Scientific Guidelines for Conducting Integrative Research Reviews

Harris M. Cooper
University of Missouri-Columbia

The inferences made in integrative research reviews are as central to the validity of behavioral science knowledge as those made in primary research. Therefore, research reviewers must pay the same attention to rigorous methodology that is required of primary researchers. This article conceptualizes the research review as a scientific inquiry involving five stages that parallel those of primary research. The functions, sources of variance, and potential threats to validity associated with each stage are described.

The behavioral sciences recently underwent a sharp increase in manpower and research (Garvey & Griffith, 1971). To accommodate this expansion, outlets for research reports became plentiful and their accessibility was facilitated by the computerized literature search. The scholarly activity affected most by the research explosion was the integrative research review, or the synthesis of separate empirical findings into a coherent whole. As the empirical base expanded, the reviewer's task became more complex while simultaneously taking on added status. Today most researchers find they cannot keep abreast of primary data reports except within a few specializations. Researchers rely heavily on integrative research reviews to define the state of knowledge.

Because of the changes in reviewing, researchers can no longer take the conclusions of reviews at face value. They must recognize that the integration of separate research projects involves scientific inferences as central to the validity of knowledge as the inferences made in primary data interpretation. While substantial attention has been paid to validity issues in primary research (Bracht & Glass, 1968; Campbell, 1969; Campbell & Stanley, 1966), behavioral scientists have no systematic guidelines for evaluating the validity of review outcomes.

This article conceptualizes the integrative review as a research process containing five stages: (1) problem formulation; (2) data collection; (3) evaluation of data points; (4) data analysis and interpretation; and (5) presentation of results. Each stage serves...
a function similar to the one it serves in primary research. Differences in review methodologies, like differences in primary research methodologies, create variation in conclusions. Most important, methodological choices at each review stage may engender threats to the validity of the review's conclusions (see Table I).

A Definition of Research Review

According to Jackson (1980):

Some [reviewers] are primarily interested in sizing up new substantive and/or methodological developments in a given field. Some are primarily interested in verifying existing theories or developing new ones. Some are interested in synthesizing knowledge from different lines of research, and still others are primarily interested in inferring generalizations about substantive issues from a set of studies directly bearing on those issues. (p. 438, italics added)

The fourth kind of review, which will be called integrative, is this paper's primary focus. The goal of an integrative review is to summarize the accumulated state of knowledge concerning the relation(s) of interest and to highlight important issues that research has left unresolved (Taveggia, 1974). From the reader's viewpoint, an integrative research review is intended to (a) replace papers that have fallen behind the research front (Price, 1965), and (b) direct future research so that it yields a maximum amount of new information.

The Problem Formulation Stage

The variables involved in a behavioral science inquiry are defined in two ways. First, the variables are given conceptual definitions. Conceptual definitions differ in abstractness: If the meaning of one concept is included in the meaning of another, the more general concept is considered more abstract (Reynolds, 1971). Second, variables are operationally defined. An operational definition relates an abstract concept to observable events. Both primary researchers and research reviewers must define concepts and specify the operations included in the definitions.

Primary researchers have little choice but to operationalize their concepts before the inquiry begins. In contrast, the research reviewer can evaluate the concept relevance of different operations as they are encountered in the search for relevant studies. Of course, most reviewers begin with some a priori specification of operations. It is not unusual, however, for a reviewer to come across concept-relevant operations during the search that were not initially considered for inclusions.

A more significant distinction between primary research and reviewing is that primary research typically involves only one or two operational definitions. Research review can, and usually does, involve many different operations defining the same concept.

How Operational Diversity Affects Review Outcomes

Operational diversity affects review outcomes in two ways. First, the operational definitions chosen by research reviewers can vary. Two reviewers using an identical label for a concept may employ 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.
<table>
<thead>
<tr>
<th>Stage Characteristics</th>
<th>Problem Formulation</th>
<th>Data Collection</th>
<th>Data Evaluation</th>
<th>Analysis and Interpretation</th>
<th>Public Presentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research Question Asked</td>
<td>What evidence should be included in the review?</td>
<td>What procedures should be used to find relevant evidence?</td>
<td>What retrieved evidence should be included in the review?</td>
<td>What procedures should be used to make inferences about the literature as a whole?</td>
<td>What information should be included in the review report?</td>
</tr>
<tr>
<td>Primary Function in Review</td>
<td>Constructing definitions that distinguish relevant from irrelevant studies.</td>
<td>Determining which sources of potentially relevant studies to examine.</td>
<td>Applying criteria to separate &quot;valid&quot; from &quot;invalid&quot; studies.</td>
<td>Synthesizing valid retrieved studies.</td>
<td>Applying editorial criteria to separate important from unimportant information.</td>
</tr>
<tr>
<td>Procedural Differences</td>
<td>Differences in included in operational definitions.</td>
<td>Differences in the research contained in sources of information.</td>
<td>Differences in quality criteria.</td>
<td>Differences in rules of inference.</td>
<td>Differences in guidelines for editorial judgment.</td>
</tr>
<tr>
<td>That Create Variation Review Conclusions</td>
<td>Differences in operational detail.</td>
<td>Differences in the influence of nonquality criteria.</td>
<td>Differences in the influence of nonquality criteria.</td>
<td>Differences in the influence of nonquality criteria.</td>
<td>Differences in the influence of nonquality criteria.</td>
</tr>
<tr>
<td>Sources of Potential Invalidity in Review Conclusions</td>
<td>Differences in included in operational definitions.</td>
<td>Differences in the influence of nonquality criteria.</td>
<td>Differences in the influence of nonquality criteria.</td>
<td>Differences in the influence of nonquality criteria.</td>
<td>Differences in the influence of nonquality criteria.</td>
</tr>
<tr>
<td></td>
<td>1. Narrow concepts might make review conclusions less definitive and robust.</td>
<td>1. Accessed studies might be qualitatively different from the target population of studies.</td>
<td>1. Nonquality factors might cause improper weighting of study information.</td>
<td>1. Rules for distinguishing patterns from noise might be inappropriate.</td>
<td>1. Omission of review procedures might make conclusions inappropriate.</td>
</tr>
<tr>
<td></td>
<td>2. Superficial operational detail might obscure interacting variables.</td>
<td>2. People sampled in accessible studies might be different from target population of people.</td>
<td>2. Omissions in study reports might make conclusions unreliable.</td>
<td>2. Review-based evidence might be used to infer causality.</td>
<td>2. Omission of review findings and study procedures might make conclusions obsolete.</td>
</tr>
</tbody>
</table>
Second, diversity among studies related to the same concept means reviewers can vary in their treatment of operations after the data archives have been searched. One reviewer may decide to meticulously identify the operational distinctions among studies while another ignores the finer points. The first reviewer might recognize that the outcome of reviewed studies is mediated by a methodological variation that the second reviewer did not examine. Therefore, two reviewers employing identical conceptual definitions and reviewing the same set of studies may still reach different conclusions.

*Threats to validity.* Reviewers who focus on a limited set of operational definitions typically do so to ensure consensus about the meaning of a concept. However, multiple realizations of concepts are desirable in behavioral research; if multiple operations produce similar results they rule out rival conceptualizations of the findings (Campbell & Fiske, 1959; Webb, Campbell, Schwartz, & Sechrest, 1973). Also, very narrow definitions provide little information about whether a finding applies across various situations. Thus, reviewers who employ broad conceptual definitions (or who believe many operations are concept-relevant) can potentially reach more definitive and robust conclusions than reviewers using narrow definitions.

The second threat to validity is associated with how study operations are treated in the review. Lack of attention to study details might mask important distinctions in results. 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). Reviewers who examine more operational details probably will produce more valid review conclusions. These reviewers present more information about the contextual variations that do and do not influence the review conclusion.

*Protecting validity.* Reviewers should protect their conclusions from threats to validity. Archive searches should begin with the broadest possible conceptual definition in mind. The reviewer should begin with a few central operations but remain open to the possibility that other relevant operations will be discovered during the search. To complement conceptual broadness, reviewers should be exhaustive in their attention to distinctions in study procedures. Any suspicion that differences in study results are associated with procedural distinctions should receive attention.

Reviewers should also allow the users of their work to assess the degree to which these validity threats are present. As completely as possible, reviewers should (a) describe all the operational variations that were considered concept-relevant, and (b) report all variations in study methods that were related to study outcomes.

The Data Collection Stage

The major decision during data collection involves choosing the population of elements that will be the referent for the inquiry. The *target population* includes those elements the inquirer hopes to represent in the study. The *accessible population* includes those elements the inquirer is pragmatically able to obtain (Bracht & Glass, 1968). Both primary research and research review involve specifying target and accessible populations and considering how they might differ from one another.

Identifying populations for research reviews is complicated by the fact that reviews involve two targets. First, the reviewer wants the findings to pertain to *all previous research* on the problem. Reviewers exert some control over whether this goal is achieved through their choice of which and how many data archives to search.
Second, the reviewer hopes that the retrieved studies will allow generalization to the unit of analysis that interests the topic area. Here, the reviewer is constrained by the types of units sampled by primary researchers.

**Techniques for Information Retrieval**

There are at least five techniques a reviewer can use to retrieve information on a research problem: (1) the “invisible college” approach; (2) the ancestry approach; (3) the descendency approach; (4) the use of abstracting services; and (5) the on-line computer search. The invisible college is the most informal approach. Crane (1969) notes that “scientists working on similar problems are usually aware of each other and in some cases attempt to systematize their contacts by exchanging reprints with one another” (p. 335). The ancestry approach retrieves information by “tracking” citations from one study to another. Most reviewers are aware of several studies bearing on their problem, and these studies provide bibliographies which cite earlier, related research. The descendency approach, or the *Science or Social Science Citation Indexes*, is employed to retrieve studies that cite papers central to a topic and then screen these for topic relevance. To use abstracting services, the reviewer selects a set of keywords or phrases and compares them with the indicators used to index studies. In the on-line computer search, the computer exhaustively scans abstracting services and citation indexes at phenomenal speed. A problem with using the computer is that it eliminates “browsing,” or following up promising leads that arise during a manual search.

**How Information Retrieval Techniques Affect Review Outcomes**

Every past study does not have an equal chance of being retrieved by the reviewer. It is likely that studies contained in all the above sources are different from studies that never become public information. For instance, Greenwald (1975) found that about half of researchers who produced a rejection of the null hypothesis would submit a report for publication, while only 6 percent who failed to reject the null hypothesis would attempt to publish. McNemar (1960) speculates that findings that contradict conventional wisdom are relatively less likely to be visible to other researchers. Studies that are not available in the retrieval system do not create variability in review conclusions because they are absent from all reviews. However, discrepancies between review conclusions are created by differences in the sources reviewers use to retrieve information. Two reviewers using different techniques to locate studies may end up with different evidence and reach different conclusions. Regrettably, little is known about how information sources differ. This problem is complicated by the fact that the effect of information source probably varies from one topic area to another. Some generic differences, however, may be offered, based on commonly held beliefs.

The studies available through invisible colleges are probably (a) more homogeneous in operations, and (b) less carefully scrutinized for methodological flaws than all relevant studies available to a reviewer. Invisible college studies are also probably more uniformly supportive of the findings of central researchers than evidence found through more diverse sources. A similar homogeneity of findings and methods might be anticipated for studies located through citation indexes. However, because citation
indexes primarily contain studies that have been scrutinized by editors or dissertation advisors, fewer methodologically flawed studies should be included than in informal communications. Searching bibliographies is likely to overrepresent published research, particularly research within a circumscribed communication network (Xhignesse & Osgood, 1967).

Finally, abstracting services and computer searches probably contain the studies most closely approximating all publicly available research. These sources have the least restrictive requirements for a study to gain entry into the system. Their limitation is that there is typically a long lag between when a study is completed and when it will appear in the abstracts. Clearly, a reviewer who is well connected to an active invisible college is more likely to retrieve current research than a reviewer who must rely solely on the abstracting services.

**Threats to validity.** It is necessary to examine the adequacy of a reviewer’s archive search with respect to two targets. The review must be evaluated by (a) how the retrieved studies might differ from all studies, and (b) how the units contained in retrieved studies might differ from all units of interest.

The first threat to validity associated with the data-gathering phase of reviewing, then, is that the studies in the review might not include, and probably will not include, all studies pertinent to the topic of interest. The second threat to validity is that the units in retrieved studies might not represent all units in the target population. The reviewer cannot be faulted for the existence of this threat if retrieval procedures were exhaustive.

**Protecting validity.** Reviewers should access as many information sources as possible to ensure that as many studies as possible are located. The biases contained in one source can be partially obviated through the use of another source. In their manuscripts, reviewers should be explicit about how studies were gathered, including information on sources, years, and keywords covered in the search. Reviewers should present other indices of potential retrieval bias if they are available. For instance, Rosenthal and Rubin (1978) distinguished journal research from dissertations in an attempt to determine if the evidence from the two sources differed. Cooper, Burger, and Good (1981) reviewed only journal studies but speculated that there was little bias because the report titles rarely mentioned the hypothesis of interest.

The research reviewer should also describe the sample characteristics of individuals used in the separate studies. Missing samples and overrepresented samples should be discussed with reference to their potential impact on the findings.

The Evaluation of Data Stage

After data are collected, critical judgments are made about the quality of individual data points. Each data point is examined in light of surrounding evidence to determine if it is contaminated by too many factors irrelevant to the problem of interest. These procedures are performed whether the data pertain to the units of interest or to the results of studies.

**How Evaluative Criteria Create Variance in Reviews**

Review outcomes can differ because reviewers differ over how reliable they think the results of individual studies are. This variance in conclusions is created by a divergence in reviewers’ criteria for evaluating the quality of research. A demonstration of differences in qualitative judgments was carried out by Gottfredson (1978).
He studied editors and authors in nine psychological journals and suggested that interjudge agreement was relatively modest. Gottfredson reported intraclass coefficients ranging from .16 to .49 on nine subscales of an evaluative device.

Variance in review conclusions also occurs because reviewers differ in the degree to which factors other than research quality affect their evaluative decisions. One extraneous factor is the reviewer's prior expectations concerning the review outcome. Lord, Ross, and Lepper (1979), for instance, found that readers rated proattitudinal studies as better conducted than counterattitudinal studies. More strikingly, the study evaluators showed polarization in attitudes even though they all read the same two research abstracts.

**Threats to validity.** The use of any evaluative criteria other than substantive methodological discriminations is a threat to the validity of a research review. As Mahoney (1977) 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 data evaluation is wholly beyond the control of the reviewer. This threat involves the potential for unreliable outcomes due to incomplete reporting by primary researchers. Many research reports omit discussion of some hypotheses that were tested. Other reports give only incomplete information on the tests that were mentioned. If a reviewer must estimate or omit what happened in these studies, wider confidence intervals must be placed around review conclusions.

**Protecting validity.** It is difficult to suggest the kinds of substantive criteria reviewers should use for judgments about methodology. The topic is far too complex and opinions are too varied for brief treatment. Some suggestions can be made, however, about how criteria should be applied. For instance, reviewers should develop evaluative criteria before the literature is searched. Criteria for excluding studies should be stated as objectively as possible. More than one evaluator should be employed, and interevaluator agreement should be quantified and reported. The persons applying the criteria should be blind to the author, outlet, and results of the study.

The Analysis and Interpretation Stage

During analysis and interpretation, the separate data points are synthesized into a unified statement about the research problem.

Until recently, there has been little similarity in how analysis and interpretation was carried out by primary researchers and integrative research reviewers. Primary behavioral researchers have been obligated to present summary statistics and to substantiate the existence of any aggregate relations with probability tests. While statistical aids to interpretation have been criticized (Bakan, 1966; Cornfield & Tukey, 1956; Lykken, 1968), most primary researchers feel uncomfortable about synthesizing data without some assistance (or credibility) supplied by statistical procedures.

Integrative reviewers have not been obligated to apply *any* standard analysis and interpretation techniques in their synthesis process. Most frequently, reviewers interpret data using inexplicit rules of inference. This potential for subjectivity has led some critics to voice considerable skepticism about the outcome of many reviews.
HARRIS M. COOPER

Gene Glass (1976) writes: "A common method for integrating several studies with inconsistent findings is to carp on the design or analysis deficiencies of all but a few studies—those remaining frequently being one's own work or that of one's students or friends—and then advance the one or two acceptable studies as the truth of the matter" (p. 4).

The Quantitative Literature Review

The information explosion in the behavioral sciences has focused considerable attention on the lack of standardization in how reviewers arrive at general conclusions. For some topic areas, a separate verbal description of all relevant studies will now be impossible. Focusing on one or two studies chosen from dozens or hundreds will fail to accurately portray the accumulated state of knowledge (Cooper, 1979). The present day reviewer also faces problems when attempting to relate variance in the results of studies with variance in procedures. The reviewer will find a distribution of results for studies sharing a particular procedural characteristic and an overlap in the distributions of results involving different procedures.

Quantitative reviewing techniques have been suggested as a remedy to this problem (Glass, McGaw, & Smith, 1981; Rosenthal, 1980). As Glass (1977) notes, "The accumulated findings of... studies should be regarded as complex data points, no more comprehensible without statistical analysis than hundreds of data points in a single study..." (p. 352). The application of quantitative inference procedures to reviewing seems to be a crucial response to the expanding literature (Cooper, 1981; Cooper & Arkin, 1981). The value of quantitative reviewing, however, has been questioned along lines similar to those used to criticize primary data analysis (Barber, 1978; Eysenck, 1978). The question is still open, and both sides will probably revise their positions before the debate is over.

How Interpretation Techniques Affect Review Outcomes

Review conclusions can differ because reviewers employ different interpretation techniques. A systematic relation that cannot be distinguished from noise under one set of rules might be identifiable under another set. While the relative validity of different inference strategies is difficult to assess, Cooper and Rosenthal (1980) did demonstrate some of the objective differences between quantitative and nonquantitative procedures for research review. They asked graduate students and faculty to summarize seven research articles on a simple hypothesis. All reviewers evaluated the same set of studies, but half used quantitative procedures and half used whatever criteria appealed to them. Quantitative reviewers reported more support for the hypothesis and a larger relationship between variables than did traditional reviewers. Quantitative reviewers also tended to view future replicative research as less necessary than traditional reviewers, although this finding did not reach statistical significance.

Threats to validity. The first threat to validity accompanying analysis and interpretation is that the rules of inference a reviewer employs may be inappropriate. In nonquantitative reviews, it is difficult to gauge the appropriateness of inference rules because they are not often made public. For quantitative reviews, the suppositions underlying 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. To be more specific, research reviews contain two different sources of evidence about relations. The first, study-based evidence, comes from single studies testing whether the relation of interest is or is not present. Research reviews also contain review-based evidence. As mentioned earlier, reviewers try to associate differences in study results with differences in study instruments, participants, and/or testing conditions (Light, 1979). Review-based evidence is unique to examinations of accumulated results.

Study-based evidence is capable of establishing causal precedence among variables while review-based evidence is always purely associational. The problem with review-based evidence is not causal direction because it would be unreasonable to argue that a study's outcome caused the investigator's choice of variables. However, another ingredient of causality, nonspuriousness, is problematic. A plethora of other procedural characteristics are potentially confounded with the original researcher's choice of variables. These third variables cannot be eliminated as an explanation for review-based findings because the reviewer did not randomly assign procedures to experiments.

Protecting validity. Recommendations about what interpretative assumptions are appropriate for a reviewer to make will depend on the purposes and the peculiarities of a problem area. This is as true of quantitative procedures as of nonquantitative procedures. However, reviewers should open their rules of inference to public inspection by stating them explicitly. If there is any evidence bearing on the validity of the rules, it should be presented.

Also, reviewers should be careful to distinguish study- from review-based inferences. The potentially more equivocal nature of review-based inferences means that if this type of evidence indicates that a relation exists, the reviewer should call for the relation to be tested within a single study.

The Public Presentation Stage

The translation of an inquirer's notes, printouts, and remembrances into a public document is a task with profound implications for the accumulation of knowledge. The importance of the public presentation of results is readily acknowledged, but suggestions about how dissemination is best carried out are limited. Apparently the scientific community gives considerable latitude to an inquirer concerning what information about a study to make public and how the information should be presented (Ziman, 1969). Both primary researchers and literature reviewers are confronted by these editorial judgments.

The literature reviewer's dilemma may be similar in kind to that of the primary researcher, but the dilemma is more dramatic in degree. Reviewers have no formal guidelines describing how to structure the final report. At best, the reviewer follows informal guidelines provided by research reviews on the same or related topics. In most cases, the reviewer chooses a format convenient for the particular review problem.

How Editorial Judgments Create Variance in Reviews

The variation that differences in editorial judgments create is not found in the magnitude or direction of conclusions, but in the particular design aspects and results
that are included in the report. One reviewer might believe that a methodological characteristic or result of the review would only clutter the manuscript. A second reviewer might think that same information would be of interest to some readers and decide that the clutter is worthwhile.

**Threats to validity.** The two threats to validity accompanying report writing relate to the different target populations of the review. First, the omission of details about how the review was conducted is a potential threat to validity. As with primary research, an incomplete report reduces the replicability of the review conclusion. A review's external validity is threatened to the extent that a complete knowledge of review procedures is not approximated in the report.

The second validity threat in report writing involves the omission of evidence about units and relations that other inquirers find important. Matheson, Bruce, and Beauchamp (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 that are (or will be) important to an area.

**Protecting validity.** This article contains many suggestions that might guide research review writing. If they are followed, reviewers will probably find themselves using the presentation format (Introduction/Methods/Results/Discussion) already familiar to behavioral scientists. However, reviewers will never be able to perfectly predict which omitted characteristic or result will eventually render their conclusions invalid or obsolete.

**Some Post Hoc Issues**

First, the five stages of reviewing contain 10 threats to validity. It is likely that other validity threats were overlooked in this treatment. Campbell and Stanley's (1966) list of validity threats to primary research was expanded by Campbell (1969), Bracht and Glass (1968), and Glass, Wilson, and Gottman (1975). It is expected that other threats to the validity of the integrative review will be uncovered.

Second, several of the threats to validity are shared by both primary research and literature review. This suggests that any threat associated with a particular research design will be applicable to the results of an integrative review in which the design is predominant.

Third, the standardization and exhaustiveness of the procedures presented here will reduce the variability in review conclusions most when applied to problem areas containing many studies. Of course, this does not mean that reviewers of highly specialized literatures or recent trends can ignore rigorous procedures. In addition, the accumulation of evidence will forever increase the irrelevance of this distinction.

Finally, it will be considerably more costly to undertake reviews using the guidelines set forth in this article. Also, reviewers cannot expect perfectly valid conclusions. The guidelines, then, must be viewed as optimal criteria which rarely will be achieved. This is the same spirit with which behavioral scientists presently undertake primary research.

**Conclusion**

This article began with the supposition that integrative research reviewing was a data-gathering exercise that needed to be evaluated against scientific criteria. It was
INTEGRATIVE RESEARCH REVIEWING

proposed that because of the growth in empirical research, the conclusions of research reviews in the behavioral sciences would become less and less trustworthy unless something is done to standardize the process and make it more rigorous. Hopefully, the conceptualization has convinced readers that it is possible and desirable to require reviews to be more scientific. Because of the increasing role that reviews play in our definition of knowledge, it seems that these adjustments in procedures are inevitable if behavioral scientists hope to retain their claim to objectivity.

References


Light, R. Capitalizing on variation: How conflicting research findings can be helpful for policy. *Educational Researcher*, 1979, 8(9), 7-11.
HARRIS M. COOPER


AUTHOR

HARRIS M. COOPER, Associate Professor of Psychology and Research Associate, Center for Research in Social Behavior, 111 East Stewart Rd., Columbia, MO 65211. Specializations: Research methodology; social psychology of education.