

Reviewing Articles for Methods

Robert J. Sternberg and Elena L. Grigorenko

Felix Flooglefitter concludes that humans and other animals are not the only ones with feelings. Plants have feelings too. How does he know? He grows a set of 100 plants in one room of a house. He creates a happy mood for these plants, piping in happy music to them, playing happy TV shows in the background, and every day giving them a positive inspirational message about his belief that they can grow to be big and tall. He places another set of 100 plants in a different room of the house. He creates a sad mood for these plants, piping in sad music, playing depressing TV shows in the background, and telling them every day that they are failures as members of the plant kingdom and that their growth will be stunted. After 2 months, he finds that the “happy” plants have indeed grown significantly more than the “unhappy” ones. He is mortified when a prestigious journal rejects an article of such monumental importance.

This obviously fictitious example points out the importance of methodological rigor in empirical studies. The study was blatantly flawed. Did Flooglefitter randomly assign plants to the two conditions? Were the rooms equivalent in their exposure to sunlight? Did he water the plants in the two rooms equally? Did he control for soil? Did he make sure room temperatures and humidity levels were the same? No matter how monumental a

finding may seem to be, if it emerges from poorly designed research, it is worthless. One really can validly conclude nothing at all.

Let us start with how not to review for methods. The way not to review is to have a set of methods you prefer—experimental ones, biological ones, or whatever—and then judge what you read in terms of the extent to which it matches your preconception of how whatever problem is being studied really should have been studied. Bad reviewers require that studies follow a methodology that they, in their infinite wisdom, have decided is the correct one. Do not be that reviewer. If you are, you will soon have nothing to review. Editors will realize your incompetence as a reviewer and react accordingly.

Ideally, researchers should use *converging operations*, or a variety of methods that together converge on a given conclusion or set of conclusions about a particular psychological phenomenon. A conclusion is strengthened if multiple methodologies lead to it. Thus the goal is not to find some “right” method but, rather, to find that a number of methods all lead in the same direction.

What you should do is review an article in terms of how well it applies whatever methodology it uses, and in terms of whether the conclusions follow from the data, given the methodology or methodologies used. No article is likely to be wholly conclusive. What is important is that the author or authors recognize what they can and cannot, however much they might wish to, conclude.

Methodology may seem like one of the boring aspects of an article to which to pay attention. But arguably it is the most important. If the methods of a study are inadequate, then the results, no matter how interesting, will not hold up. A study cannot yield compelling data if the methods were not adequate to the questions being investigated. Hence, methods are key to the publishability of an article.

At the same time, reviewers can become so attuned to methods that they lose the forest for the trees. Remember that methods, although important, are not the be-all and end-all of an article. Especially in new areas of research, adequate methods may just be forming. Reviewers typically are more forgiving in a brand-new area of research than one in which the paradigms are

highly developed. But if the methods are inadequate, the study probably will not be publishable, no matter how new the area of research.

When you review, be constructive. If there are methodological flaws, point them out, but also point out how the methods could be improved. Investigators can often conduct follow-up studies that correct methodological weaknesses of earlier studies, and thereby turn articles that are not publishable into ones that are. As a reviewer, your job is not just to be critical but to be a constructive critic who helps the author or authors of an article improve on the work they have done in the submitted article, in future articles, or ideally, in both.

Authors sometimes remember, over a period of years, referees who helped them improve the methodologies by which they studied the phenomena of interest to them. You can become one of those memorable referees if you give the time and effort to your review.

There are many things you need to watch out for when you review an article for methods. In this chapter, we can only begin to scratch the surface in mentioning these things.

Participants

The following is a list of questions to ask when you read about the participants in a study.

1. *Is the number of participants sufficient to generate the power needed to test the stated hypotheses?* In general, the more error variance there is in the data, the larger the N an investigator will need to be able to draw sound conclusions. The more treatment variance there is in the data, the smaller the N an investigator will need. So numbers of participants need to be evaluated relative to the variability of the data.
2. *Are the participants drawn from an appropriate population for testing?* Many researchers test college sophomores or other samples of convenience and then are rather quick to draw conclusions that extend well

beyond their sample population. You should evaluate whether the sample population is appropriate for the study and whether the conclusions are generalized to an appropriate level. For example, if someone wants to study conceptions of the nature of romantic love, testing only college sophomores might give a biased view because of the relatively limited experience such students are likely to have had with the phenomenon under investigation.

3. *Are participants balanced in ways that are relevant to the study question?* If all or most of the participants are of one kind or another, then there may be a balance problem. For example, in the study of conceptions of romantic love, it might be a problem if 80% of the sample are men (or women). If one wishes to measure a prejudice of some kind, it might be important to balance urban, suburban, and rural participants. Be sure to check for balance of participants.
 4. *If the study is longitudinal, be especially aware of two issues.*
 - a. *What is the dropout rate?* Conclusions of longitudinal studies can be compromised by selective dropout. Over time, participants in longitudinal studies often tend to drop out. Is the dropout rate so high as to compromise conclusions? Such a dropout rate can be especially problematic if it is selective. So ask whether there is any evidence of selective dropout, and if so, what has been done to correct for it.
 - b. *Are there secular effects that might compromise conclusions?* Different age cohorts live through different experiences. For example, people born in 1920 would have lived through World War II. Those born in 1960 grew up during a time of societal ferment. Those born in 1980 have grown up in an era of widespread computer use. The point is that the experiences of different cohorts may be different, and conclusions drawn from one cohort may not be generalizable to another one. Moreover,
-

there are secular changes in societies. For example, Flynn (1987) found that IQs rose at a rate of about 3 points per 10 years during much of the 20th century. So test norms for one testing of a cohort might not apply in a later testing.

5. *If the study is cross-sectional, be especially aware of two issues.*
 - a. *Are the cohorts compared truly comparable?* Precisely because cohorts have lived through different experiences, comparisons of psychological functioning across cohorts can be fraught with danger. For example, younger participants may be at an advantage in computerized testing simply because they are more familiar with the use of computers. Older participants may be further away from their school years and hence have forgotten much of what they learned during that time and, for that reason alone, may do worse than younger cohorts on tests requiring skills learned in school.
 - b. *Are the tests being used comparable for the different age groups?* If the same tests are used to ensure item comparability, then the tests may be too hard for young participants and too easy for older ones. If different tests are used, the scores on the tests simply may not be directly comparable for the different age groups. What is being done to ensure comparability?

Materials

The following are questions to ask when you read about the materials in a study.

1. *Are the materials truly appropriate for the developmental, linguistic, and cultural backgrounds of the participants being tested?* Investigators are not always sensitive to the issue of appropriateness of the materials they use. For example, materials in English may present

totally different challenges for participants whose first language is not English than for native English speakers. Sometimes the problem is not linguistic but cultural. Use of certain materials may make assumptions about cultural customs or even knowledge that is not in fact shared across cultures. For example, participants outside the United States cannot be expected to have the same knowledge of American history and geography as would participants in the United States, just as people in the United States would lack information about the history and geography of other peoples. Similarly, one needs to make sure that the materials are developmentally appropriate for the sample in the study. This consideration is especially relevant when the samples in the study are drawn from special populations (e.g., individuals with disabilities).

2. *Do participants truly understand the materials the way the investigator thinks they do?* Piaget (1972) made some seemingly reasonable but incorrect inferences from children's behavior because he assumed that the children understood the materials the way he did. For example, when asked, "Which do I have, more marbles or more green marbles?" children may become confused and give a wrong answer simply because they do not see the question as making sense. Cole, Gay, Glick, and Sharp (1971) found that Kpelle tribesmen thought that keyed answers to certain questions were stupid, whereas their own unkeyed answers were smart. They did not understand the tasks the way the investigators did. Similarly, Luria (1976) found that peasant farmers often simply refused to accept logical-reasoning problems that violated their understandings of the world.
 3. *Were the materials motivating to participants, or at least, equally motivating?* Participants may fail to do well on a task because they become bored with it. Even worse, different subgroups of participants may become differentially bored, leading to results that dif-
-

fer across these subgroups as a result of motivation for the task rather than whatever variable is supposedly being manipulated in the study. Make sure, therefore, that motivation, and even more important, differential motivation, is not a factor underlying the results, unless that is indeed what is being studied.

4. *Were the materials adequate for representation of concepts in the study?* Make sure that materials in the study are relevant to the description of the basic study concepts in the introduction. It might be the case that the materials do not map on the concepts the author or authors discussed in the introduction. Of course, this mismatch will create difficulties for interpreting the findings; therefore, the reviewer should make sure that there is no inconsistency in the way the main concepts of the study are introduced and measured.
5. *Was the description of materials satisfactory?* It is the task of the reviewer to make sure that the description of materials in the article is clear and comprehensive. If the authors use well-known instruments and tasks, only a brief description of these instruments is warranted. However, if some or all instruments and tasks are original, then they should be described in detail so the readers can evaluate the properties of these instruments and tasks and other researchers can recreate them if needed for result verification and replication purposes.

Design

When considering design, ask yourself these questions:

1. *How were participants assigned to groups?* Was assignment random? If yes, how was randomization carried out and does this study present estimates of power conducted at the level of randomization? In

other words, if a researcher randomizes schools, are there enough schools in each group to verify the hypothesis of the study? If a researcher randomizes school districts, are there enough districts to detect the effect of interest? If not, what provisions were made, such as analysis of covariance or a pretest-posttest design, to ensure that obtained differences are not due to prior differences in groups rather than to any treatment that may have been introduced?

2. *Were there adequate control groups?* Many investigators have invested huge amounts of time in conducting investigations in which no clear conclusions could be drawn for lack of adequate control groups. For example, merely showing that a treatment improved learning from pretest to posttest is not an adequate design because participants may simply have become better at the task over time or they may show a practice effect on the posttest that has nothing to do with the treatment administered. It is not enough merely for an investigator to say he or she had a control group. As a referee, you must ensure that the comparison group adequately controlled for plausible confounding variables. If no such group appeared, then you need to ensure that adequate statistical manipulations were introduced to factor out the effects of extraneous variables. When a control group is present, it is important to make sure that treatment (experimental) and comparison (control) groups are comparable. Typically, the researchers are expected to ensure and demonstrate the comparability of treatment and comparison groups. To ensure this comparability, many researchers specify the parameters of matching (e.g., gender, age, and ethnicity) and explain how the recruitment of the comparison group was carried out. However, it is the reviewer's responsibility to verify the soundness of the research design, including the author's strategies for the recruitment of control groups or statistical control for possible confounding effects.
-

3. *Was the overall design well thought through?* It is important to think of all possible aspects of the design that are relevant to the data. For example, if a study is a double-blind treatment control study, then the reviewer should pay attention to precautions introduced by the author to prevent the revelation of group assignments prior to the completion of the study, especially to the individuals who work in the field. If it is an educational interventional study, an element of fidelity evaluation should be built in so that the researchers have control over how their intervention will be delivered in classrooms. If the generated data require expert judges (e.g., scoring children's compositions or drawings), make sure that the authors established reliability in evaluations.
4. *Was the dependent variable (or were the dependent variables) adequate for the hypotheses being investigated and the conclusions being drawn?* You must first ensure that the dependent variable is clearly specified. You must then make sure that it was adequate and appropriate for the hypothesis under investigation. For example, if conclusions are to be drawn about personality, was the test of personality reliable and well validated? Was it sufficiently broad to cover all personality traits that might be relevant to the questions being posed? Or, if an investigator is looking at intelligence but only uses a limited test of intelligence, can one be reasonably confident that the test measured intelligence as a whole rather than just a part of it?
5. *Are the independent variables adequate for the hypotheses being investigated and the conclusions being drawn?* First make sure that the independent variables are clearly specified. Then make sure that they correspond adequately to the underlying constructs they are supposed to represent. If they are supposed to be orthogonal, are they? Is the design clean, in the sense that it is possible to test clearly for their effects

independently of each other and in interaction? If the independent variables are correlated, what provisions have been made, such as a correlational design, to tease out their separate effects?

Procedure

The following list presents some questions to ask about the procedure.

1. *Is the procedure clearly specified?* It is essential that someone reading the study be able to understand the procedure adequately enough to be able to replicate it. If you could not replicate it on the basis of the information given, then the account is not adequate.
 2. *Are there missing steps?* Sometimes investigators have weak points in their procedure that they are not eager for referees to spot. They may, intentionally or unintentionally, obfuscate certain details regarding these points of procedure. Make sure that all steps are specified without equivocation or ambiguity.
 3. *Is the procedure appropriate for the participants and hypotheses being tested?* You cannot assume that the procedure is appropriate because the researcher believes it was or tries to convince you it was. It is up to you to ensure that the procedure did indeed accomplish what the investigator claims it accomplished.
 4. *Is the equipment described?* Often studies require the use of particular equipment (e.g., MRI scanners, genotyping systems, or GPS devices). Make sure that all instruments essential to the work are properly described. Depending on the equipment used, results might vary from one lab to another. Therefore, all relevant machinery and all specific parameters for running this machinery should be clearly described and labeled with the manufacturer's name and location along with the model number.
-

5. *Was there adequate informed consent and debriefing?* Although human-subjects committees approve almost all studies that are submitted to journals these days, you cannot merely assume that informed consent and debriefing were adequate. You need to ensure that appropriate steps were taken.
6. *Was anything unethical done in the study?* Again, although a human-subjects committee probably approved the study, it is up to you to ensure that all appropriate ethical safeguards were introduced into the procedures of the study.
7. *Was the procedure appropriate for the ages, cultural backgrounds, or other demographic aspects of the sample or samples being tested?* As a referee, you need to ensure that the procedures were appropriate to the participant groups being tested. You cannot merely assume that they were because the investigators believed them to be.
8. *If errors were made in procedures, were they of a nature that would fatally compromise the study?* Sometimes, clocks stop working, computers fail, investigators deliver the wrong stimuli to the wrong participants, and so forth. The investigators are obligated to notify readers of such facts. You need to ask yourself the extent to which the study has been compromised as a result of errors, and whether the errors might have fatally jeopardized the study.

Hypotheses

Sometimes hypotheses are specified in the method section and sometimes they are not. But wherever they appear, they are potentially critical to the study. So ask yourself the following questions.

1. *Is it clear how the hypotheses follow from whatever theory is being offered?* Hypotheses should not merely come out of the blue. They should fit into some theoretical

framework. Do they? And if so, do they truly follow from the theory?

2. *Are the hypotheses consistent with past data? If not, are reasons for the divergence clearly explained?* Hypotheses may or may not fit with past data. If they do, it should be made clear how they fit. If they do not, then the investigator needs clearly to explain why.
3. *Are the hypotheses plausible?* Make sure that any hypotheses presented make sense. If they do not, challenge them.
4. *Are the hypotheses interesting?* If the hypotheses are trivial, then the study may lack substance, regardless of what the findings are. So it is incumbent on you to ensure that the hypotheses are interesting and not merely trivial extensions of past work or of ideas that will make no real new contribution to the literature.
5. *Are the hypotheses mutually consistent?* Do the hypotheses fit with each other, or do any of them explicitly or implicitly contradict one another?
6. *Are the hypotheses testable?* In other words, are they framed in such a manner that they can be verified?

When you write a review, appreciate the fact that today you review someone's work and tomorrow your work will be reviewed. If you want to receive comprehensive and constructive comments when you are reviewed, make sure to be thorough and comprehensive when you review your colleagues' work!

References

- Cole, M., Gay, J., Glick, J., & Sharp, D. W. (1971). *The cultural context of learning and thinking*. New York: Basic Books.
- Flynn, J. R. (1987). Massive IQ gains in 14 nations. *Psychological Bulletin*, 101, 171–191.
- Luria, A. R. (1976). *Cognitive development: Its cultural and social foundations*. Cambridge, MA: Harvard University Press.
- Piaget, J. (1972). *The psychology of intelligence*. Totowa, NJ: Littlefield Adams.
-