same methodological flaw, however, it may be impossible to determine how this particular threat influences effect size (Cook & Leviton, 1980; Jackson, 1980). When there is little or no variance in methodological quality, one has no way of examining quality as an independent variable. One needs a sufficient number of high quality studies to use as a baseline against which to compare studies of poorer quality. Without this high quality baseline, the internal validity of research synthesis is suspect. Therefore, when the range of methodological quality in the primary studies is restricted to the low end of the continuum, one may increase the internal validity of research synthesis by using only those studies of relatively higher quality (Bryant & Wortman, 1984).

This represents a type of purposive sampling plan (Cook & Campbell, 1979; Sudman, 1976), whereby one chooses which studies to include on the basis of their methodological rigor rather than trying to insure representativeness. One's choice of sampling strategy in research synthesis may thus sometimes depend on whether it is more important to draw unequivocal conclusions that have limited generalizability or equivocal conclusions that are widely generalizable. I will return to this notion of purposive sampling when I discuss external validity in research synthesis.

Statistical conclusion validity. Whereas internal validity concerns the question of whether some aspect of the treatment produced observed outcomes, statistical conclusion validity concerns the question of whether one's inferential statistics are capable of detecting a cause-and-effect relationship (Cook & Campbell, 1979). Recent work on the statistical theory underlying estimates of effect size suggests several strategies for maximizing the validity of statistical conclusions in research synthesis.

One way to improve statistical conclusion validity in research synthesis is to use estimators of effect size that have less statistical bias. For example, one can obtain a less biased estimator of effect size by using the pooled within-groups standard deviation as the unit of standardization in the denominator, rather than using the control group's standard deviation, as is typically done (Hedges, 1981, 1982; Hunter et al., 1982). Furthermore, if the studies included have different sample sizes, then one can obtain a more precise estimate of overall treatment effects by weighting each study's effect size according to the size of its sample (see Hedges, 1982, and Hunter et al., 1982, for formulas of weighted estimators). Other investigators have developed procedures to correct estimates of effect size for unreliability in both the outcome measures of the primary studies (Hunter et al., 1982) as well as the coding of variables in research synthesis (Orwin & Cordray, 1985).

Special problems with statistical conclusion validity arise when the research literature being synthesized is quasi-experimental (Bryant & Wortman, 1984). If treatment and control groups have not been randomly assigned, then one cannot assume that these groups are equivalent at the pretest. In these cases of selection bias, it may be unreasonable to use the traditional estimate of effect size (Cohen, 1977; Glass, 1976; Glass, McGaw, & Smith, 1981), which assumes pretest equivalence. However, if pretest measures are available, then one may calculate an effect size for the pretest and use it to adjust the posttest effect-size for initial between-groups differences. Wortman and Bryant (1985) have shown that this pretest-adjusted effect size is a more accurate estimate of treatment effects in quasi-experiments than is the traditional posttest effect-size.

Another way to improve statistical conclusion validity in research synthesis is to improve our procedures for identifying relationships between independent and dependent variables. Before pooling effect sizes to calculate an overall effect-size, for example, one can statistically test the homogeneity of studies' outcomes (see Hedges, 1982; Hunter et al., 1982). If one rejects the null hypothesis that sampling error alone accounts for observed variation in effect sizes, then it is unlikely that all studies come from the same underlying population, and one should distrust a single overall effect-size. If one fails to reject this null hypothesis, on the other hand, then sampling error alone may account for observed variation in effect sizes, and it may be unreasonable to search for variables that moderate treatment effects. Furthermore, when doing multiple statistical comparisons in research synthesis, one should either correct the alpha level for the per comparison error rate (Ryan, 1959) or use multivariate tests that do so (Harris, 1975), to avoid the so-called "fishing problem" (Cook & Campbell, 1979). Besides avoiding Type I errors of capitalizing on chance, one may avoid Type II errors that result from low statistical power (Cohen, 1977) by abstaining from research synthesis all together when too few studies exist on the particular topic (Cook & Leviton, 1980).

An additional problem involves the unit of analysis in research synthesis. Multiple outcomes from a single study (e.g., multiple treatment or control groups, multiple dependent measures, or measures taken at multiple points in time) must be treated as being nonindependent. Thus one should either average multiple outcomes within studies to compute an overall effect-size or compare multiple outcomes within studies to search for moderator variables (Landman & Dawes, 1982).

Construct validity. Construct validity concerns the degree to which the particular treatments and outcome measures are valid operationalizations of the constructs supposedly underlying the independent and dependent variables (Cook & Campbell, 1979). To improve construct validity in research synthesis, one should at the outset explicitly specify the range of treatments, comparison groups, and outcome measures that will be considered relevant (Bryant & Wortman, 1984; Cooper, 1982). Furthermore, in order to avoid lumping together "apples and oranges" (Gallo, 1978; Wortman, 1982), one should consider different forms of treatment separately (Light & Pillemer, 1984) and should divide studies into clusters according to the measurement instruments used (Feldman, 1971). Previous theory and research may be useful in deciding how to stratify treatments and outcome measures. For example, a recent synthesis of research on age differences in subjective well-being (Stock, Okun, Haring, & Witter, 1983) combined into one global index five related types of outcome measures--life satisfaction, happiness, morale, quality of life, and well-being. Recent theory and research on subjective mental health (Bryant & Veroff, 1984; Veit & Ware, 1983), however, suggest that these are clearly distinct constructs that should be considered separately.

Perhaps the most serious threat to the construct validity of research synthesis is the difficulty of assessing the strength or "dosage" (Quay, 1977; Sechrest, West, Phillips, Redner, & Yeaton, 1979) of the treatment. Often one only finds significant main effects or interactions when the appropriate levels of independent variables have been implemented (Cooper & Arkin, 1981). For example, programs designed to promote preventive health behaviors by arousing fear may only work when they elicit moderate levels of fear and may be relatively ineffective when they involve either low or high levels of fear (Janis & Feshbach, 1953). This suggests that research synthesis should incorporate qualitative information about the strength of the treatment as implemented in the primary studies (Light & Pillemer, 1984; Morra, 1983). This would enable us to distinguish studies involving stronger treatments from studies involving milder treatments and to specify the level at which a particular treatment works best.

A related strategy for improving construct validity in research synthesis is to decompose the "treatment package" (Quay, 1977) into its composite constructs. This involves identifying different conceptual components of a particular treatment program and keeping track of the levels at which each of these components has been implemented across studies. This approach enables one to pinpoint the specific ingredients that maximize a program's impact. Synthesizing research on hospital patient education programs, for example, Devine and Cook (1983) identified three common components of treatment interventions: (1) providing patients with information about medical procedures, (2) attempting to increase patients' feelings of control, and (3) attempting to reduce patients' levels of anxiety. The programs most effective in reducing length of stay were those that incorporated all three of these components; programs that involved only one or two of these components were less effective. This illustrates how decomposing a treatment into its composite constructs can help us specify precisely which of these constructs are responsible for a program's effectiveness.

External validity. External validity concerns the degree of confidence that one has in generalizing results to different populations, settings, or time periods (Campbell & Stanley, 1966; Cook & Campbell, 1979). External validity is especially problematic in research synthesis because there is no single definitive list of all existing studies on a given topic. This typically precludes an exhaustive sampling of all existing studies and prevents one from determining the representativeness of one's final sample of studies (Feldman, 1971).

In discussing internal validity, I suggested that one may decide to sample only studies of relatively higher methodological quality when the range of design quality is restricted to the low end of the continuum and one wishes to place more weight on internal validity than on external validity. I will now propose two other types of purposive sampling strategies that can be used in research synthesis.

The first purposive sampling strategy is to sample for modal instances (Cook & Campbell, 1979; St. Pierre & Cook, 1984). This involves limiting the sample of studies to those using the most widely representative populations, settings, or forms of treatment implementation. This sampling plan provides program developers with information that is generalizable solely to the cases that are most typical. The task here is to define the variables across which one wishes to generalize and then to select instances at the mode of each of these variables (St. Pierre & Cook, 1984). Imagine, for example, that one has been commissioned to synthesize research on school desegregation for the legislature of a particular New England State. In this case, one might decide to sample only studies of two-way cross-district busing programs in large, urban settings, if this was the modal instance for the particular state.

Another type of purposive sampling strategy is to sample on implementation (Cook & Campbell, 1979; St. Pierre & Cook, 1984). This involves selecting only studies in which the particular program is developed and mature enough to be well-implemented or to be transferable from one location to another. For example, in synthesizing research on alternative health care programs for state-subsidized nursing homes, one might decide to include only studies of programs that one feels have been developed clearly and fully enough to be transferred to the particular sites one has in mind. Alternatively, one may chose to sample only studies in which the program has been either particularly well-implemented, moderately well-implemented, or poorly implemented, to determine how well it works at different levels of implementation. This represents one way of incorporating crucial qualitative information about the integrity or fidelity of the treatment (Gottfredson, 1984; Quay, 1977; Sechrest et al., 1979).

As is the case with primary research, the ultimate test of the external validity of research synthesis is independent replication. Thus, in the long run, the best way to enhance external validity in research synthesis may be to improve the ability of others to duplicate (1) our procedures for selecting relevant studies and relevant comparisons within studies and (2) our criteria for coding quantitative and qualitative information from primary studies. One strategy for improving replicability is to make explicit the many subjective judgments that one must necessarily make in research synthesis (Cooper, 1982; Wortman & Bryant, 1985). Without knowing the specific criteria by which a particular researcher has resolved these inevitable uncertainties, independent replication remains impossible.

Another way to enhance the external validity of research synthesis is to establish formal archives of published and unpublished reports on selected topics. In fact, this is currently being done in the field of education by the Educational Resources Information Center (ERIC). This helps to promote independent reanalyses by providing others with a comparable sample of studies for research synthesis.

Perhaps the most efficient method of improving external validity, however, would involve making public the actual raw data from research syntheses. Just as archiving primary research data facilitates more effective secondary analysis (Bowering, 1984; Bryant & Wortman, 1978), so may archiving the data from research synthesis promote more valid reanalyses of the same data base (Bryant & Wortman, 1984).

In conclusion, I have argued that we can improve the quality of research synthesis by controlling for threats to its internal, statistical conclusion, construct, and external validity. I have considered major threats to each type of validity and have suggested specific strategies for avoiding these pitfalls. There are, however, other potential threats to validity in research synthesis about which we know very little. For example, small sample sizes reduce statistical power and undermine statistical conclusion validity (Cook & Campbell, 1979); however, no formal rules have been established for deciding on the minimum number of studies required for research synthesis. Future work should explore whether power curves (Cohen, 1977; Feldt & Mahmoud, 1958) for determining the number of subjects to include in primary research can be used to determine the number of studies to include in research synthesis (Bryant & Wortman, 1984). In addition, we know very little about how artifacts such as sampling error influence estimates of effect size. Monte Carlo "simulation" studies are clearly needed to test the susceptibility of statistical procedures in research synthesis to Type I and Type II errors. Only by carefully considering sources of error and bias in research synthesis can we maximize its ability to provide us with valid conclusions.

References

Arkin, R., Cooper, H., & Kolditz, T. (1980). A statistical review of the literature concerning the self-serving bias in interpersonal influence situations. *Journal of Personality*, *48*, 435-448.

Bowering, D.J. (1984). Impact analysis using an integrated data base: A case study. In D.J. Bowering (Ed.), *New directions for program evaluation*, vol. 22. San Francisco: Jossey-Bass.

Bryant, F.B. & Veroff, J. (1984). Dimensions of subjective mental health in American men and women. *Journal of Health and Social Behavior*, *25*, 116-135.

Bryant, F.B. & Wortman, P.M. (1978). Secondary analysis: The case for data archives. *American Psychologist*, *33*, 381-387.

Bryant, F.B. & Wortman, P.M. (1984). Methodological issues in the meta-analysis of quasi-experiments. In W.H.Yeaton & P.M. Wortman (Eds.), *New directions for program evaluation*. San Francisco: Jossey-Bass.

Campbell, D.T., & Stanley, J.C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Cook, T.D., & Leviton, L. (1980). Reviewing the literature: A comparison of traditional methods with meta-analysis. *Journal of Personality*,*48*, 449-471.

Cooper, H.M. (1982). Scientific guidelines for conducting integrative research reviews. *Review of Educational Research*, *52*, 291-302

Cooper, H., & Arkin, R.M. (1981). On quantitative reviewing. *Journal of Personality*, *49*, 225-230.

Devine, E.C., & Cook, T.D. (1983). A meta-analytic analysis of the effects of psycho-educational interventions on length of post-surgical hospital stay. *Nursing Research*, *32*, 267-274.

Feldman, K.A. (1971). Using the work of others: Some observations on reviewing and integrating. *Sociology of Education*, *44*, 85-102.

Gallo, P.S. (1978). Meta-analysis: A mixed meta-phor. *American Psychologist*, *33*, 517.

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

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

Gottfredson, G.D. (1984). A theory-ridden approach to program evaluation: A method for stimulating researcher-implementer collaboration. *American Psychologist*, *39*, 1101-1112.

Harris, R.J. (1975). *A primer of multivariate statistics.* New York: Academic Press.

Hedges, L.V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin*, *92*, 490-499.

Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107-128

Hedges, L.V., & Olkin, I. (1980). Vote counting methods in research synthesis. *Psychological Bulletin*, *88*, 359-369.

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

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

Janis, I.L., & Feshbach, S. (1953). Effects of fear-arousing communications. *Journal of Abnormal and Social Psychology*, *48*, 78-92.

Landman, J.T., & Dawes, R.M. (1982). Psychotherapy outcome: Smith and Glass' conclusions stand up under scrutiny. *American Psychologist*, *37*, 504-515.

Light, R.J., & Smith, P.V. (1971). Accumulating evidence: Procedures for resolving contradictions among different research studies. *Harvard*

Educational Review, 41, 429-471.

Light, R.J. & Pillemer, D.B. (1984). Summing up: The science of reviewing research. Cambridge, MA: Harvard University Press.

Messick, S., & Jungeblut, A. (1981). Time and method coaching for the SAT. Psychological Bulletin, 89, 191-216.

Morra, L. (1983). The evaluation synthesis. Washington, D.C.: United States General Accounting Office.

Oliver, L.W., & Spokane, A.R. (1983). Research integration: Approaches, problems, and recommendations for research reporting. Journal of Counseling Psychology, 30, 252-257.

Orwin, R., & Cordray, D.S. (1985)). The effects of deficient reporting on meta-analysis: A conceptual framework and reanalysis. Psychological Bulletin,97, 134-147.

Quay, H.C. (1977). The three faces of evaluation: What can be expected to work. Criminal Justice and Behavior, 4, 341-354.

Ryan, T.A. (1959). Multiple comparisons in psychological research. Psychological Bulletin, 56, 26-47.

Sechrest, L., West, S.G., Phillips, M.A., Redner, R., & Yeaton, W. (1979). Some neglected problems in evaluation research: Strength and integrity of treatments. In L. Sechrest, S.G. West, M.A. Phillips, R. Redner, & W. Yeaton (Eds.), Evaluation studies review annual, vol.4. Beverly Hills, CA: Sage.

Shadish, W.R., Jr. (1982). A review and critique of controlled studies of the effectiveness of preventive child health care. Health Policy Quarterly, 2, 24-52.

St. Pierre, R.G., & Cook, T.D. (1984). Sampling strategy in the design of program evaluations. In R.F. Connor (Ed.,) Evaluation studies review annual, vol.9. Beverly Hills, CA: Sage.

Stock, W.A., Okum, M.A., Haring , M.J., Miller, W. Kinney, C., & Cuervorst, R.W. (1982). Rigor in data synthesis: A case study of reliability in meta-analysis. *Educational Researcher*, *11*, 10-20.

Stock, W.A., Okun, M.A., Haring, M.J. & Witter R.A. (1983). Age differences in subjective well-being: A meta-analysis. In R.J. Light (Ed.), *Evaluation studies review annual*, vol.8. Beverly Hills, CA: Sage.

Sudman, S. (1976). *Applied sampling*. New York: Academic Press.

Taveggia, T. (1974). Resolving research controversy through empirical cumulation. *Sociological Methods and Research*, *2*, 395-407.

Veit, C.T. & Ware, J.E., Jr. (1983). The structure of psychological distress and well-being in general populations. *Journal of Consulting and Clinical Psychology*, *51*, 730-742.

Walberg, H.J., & Haertel, E.H. (1980). Research integration: The state of the art. *Evaluation in Education: An International Review Series*, *4*, 1-142.

Wortman, P.M. (1983). Evaluation research: A methodological perspective. *Annual Review of Psychology*, *34*, 223-250.

Wortman, P.M., & Bryant, F.B. (1985). School desegregation and black achievement: An integrative review. *Sociological Methods and Research*, *13*, 284-324.