Chapter 10

Reviewing Empirically Based Manuscripts: Perspectives on Process*

by
Donald P. Schwab

Graduate School of Business and
Industrial Relations Research Institute
University of Wisconsin, Madison

While most Ph.D. programs train students to critically evaluate published research, I know of no program that provides formal training in the review and evaluation of manuscripts for purposes of making a publication recommendation. This paper is not offered as a tutorial on how such training might proceed, nor as a model of behavior for fledgling reviewers. Rather, it is designed to describe one person's approach to reviewing. Undoubtedly, part of my approach is idiosyncratic, not characteristic of reviewers in general. But I suspect that the criteria I use to evaluate manuscripts, the weights assigned to these criteria, and even some of the characteristics of the reviewing process itself are, in the main, descriptive of many reviewers.

Hopefully, the title calls attention to the fact that this paper focuses on the review and evaluation of *empirically* based manuscripts. Frankly, I have chosen to confine my comments to this type of review not only because I have greater experience conducting such reviews, but also because such reviews are easier to perform, and the criteria and review process are easier to describe. Two reasons account for this: First, empirical studies tend to be written in a

* Financial support in preparing this manuscript from the Graduate School, University of Wisconsin, Madison and helpful comments on an earlier draft from R. J. Aldag, L. L. Cummings and J. Paul Peter are gratefully acknowledged.

standardized fashion (introduction, method, results, and discussion) so that one can review them using a fairly standardized approach. Second and more important, we share belief systems about appropriate empirical procedures to a greater degree than we share values regarding appropriate norms for the conduct and evaluation of conceptual and literature review efforts (although a common methodology for the latter, meta-analysis, appears to be emerging).

The paper is organized into two major sections. The first describes the criteria and weighting system I employ when going through the evaluation process. Of course, editors are in a far better position to provide data on criteria that are *representative* of reviewers in general (Campbell, 1982). A description of the criteria I employ is included largely to facilitate an understanding of the second major section of the paper having to do with the process employed to conduct evaluations. I believe that the process is, to a considerable extent, determined by the criteria and weighting schema used.

It is the second part of the paper that may be of most interest to readers since the reviewing process has received less attention than appropriate evaluation criteria. Any research methods volume can tell us much about criteria, and even something about probable weighting of those criteria. However, processes to bring these criteria to bear in evaluating a manuscript for publication are difficult to even infer from such volumes.

I frankly most enjoyed writing the process section. I have given considerable thought to the criteria I employ in evaluating research. Criteria are often discussed with other reviewers. However, before being asked to write this paper, I had given little thought to the process I employ. It has been educational to reflect on that process.

Evaluation Criteria

Three decreasingly well defined criteria dominate my recommendations to journal editors. These are *technical merit, craftsmanship,* and *significance.* In terms of my final accept (or permit a rewrite followed by another review)/reject recommendation, these three are clearly weighted in the order listed. My reasons for this ordering will become clearer after an understanding of what these criteria mean.

Technical Merit

Campbell and Stanley (1966, pp. 1–6) called attention to the distinction between *internal* validity (the confidence one has in the causal model suggested *within* the study itself) and *external* validity (generalizability of the causal model *beyond* the study). They argued then that internal validity is primary. If we cannot be confident of the causal model within the study itself, generalization is irrelevant. Their reasoning forms an important part of my evaluative framework.

Thus, when examining a manuscript I pay close attention to causal issues.

How compelling are the theoretical hypotheses in terms of differentiating between independent (influencing) variables and dependent (consequential) variables? In what sequence was the data collected (dependent variables sometimes precede independent variables)? What sort of controls (statistical and especially design) have been built into the study? Unsatisfactory answers to these types of questions account for a substantial percentage of my negative decisions.

The initial Campbell and Stanley typology has been elaborated on by Cook and Campbell (1979, pp. 37–94) with the addition of statistical conclusion validity (can the study identify population covariation?), and construct validity (what is the correspondence between operationalizations and the conceptual meaning of variables?). Of the two, I place highest priority on the latter (see Schwab, 1980). I do so because of my belief that investigators and readers *think* in terms of constructs and relationships between constructs—not operations. That is, the inference is *always* made in studies of organization behavior (and the social sciences generally) that the operations employed measure the constructs. *Results are inevitably thought of in terms of the constructs.* If there are reasons to suspect that the operations are not representative of the constructs, I therefore believe the study should not be published.

Statistical conclusion validity, alternatively, does not figure prominently in my weighting schema. Statistical conclusion validity addresses the issue of the legitimacy of generalizing from sample to population, an issue involving statistical inference. Indeed Cook and Campbell discuss threats to statistical conclusion validity in terms of errors of statistical inference.

Of course, we know that almost all of the empirical studies published in our journals use *convenience,* not probability samples. From a statistical inference point of view, we conduct and publish *case studies.* Thus, if one took generalization to a population using statistical inference seriously, one would recommend rejecting nearly all manuscripts submitted.

I accept the view that case studies are worth publishing. Moreover, I accept the convention (and fiction) that tests of statistical significance within such studies provide evidence on whether the observed effect size is large enough to be considered as supportive of an hypothesis. (I do, however, try to insist that studies report effect size *magnitudes* as well.) My only excuse for accepting this convention is that if I did not, my publishing (and reviewing) career would no doubt end.

Consequently, the two most important components of technical merit from my perspective are internal validity (as defined by Cook and Campbell) and construct validity (which, because of the way we think about research, is as essential as internal validity). Statistical conclusion validity, alternatively, is not as important, given my acceptance of the type of research that we actually do. In this view, generalization to populations becomes an element of external validity. Accumulation *across* studies is necessary not only to make inferences about generalizations over settings and times, but also across the objects of the statistical analyses.

Craftsmanship

Craftsmanship is my term and hence requires some explanation. To me, a major component of craftsmanship involves writing. Is the manuscript clear (i.e., understandable)? Two elements of clear exposition come readily to mind because they frequently are sources of difficulty. The first involves the organization of the manuscript and the lack of redundancy (the two are correlated). The second, lesser problem, is lack of clarity, making it difficult to follow a manuscript. For example, authors frequently use synonyms to characterize constructs, and this can be confusing. "Employee satisfaction" in one sentence becomes "affect" in the next, and "job attitudes" elsewhere. One had best throw away the thesaurus when writing research reports. The objective is to communicate clearly and unambiguously—not cleverly, or even interestingly (although clear writing does not have to be uninteresting).

There are also substantive and technical components to craftsmanship. Examples of the former include whether or not the manuscript shows evidence that the author is aware of, and knowledgeable about relevant prior research and theory on the topic studied; and whether the references are appropriate and current. A major example of the latter is whether or not the statistics that can be calculated from the data reported in the manuscript (e.g., statistical significance levels for bivariate correlation coefficients) are correctly calculated by the author. Also included as a part of craftsmanship is how closely the manuscript as a whole (including tables and references) conform to the publication requirements of the journal.

My weighting of craftsmanship may be too great because I clearly make attributions about the integrity of the entire research project from that small part of the study revealed to the reviewer (i.e., the manuscript). If the manuscript itself is not carefully crafted, then I worry about the integrety of the entire research study.

Significance

When I first started reviewing, technical issues dominated my approach to reviewing almost exclusively. If I could not find a technical reason to recommend that a manuscript be rejected, I was loath to employ any other criterion, including craftsmanship (I simply recommended that changes be made in the manuscript). I was especially reluctant to make a judgment about the significance or importance of the study. I am still reticent about making such a decision since the importance of a study is, to a considerable extent, an inherently subjective judgment.

Increasingly, however, I have assumed some responsibility for making a decision about the significance of a manuscript when making my final decision about a recommendation. My thoughts here are not as clearly formalized as with either technical merit or craftsmanship. Nevertheless, I expect a publishable study to contribute in at least one of three not mutually exclusive ways. Assuming

that the study has technical merit and acceptable craftsmanship, I judge a study to be significant if it (a) tests a reasonable theory in some nontrivial and nonredundant manner, (b) improves on an existing body of empirical research, theoretically based or not, or (c) has implications for some important policy issue.

These criteria are admittedly vague, and not very rigorous. About the only kinds of studies that fail all three tests are pure data crunching pieces that have no obvious relevance to prior theory or research (e.g., studies that are of interest only to the organization that sponsored the research), or studies that do not make any contribution beyond what is already known from the empirical literature. Nevertheless, I occasionally reach such a conclusion.

The Process

After being asked to write this paper, my procedure for generating this section was to reflect on the process I employ when reviewing manuscripts (it might be characterized as introspective-protocol-recall). This recall was supplemented by more consciously reflecting and taking notes on the process during some recent reviewing. Such a procedure is certainly accurate about major points (e.g., I take notes as I read a manuscript and more often than not, read it two times before writing up my evaluation). It may, however, be subject to some distortion regarding the more subtle points and the frequency with which certain events occur (no one tried to teach me introspective data generation and analysis when I was a graduate student). No doubt the process has evolved over time, but I do not trust my ability to recall details regarding how I started out reviewing some 10 to 12 years ago, nor much about how that process has changed in the intervening time period. Hopefully, what appears here is a reasonably accurate description of the process as I currently perform it.

However, a final caveat: After reading the description below and despite all the qualifying terms ("typically," "usually," "often," etc.), the process still appears more systematic and organized than it really is. Sometimes I conduct a review almost exactly as it is characterized below. More often there are deviations of one sort or another from the process described.

Decision Heuristics

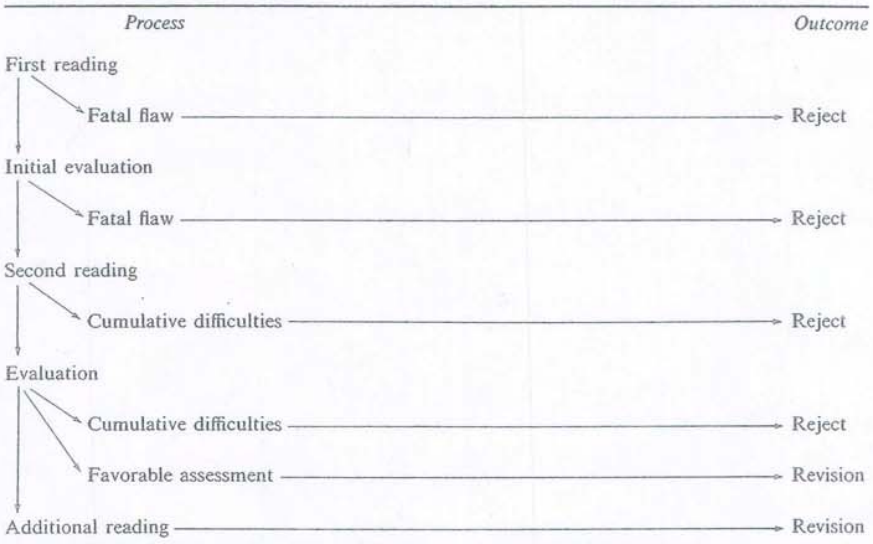
One has only to know something of the acceptance base rate to have priors that any particular manuscript will not be acceptable. The general process I use to review manuscripts reflects this fact. Figure 1 shows a schematic of the process and the outcomes that follow. Taking the latter first, editors generally entertain one of three alternative recommendations from their reviewers: (a) rejection of the manuscript out of hand, (b) manuscript revision for rereview or, less frequently, for acceptance without further review, and (c) acceptance without revision.

The procedure I use can lead to rejection at nearly any step in the process. During the first reading and initial evaluation, rejection may occur because

the manuscript contains a *fatal flaw.* Fatal flaws represent uncorrectable difficulties with the manuscript so serious that a rejection recommendation is warranted. Thus fatal flaws are a sufficient reason for rejection even if the paper is otherwise acceptable, or could otherwise be made acceptable. Given the criteria discussed above, fatal flaws most frequently involve issues of technical merit, especially internal or construct validity. I do not recall finding a fatal flaw with respect to external validity and have never found a fatal statistical flaw. (After all, if the problem was only statistical, it could be corrected.)

In one, perhaps not very professional sense, the fatal flaw is the reviewer's salvation. It allows us to efficiently evaluate a study since we do not have to document all of the other difficulties of the manuscript, repairable or not. Indeed, I may not even read the entire manuscript if a fatal flaw is encountered.

FIGURE 1
Overview of the Review Process



It is not necessary for a manuscript to contain a fatal flaw to warrant rejection. Cumulative difficulties, often spanning a gamut of methodological and substantive concerns may also lead to a recommendation that the manuscript be rejected. That judgment, given the procedure I use, is usually reserved to a second reading or an evaluation that follows it. If after two readings and two evaluations I have not concluded that the manuscript should be rejected, subsequent readings (typically of only part of the paper) are likely to lead only to recommended revisions; not rejection. I don't recall ever recommending that a manuscript be accepted without some revision on initial submission.

First Reading

Scratch paper on the desk, pencil in hand, I prepare to review a manuscript. How do I proceed with the review? What, if anything, goes through my mind as I proceed?

In spite of my knowledge of the acceptance base rate, I generally begin reading a manuscript in a positive or at least not negative frame of mind. Perhaps this study will be the exception, a study that I can learn from and that will make a contribution to our understanding of the field. During this initial reading I seldom evaluate the significance or importance of the topic being addressed (criteria employed only after deciding the study has technical merit).

I tend to remain positive, more or less uncritical, throughout my reading of the introductory material as long as it is well crafted. If, however, the introduction is poorly crafted, my attitude becomes more critical, even while reading the introduction. I tend to get more suspicious, looking for things that might be incorrect in the paper. Are there inconsistencies in the arguments? Are the citations appropriate? These sorts of evaluative assessments undoubtedly always float close to the surface of my consciousness. A poorly crafted manuscript simply draws them into my conscious mind state.

Manuscripts often are poorly crafted, although I sense some improvement on this dimension since I began reviewing. Campbell (1982) listed writing (a major element of craftsmanship) as the third most frequent reason manuscripts were rejected. In my own case, I do not know what percentage of manuscripts I have recommended be rejected on the basis of writing or craftsmanship alone; certainly some, but typically these are not the only reasons for recommending rejection. In theory at least, writing and other elements of craftsmanship could be improved (hence not fatal) and so I am reluctant to recommend rejection on the basis of these issues alone.

There is a point, however, where I feel the burden of proof shifts from the reviewer to the submitter. If the manuscript is so incompetently or indifferently crafted that it is difficult or impossible to determine what was done in the study, then I believe the manuscript can be legitimately rejected without further review. At the very least, a poorly crafted manuscript gets the critical juices flowing, and sometimes (although less so recently) leads to a review that is not constructive.

This perspective may bother some readers, although I suspect many reviewers share my sentiments. The sort of model editors encourage is that reviewers always cast their comments in a constructive vein, and that the review be as thorough and complete as possible. All difficulties with the study should be identified, and where possible, helpful suggestions be made about how the manuscript under review, or subsequent research, could be improved.

This is an attractive normative model. I still feel embarrassed by the reminder from an editor that careers are at stake as a function of the review process. This admonition followed a particularly scathing review I submitted. Nevertheless, I still make attributions based on the craftsmanship demonstrated in a manuscript. And, I can still become very critical of an author if I see a manuscript

a second time (often submitted to a different journal) that has not corrected errors identified in an earlier review.

Even if the manuscript is well crafted through the introduction, I assume a more critical (not negative) frame of mind on reaching the method section since this is where (given the weighting model I employ) one is most likely to find fatal flaws. Emphasis, of course, is placed on an examination of issues having to do with internal and construct validity.

Assuming the method section survives this initial examination for fatal flaws, I proceed to the results section. This section is usually acceptable, or at least salvageable. This is probably because the results section is usually so short and straightforward. If an ANOVA has been promised in the method section, it is usually delivered in the results section. Often there are stylistic problems in presenting the results, but usually these are not serious. There are, too frequently, statistical calculation problems as well (an element of craftsmanship). I tend to look for such errors more religiously if the manuscript is otherwise deficient (although sometimes not at all if the manuscript is fatally flawed).

Initial Evaluation

After reading the results section the first time, I usually stop to make an initial evaluation of the manuscript. This is undoubtedly a change from when I was a neophyte reviewer and read straight through a manuscript before pausing to reflect on what contribution it might suggest. I no longer do that very often. Regardless of what the author has to say in the discussion, I want to reach my own conclusions about the study. Thus, I usually do not read the discussion section the first time through.

Instead, I usually take some kind of a break after completing the results section. It is nice when my first reading ends around lunch or supper time. But since that does not work out too often, I may simply walk up and down the hall, or more likely, go bother a colleague for a while. The point of this activity is to give me an opportunity to think about the design and analytical procedures employed in the study. Does the design address the issues that the author purports to study? Are the measuring procedures likely to be appropriate? If the study is experimental or quasi-experimental, do the manipulations likely get at the theoretical constructs of interest? Are the statistical procedures appropriate for the questions addressed?

Often I will decide during this initial period of evaluation that the manuscript is fatally flawed, should not be accepted because of cumulative difficulties, or that it needs revision before a decision is appropriate. As a very rough estimate, I would guess that in over 50 percent of my reviews a decision has been made to recommend rejection of the manuscript by this stage of the review process. Correctly or incorrectly, I have convinced myself that one or more fatal flaws exist, or that the cumulative effects of the difficulties in the manuscript warrant recommending that it be rejected.

Second Reading

If my decision is negative, then the second reading of the manuscript is designed to justify that decision. I make sure that my original understanding of the manuscript is correct, but now I also begin to take notes and write down points that I will probably want to make in the review itself. If the study has a fatal flaw(s), then I may be content to only point it (them) out. On other occasions I will point out other difficulties as well. My choice of approaches on this issue may not be very professional since it is largely a function of two factors. First, if I feel the author is trying to do good research I am more inclined to try to be helpful, even though my decision on the manuscript at hand is negative. I must confess that my attributions about the author's intentions may not be very accurate, and again emphasize craftsmanship. A major clue that the author is not very serious are the presence of obvious errors. These may be typographical errors, errors in citations, calculation errors in tables, unexplained sample size differences, and the like. Very many of these suggest to me that the author did not care very much for the quality of the study performed.

Second, additional comments designed to be helpful, as opposed to demonstrating why I have found the manuscript to be unacceptable, depend on the time I have to devote to the review. This, in turn, is a function of how many other reviews I have to perform and on when the reviews are due. I usually allocate certain days or parts of days to conduct reviews (e.g., Saturday mornings) so the time constraints I experience tend to center on reviews rather than on other activities.

If my initial decision is not negative, the second reading proceeds somewhat as the first, only this time it is more critical. In reviewing the introduction, I try to more carefully assess the appropriateness of the author's arguments and hypotheses, both in terms of existing literature and in terms of the current study. I also more critically evaluate the way the paper is written. However, I still do not reflect much on the significance of the issues addressed. I generally want to examine the methodology a second time before passing judgment on the importance of getting this manuscript into the literature.

My second reading of the method and results section is analogous to my second reading of the introductory material. I simply go through these sections more critically. In this reading, I may occasionally decide I missed a fatal flaw in the first reading, but that is unusual. I am, however, quite likely to decide at this point that the composite of less than fatal flaws makes the manuscript unpublishable. If so, my mind set and note writing assume the characteristics of decision confirmation as described above.

If, after reading through to the discussion section a second time, I have not reached a negative decision, then it is at this point that the author's interpretation of the findings becomes an issue. I tend to focus on one central issue when reading the discussion section. Is the discussion consistent with the methodology and findings? Frequently the answer to this question must be negative. Nonstatistically significant findings mystically reemerge as important in the dis-

cussion. More frequently, statistically significant findings that explain virtually no variance in the dependent variable become of practical significance in the discussion. It is not entirely uncommon to find discussion of variables that were not even included in the study!

The most common difficulty of this sort, indeed almost universally so, is that study findings are generalized far beyond the subjects and environment studied. If there is one affliction that is common to almost all researchers in our field (myself no doubt included), it is the penchant to infer more generalizability in our studies than is warranted by our procedures. Findings obtained using scales of questionable validity are suddenly discussed as if the constructs were perfectly measured. Surveys conducted on a convenience sample of employees from a single organization are discussed as if the findings apply to all employees in all organizations. Results obtained from a short experimental exercise (often conducted on college students) are discussed as if the findings apply to real organizations and real employees. The transition from study to generalization frequently appears to be a leap of blind faith that the author encourages the reviewers and readers to take.

Despite the fact that discussion sections frequently call for revision, I cannot recall ever finding one to be fatally flawed. Again, if the study is otherwise meritorious, the discussion can be repaired. Nor has it been my experience to find the author's discussion of the shortcomings of the study (a practice I strongly endorse—it is a component that strengthens the discussion section) so persuasive that my own decision is modified. (However, my independent judgment may have led me to conclude that the issues raised by the author make the manuscript unpublishable.) In short, the discussion section generally does not figure prominently in my accept/reject recommendation even though it often does call for suggested modifications.

The Decision

If I have not decided at some prior step to recommend that the manuscript be rejected on technical (method and measurement) grounds, it is at this point that I pass judgment on the significance of the study. If this study was published, would it make a contribution to our understanding of work force behavior? Is the study significant enough to warrant publication?

For the most part, this decision is noncompensatory. That is, the manuscript must demonstrate technical merit, craftsmanship, and be significant (as defined). A technically excellent study will not compensate for a study on a trivial issue. However, over a relatively narrow range, I believe a compensatory model explains my behavior. That is, I do trade off among the three criteria to some extent. If a study is very well done, well crafted, I'm likely to require less significance than if I believe the study has some nonfatal flaws. One can certainly argue with this judgment, but I believe it accurately reflects my decision process.

My judgment regarding the study's significance is usually made immediately following my decision that the study has technical merit, and this is probably

unfortunate. At least in several instances I have recommended to an editor that a manuscript be encouraged (i.e., it had technical merit, adequate craftsmanship, and was significant), only to decide later (overnight or when the revision was resubmitted) that the study really was not worth publishing because it lacked significance. That is, even though the study was well done, its findings contributed little or nothing to the field.

Discussion

Writing this paper has been an unusual experience. I am accustomed to writing reports based on data collection and analysis, reviewing literature, and in a few cases, making theoretical and analytical arguments. I am unaccustomed to simply describing behaviors, especially my own. Use of the word "I" in writing, especially so frequently, is embarrassing. Nevertheless, as noted earlier, the exercise has been enjoyable, and I feel that I have gained some insight about my reviewing process.

Over the years I have discussed evaluation criteria with many others who review for the same journals. With some understandable differences on specifics, my perception of these conversations is that we tend to share the general evaluative model described here. Nearly all seem to weight craftsmanship fairly heavily, although we do not share a common term for the construct. It is true that some reviewers place relatively greater weight on significance and less on technical merit. However, even here I believe that most reviewers of empirical manuscripts place heavy weight on technical issues. After all, we are best trained to make "good-bad" judgments of a defensible nature on technical matters.

I have not discussed the process used to evaluate manuscripts with others. Indeed, as noted earlier, I had not thought about it until being asked to write this paper. In retrospect, I believe that the process I employ is largely a function of the criteria employed and their weights. Because technical issues dominate my evaluative model, it is quite reasonable to concentrate on research method first. While I probably did not examine a manuscript so selectively on first reading when I started out reviewing, the process evolved quite naturally over time. To the extent that other reviewers share my evaluative model, it is reasonable to suppose that their process of reviewing is similar to the one described here.

References

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

Campbell, J. P. (1982). Editorial: Some remarks from the outgoing editor. *Journal of Applied Psychology, 67,* 691–700.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston: Houghton Mifflin.

Schwab, D. P. (1980). Construct validity in organizational behavior. In Staw, B., and Cummings, L. L. *Research in organizational behavior, 2,* 3–43. Greenwich, CT: JAI Press.