DOCUMENT RESUME

ED 216 032 TM 820 188

AUTHOR Cooper, Harris
TITLE Scientific Guidelines for Conducting Integrative Literature Reviews.
PUB DATE Mar 82
NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Behavioral Science Research; Data Analysis; Data Collection; Evaluation; *Literature Reviews; *Models; Research Reports; *Validity
IDENTIFIERS *Integrative Processes

ABSTRACT

Inferences made in integrative research reviews are as important to the validity of behavioral science knowledge as are those in primary research. The research review is conceptualized as a scientific inquiry involving five stages paralleling those of primary research. Problem formulation is the stage when variables are defined conceptually and operationally. In the data collection stage, an inquirer must decide on the population of elements that will be the referent for the inquiry. Critical judgments about the quality of data points occur during data evaluation. Data collected by the researcher are synthesized into a unified statement during analysis and interpretation. Presentation of the review in a public document is the final stage. One study asked reviewers to integrate literatures that vary in findings and operational homogeneity of their studies. Some were requested to make formal judgments of research quality. The other study manipulated literature size, findings, and the reviewer's analytic interpretation strategy. The dependent variables in both studies were reviewer perceptions about the tested hypotheses and recommendations for future research. (DWH)

Scientific Guidelines for Conducting Integrative Literature Reviews

Harris Cooper
Center for Research in Social Behavior
University of Missouri-Columbia
111 East Stewart Road
Columbia, MO 65211

The inferences made in integrative research reviews are as central to the validity of behavioral science knowledge as those inferences made in primary research. Therefore, research reviewers must pay the same attention to rigorous methodology that is required of primary researchers. This paper is based on a conceptualization of the research review as a scientific inquiry involving five stages which parallel those of primary research: (a) problem formulation; (b) data collection; (c) evaluation of data points; (d) data analysis and interpretation, and; (e) public presentation.

The results of two studies will be reported which experimentally examine different facets of literature reviewing. In Study I, reviewers are asked to integrate literatures that vary in the findings and operational homogeneity of their constituent studies. Some reviewers are also asked to make formal judgments of research quality. These manipulations (along with several individual differences among reviewers) are tested as antecedents to the reviewer's (a) decisions about tested hypotheses, and (b) recommendations for future research. Study II manipulates the literature size and findings and the reviewers analytic interpretation strategy (statistical versus traditional). The dependent variables are again reviewer perceptions about the hypothesis and needed future research.

The studies examine aspects of research reviewing that have never before received systematic study. There results should have considerable impact on future discussions about how to carry out this critical scientific activity.

Scientific Guidelines for Conducting Integrative Literature Reviews

Harris M. Cooper
Center for Research in Social Behavior
111 East Stewart Road
University of Missouri-Columbia
Columbia, MO 65211

A paper presented in M. Steinkamp (Chair), Meta-analysis and the synthesis of researc' findings. Symposium presented at the meeting of the American Educational Research Association, New York, 1982.

Running Head: Scientific idelines

Scientific Guidelines for Conducting Integrative Literature Reviews

The use of quantitative procedures in literature reviews has increased largely in response to increases in the number of studies relevant to particular research hypotheses. Integrative reviewing, however, contains many decision points, in addition to how to synthesize studies, and all of these have been affected by the expanding research evidence. Critics of meta-analysis have pointed to problems in other phases of reviewing caused by large literatures and have accused the quantitative procedures of either creating or inadequately addressing these problems. In fact, neither is the case. The synthesis of studies is only one of several independent activities involved in literature reviewing. I would like to present a model of the literature review that (1) conceptualizes research integration as a scientific process and (2) suggests systematic guidelines for evaluating the validity of review outcomes. After describing the model, examples of reviews that were attempts to employ the guidelines will be presented.

The Stages of Integrative Research Review

Figure 1 describes the integrative review process (Cooper, in press). Five stages are identified by their particular research function. Sources of variance and potential threats to the validity of review outcomes which are associated with each stage are listed.

Place Figure 1 about here

Problem formulation stage. The first stage in the integrative review is the problem formulation stage. During problem formulation, the variables involved in the inquiry are defined in two different ways, conceptually and operationally.

The first source of variance in reviews enters during concept identification. Two reviewers using an identical label for an abstract concept can employ very different operational definitions or levels of abstraction. Each definition may contain some operations excluded by the other, or one reviewer's definition may completely contain the other.

Multiple operations for constructs also affect review outcomes by making it possible for reviewers to vary in the attention they pay to methodological distinctions in the literature. This variation is attributable to differences in the way operations are treated after the relevant literature has been retrieved. Two reviewers employing identical conceptual definitions and reviewing the same set of studies can still reach decidedly different conclusions. If one reviewer retrieved more method information and recognized a method-dependent relation that another reviewer did not test, the two conclusions could be orthogonal to one another.

Each source of variance introduces a potential threat to the validity of a review's conclusions. First, reviewers who focus on only a few empirical realizations leave open rival interpretations for the findings. Also, very narrow conceptualizations provide little information about how many different contexts a finding applies to. Therefore, reviewers who employ broad conceptual definitions can potentially produce more valid conclusions than reviewers using narrow definitions.

The word potentially was used because of the second threat to validity associated with problem definition. As Presby (1978) notes, "... differences (in studies) are cancelled in the use of very broad categories, which leads to the erroneous conclusion that research results indicate negligible differences in outcomes..." (p. 514). We can assume, therefore, that reviewers who examine more operational details within their broad constructs will probably produce more externally valid conclusions. These reviewers present more information about contextual variations that do and do not influence the review outcome.

Data collection stage. The major decision an inquirer makes during the data collection stage involves the population of elements that will be the referent for the inquiry.

Identifying populations for research reviews is complicated by the fact that reviews involve two targets. First, the inquirer wants the review findings to pertain to all previous research on the problem. Reviewers can exert some control over whether this goal is achieved through their choice of information sources. In addition, the reviewer hopes that the included studies will allow generalizations to the individuals that interest the topic area. The reviewer's influence is constrained at this point by the types of individuals who were sampled by primary researchers.

Discrepancies in review conclusions are created by differences in the channels reviewers use to retrieve information, such as invisible colleges, citation indexes and abstracting services. The studies available through different sources often are different from one another and Smith (1980) has demonstrated this empirically. It is likely

that two reviewers who use different techniques to locate studies will end up with different evidence and will potentially reach different conclusions.

The first threat to data gathering validity, then, is that the review may not include, and probably will not include, all studies pertinent to the topic of interest. A reviewer who has utilized the broadest sources of information is most likely to retrieve a set of results which resembles the entire population of previous research.

The second threat to validity occurring during data gathering is that the individuals in the retrieved studies may not represent all individuals in the target population. The reviewer cannot be faulted for the existence of this threat if retrieval procedures were exhaustive. However, reviewers who qualify conclusions with information about the kinds of people missing or overrepresented in studies probably run less risk of overgeneralization.

Data evaluation stage. After data is collected, the inquirer makes critical judgments about the quality of individual data points. Each data point is examined in light of surrounding evidence to determine whether it is contaminated by factors irrelevant to the problem under consideration.

The first source of variance introduced during data examination is created by divergence in reviewers' criteria for evaluating the quality of research. For instance, Gottfredson (1978) studied editors and authors in nine APA journals and suggested that interjudge agreement on quality was "relatively modest" (p. 928).

Another source of variance in review conclusions is the degree to which factors other than research quality affect evaluative decisions. To demonstrate, Mahoney (1977) found that the methods, discussion and contribution of manuscripts were evaluated more favorably if the study confirmed the reviewer's predisposition about the result.

The use of any evaluative criteria other than methodological quality ought to be considered a threat to the internal validity of a research review. As Mahoney states, "To the extent that researchers display (confirmatory) bias, our adequate understanding of the processes and parameters of human adaptation may be seriously jeopardized" (p. 162).

A second threat to validity during evaluation is wholly beyond the control of the reviewer. This threat involves incomplete reporting by primary researchers. If a reviewer must estimate or omit what happened in these studies, wider confidence intervals must be placed around review conclusions.

Analysis and interpretation stage. During analysis and interpretation, the separate data points collected by the inquirer are synthesized into a unified statement about the research problem. Interpretation demands that the inquirer distinguish systematic data patterns from "noise" or chance fluctuation. To carry out this function, the inquirer must apply some rules of inference.

Review conclusions can differ because reviewers employ different analytic interpretation techniques. A systematic relation which cannot be distinguished from noise under one set of rules may be differentiated under another set.

The first threat to validity accompanying the analysis and interpretation stage involves the rule of inference that a reviewer employs. In non-quantitative reviews, it is difficult to gauge the appropriateness of inference rules because they are not very often made explicit. For quantitative reviews, the suppositions of statistical tests are generally known and some statistical biases in reviews can be removed. Regardless of the strategy used for analysis and interpretation, the possibility always exists that the reviewer has used an invalid rule for inferring a characteristic of the target population.

The second threat to validity is the misinterpretation of review-based evidence as supporting statements about causality. In order to explain method-generated variance in study outcomes, reviewers will try to associate the differences in results with differences in study procedures. While the reviewer may be tempted to do so, he or she cannot rule out the possibility that the review-generated relation is spurious. Many other variables are confounded with the original experimenters' choice of a study procedure. Spurious relations are possible because the reviewer did not randomly assign procedures to experiments.

Public presentation stage. Finally, the production of a public document describing the review is a task with profound implications for the accumulation of knowledge.

Two threats to validity accompany report writing. First, the omission of details about how the review was conducted reduces the replicability of the review conclusion. Without sufficient detail, the reader is unable to ascertain whether a personal search of the literature would lead to a similar conclusion.

The second threat involves the omission of evidence that other inquirers find important. Matheson et al. (1978) observe that "as research on a specific behavior progresses, more details concerning the experimental conditions are found to be relevant" (p. 265). A review will quickly become obsolete if it does not address the variables and relations which are (or will be) important to an area.

Examples of Reviews

The supposition underlying this model of literature reviewing is that it is a data-gathering exercise which needs to be evaluated against scientific criteria. As with primary research, reviewers must take precautions to avoid bias in conducting their study. Equally important, the reviewer must produce a report which allows readers to assess the review's validity and to conduct direct replications, if they so desire. Also similar to primary research, the perfect literature review does not exist. Many reviewers have, however, applied the extra time, effort and expense needed to produce reviews with considerably greater validity and replicability than has traditionally been acceptable. My colleagues on today's panel are among these reviewers, as are many others who predate the "review-as-research" notion (see Glass, McGaw & Smith, 1981; Rosenthal, 1980). In the time remaining, I would like to briefly describe the efforts of three of my students who conducted reviews in very different areas using different techniques. What the reviews have in common is that they all attempted to apply the guidelines described earlier.

The first review was on the relation between locus of control (or a person's belief about whether or not they control the things

that happen to them) and academic achievement. This review was conducted by Maureen Findley, a graduate student in social psychology (Findley & Cooper, in press). Five previous reviews of the locus of control-achievement relation concluded that a positive association existed between the variables but the reviews differed in their confidence in this conclusion. The reviews also varied in their target populations, with some focusing on children, some on adults and some on all age groups. Finally, the reviews differed in the mediators they suggested might affect the size or existence of the relation. Of the five reviews, the most exhaustive contained 36 empirical studies. Maureen's goal was to comprehensively search the literature and examine all target populations and suggested mediators in a single review.

Three data bases, Psychological Abstracts, ERIC, and Dissertation Abstracts International, were searched by computer. The index words "achievement" or "performance" were crossed with "locus of control" or "internal-external." Eight hundred and two studies were located which either mentioned these terms in their title or abstract or were so classified by a person who read the entire document. The titles and abstracts of these studies were provided by the computer and these were used to reduce the number of potentially relevant reports. Ultimately, 208 studies were examined in their entirety and 98 relevant studies were found, nearly three times as many as the next most exhaustive review. The 98 studies contained 275 tests of the hypothesis.

Across all studies, the average correlation between locus of control and achievement was $r = +.18$. This combined result would require over 3,000 unretrieved, null-summing studies to be reversed at

the $p < .05$ level of significance (see Cooper, 1979). Table 1 provides a breakdown of these results according to the six mediating variables most frequently suggested by previous reviewers. The analyses of differences in correlations across studies revealed that male samples produced stronger relations than female samples and that junior high school students produced stronger relations than either elementary school or college students. In addition, measures of locus of control specific to academic outcomes tended to show stronger associations with achievement than more general locus of control measures and stronger relations were found in studies employing standardized as opposed to informal assessments of achievement.

Place Table 1 about here

Maureen's discussion of the results was able to evaluate the magnitude of the locus of control relation and pay particular attention to conflicts in the results of the previous reviews.

Ken Ottenbacher, a graduate student in special education, conducted a review of studies testing the effectiveness of drug treatments of hyperactivity in children. Ken was able to locate 61 studies that met very stringent criteria. All 61 studies employed two-group comparisons between a drug condition, a no treatment control condition or a placebo condition. In addition, all studies used random assignment of children to conditions and a double blind procedure in administering the treatment and recording the dependent variable.

Table 2 presents a stem-and-leaf display of the 408 separate d-indexes uncovered by the literature search. Before synthesis, Hedges'

correction factor was employed. Most interesting in this review was the comparison of the effects of placebo and drug treatments. The mean d-index for comparisons of drug treatments versus no treatment control groups was +1.21. Placebo groups versus no treatment controls produced a mean effect of d = +.19 while drug versus placebo comparisons revealed a mean d-index of +.69. This analysis led to a conclusion that about 3/4 of the drug effect was probably due to the drug itself and 1/4 due to the expectancies that surround drug therapies. Ken also found that stimulant drugs were more effective than nonstimulants in reducing hyperactivity and that the effect of drug therapy was unrelated to the age or I.Q. of the child or to how hyperactivity was measured.

Place Table 2 about here

A third review was conducted by Julie Yu, a graduate student in Marketing. Julie was interested in how response rates to questionnaires were affected by the research design. She examined over a dozen different techniques that surveyers use to increase whether or not an individual agrees to complete a questionnaire.

The unique aspect of Julie's task was that all studies employed identical dependent variables, namely the percentage of contacted individuals who agree to respond. Thus, rather than working with study probabilities and effect sizes, it was possible to directly combine raw data. Literature searches of BRS/Inform, Management Contents, Psychological Abstracts and the Social Science Citation Index uncovered 25 relevant studies and 60 more were found through a manual search of references in bibliographies.

Table 3 examines the effect of a monetary incentive on response rate. For each condition, a weighted average response rate is given along with the number of contacts and separate response rates the average rate is based on. The standard deviation of the rates is also presented. Four rates were differentiated. Experimental and control rates are based on studies that explicitly manipulated the presence or absence of monetary incentives. The without control rate is based on studies in which all participants received an incentive and the absent rate is based on studies in which no participant was paid.

Place Table 3 about here

A chi-square statistic for the experiment versus control frequencies of responding was highly significant with an associated phi-coefficient of +.15. A descriptive correlation was also generated by pairing response rates with the amount of money offered. This correlation equalled +.61, indicating greater monetary incentives led to higher responding. These analysis procedures were applied to each technique. Julie found significantly higher response rates associated with both prepaid and promised monetary incentives, nonmonetary rewards, preliminary notification, personalization of the request, and follow-up contacts. The effects of a cover letter, assurances of anonymity, providing a deadline, and providing return postage were all nonsignificant.

Conclusion

Obviously, this discussion has not done justice to the detail and complexity of these reviews, but I hope the general point is clear. More scientific guidelines for conducting integrative research reviews

are not only desirable but they are feasible to apply. Also, rigorous criteria will not produce reviews that are uncreative or mechanical in nature. The expertise and intuition of the reviewer will be challenged to capitalize on the opportunities for mining information unique to each problem area. Scientific reviews, however, should have much greater potential for creating consensus among scholars for focusing discussion on specific and testable areas of disagreement where conflict does exist.

References

Cooper, H. Statistically combining independent studies: A meta-analysis of sex differences in conformity research. Journal of Personality and Social Psychology, 1979, 37, 131-146.

Cooper, H. Scientific guidelines for conducting integrative research reviews. In press, Review of Education Research.

Findley, M. and Cooper, H. The relation between locus of control and achievement. In press, Journal of Personality and Social Psychology.

Glass, G., McGaw, B. & Smith, M. Meta-Analysis in Social Research. Beverly Hills: Sage, 1981.

Gottfredson, S. Evaluating psychological research reports. American Psychologist, 1978, 33, 920-934.

Mahoney, M. Publication prejudices: An experimental study of confirmatory bias in the peer review system. Cognitive Therapy and Research, 1977, 1, 161-175.

Matheson, D., Bruce, R. & Beauchamp, K. Experimental psychology (3rd ed.). New York: Holt, Rinehart & Winston, 1978.

Ottenbacher, K. and Cooper, H. Drug treatment of hyperactivity. Manuscript under review, 1981.

Presby, S. Overly broad categories obscure important differences between therapies. American Psychologist, 1978, 33, 514-515.

Rosenthal, R. Summarizing significance levels. New Directions for Methodology of Social and Behavioral Science, 1980, 5, 33-46.

Smith, M. Publication bias and meta-analysis. Evaluations in Education, 1980, 4, 22-24.

Yu, J. and Cooper, H. A literature review of research design effects on response rates to questionnaires. Manuscript under review, 1981.

Figure 1

The Literature 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 Which Create Variation in Review Conclusions	1. Differences in abstractness of definition. 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 may make review conclusions less general. 2. Superficial operational detail may obscure interacting variables.	1. Accessed studies may be qualitatively different from the target population of studies. 2. People sampled in accessible studies may be different from target population of people.	1. Nonquality factors may cause improper weighting of study information. 2. Omissions in study reports may make conclusions unreliable.	1. Rules for distinguishing patterns from noise may be inappropriate. 2. Review-based evidence may be used to infer causality.	1. Omission of review procedures may make conclusions irreproducible. 2. Omission of review findings and study procedures may make conclusions obsolete.

Table 1

Average Effect Size for Subgroupings of Study Characteristics

Characteristics	Average Correlations	SD	N[1]
Gender			
Males	+.20	.14	27
Females	+.11	.18	18
Age			
College	+.14	.15	32
High school	+.23	.10	8
Junior high	+.35	.22	7
4th-6th	+.24	.15	21
1st-3rd grade	+.04	.06	4
Race			
Black	+.25	.47	3
White	+.25	.17	8
Socioeconomic Status			
Middle class	+.26	.11	9
Lower class	+.35	.34	4
Locus of Control Measure			
General	+.18	.16	15
Specific	+.30	.22	12
Achievement Measures			
Classroom-related	+.16	.15	45
Standardized Achievement	+.21	.17	36
Standardized Intelligence	+.24	.15	12

Note 1. N is the number of studies upon which the average correlation and SD are based.

From: Findley, M. and Cooper, H. The relation between locus of control and achievement. In press, Journal of Personality and Social Psychology.

Table 2

d-indexes for the comparisons of drug versus control, drug versus placebo, and placebo versus control.

Stem	Drug vs Control	Drug vs Placebo	Placebo vs Control	Total[1]
2.1	5			5
2.0		8		8
1.9	4			4
1.8				
1.7	0	24		024
1.6		1		1
1.5	5	0688		05688
1.4	8	126	3	12368
1.3	2	02		022
1.2				
1.1	16	8		168
1.0		8	6	68
.9	39	12399	3	12333999
.8	4	05	9	0549
.7	9	5		59
.6		1125779		1125779
.5	09	07	0	00079
.4		2489		2489
.3	46	168	8	146688
.2		478		78
.1		56		56
.0		005	00000	00000005
Maximum	2.77	2.08	1.43	2.77
Q_3	1.55	1.30	.93	1.30
Median	1.10	.69	.19	.80
Q_1	.59	.42	.00	.38
Minimum	.34	.00	-1.30	-1.30
Mean	1.21	.84	.32	.84
SD	.67	.54	.72	.60

[1] Note: Two values, 2.77 and -1.30, are not included in the table.

From: Ottenbacher, K. and Cooper, H. Drug treatment of hyperactivity. Manuscript under review, 1981.

Table 3

Effect of Monetary Incentives on Response Rate

Monetary Incentive	Weighted Average Response Rate	Number of Contacts	Number of Response Rates	SD of Response Rates
Experimental	51.9	5,021	48	21.2
Control	39.0	2,794	30	20.7
W/o Control	48.3	3,056	20	17.9
Absent	20.1	961	3	21.8
Amount:				
$ 0.10	41.6	1,484	17	9.5
$ 0.25	54.2	2,549	12	25.0
$ 0.50	34.7	1,035	9	12.9
$ 1.00	35.9	697	5	19.9
$ 2.00	41.0	200	1	0.0
$ 3.00	40.5	200	1	0.0
$ 5.00	61.4	1,062	15	14.3
$10.00	82.0	314	2	5.9
$25.00	54.1	205	2	22.5
$50.00	75.0	83	1	0.0

From: Yu, J. and Cooper, H. A literature review of research design effects on response rates to questionnaires. Manuscript under review, 1981.