Psychological Assessment
1995, Vol. 7, No. 3, 228-237

1040-3590/95/$3.00

Preparing and Evaluating Research Reports

Alan E. Kazdin
Yale University

Preparation of research reports for journal publication or dissemination in some other form is a central part of the research process. This article discusses preparation of the report in light of how the information is likely to be evaluated and how the report contributes to the research process. The focus is on three essential features: description, explanation, and contextualization of the study. These features are elaborated by reviewing the contents of each section of the manuscript and questions to guide authors and reviewers for preparing and evaluating the report. Emphasis is placed on conveying the rationale for decisions made in the design, execution, and analysis of the study. Common issues related to the interpretation of assessment studies, including test validity data, the relation of constructs and measures, and sampling, are highlighted as well.

The research process consists of the design and execution of the study, analysis of the results, and preparation of the report (e.g., journal article). The final step seems straightforward and relatively easy, given the nature and scope of the other steps. In fact, one often refers to preparation of the article as merely "writing up the results." Yet the implied simplicity of the task belies the significance of the product in the research process. The article is not the final step in this process. Rather, it is an important beginning. The article is often a launching platform for the next study for the authors themselves or for others in the field who are interested in pursuing the findings. Thus, the report is central to the research process.

The article itself is not only a description of what was accomplished, but it also conveys the extent to which the design, execution, and analyses were well conceived and appropriate. Recognition of this facet of the report is the reason why faculty require students in training to write a proposal of the study in advance of its execution. At the proposal stage, faculty can examine the thought processes, design, planned execution, and data analyses and make the necessary changes in advance. Even so, writing the full article at the completion of the study raises special issues. At that point, the authors evaluate critical issues, see the shortcomings of the design, and struggle with any clashes or ambiguities of the findings in light of the hypotheses.

The purpose of this article is to discuss the preparation and evaluation of research reports (articles) for publication.[1] Guidelines are presented to facilitate preparation of research articles. The guidelines cover the types of details that are to be included, but more important, the rationale, logic, and flow of the article to facilitate communication and to advance the next stage of the research process. Thus, preparation of a research report involves many of the same considerations that underlie the design and plan of the research.

Completion of this research was supported by Research Scientist Award MH00353 and Grant MH35408 from the National Institute of Mental Health.

Correspondence concerning this article should be addressed to Alan E. Kazdin, Department of Psychology, Yale University, P.O. Box 208205, New Haven, Connecticut 06520-8205.

Reports of empirical studies have many characteristics in common, whether or not they focus on assessment. Even so, the present focus will emphasize studies that are designed to evaluate assessment devices, constructs that the measures are intended to reflect, and studies of test validation. Issues that commonly emerge in articles of assessment and hence the design of assessment studies are highlighted as well.

Guidelines for Preparing Reports for Publication

Preparation of the report for publication involves three interrelated tasks, which I shall refer to as description, explanation, and contextualization. Failure to appreciate or to accomplish these tasks serves as a main source of frustration for authors, as their articles traverse the process of manuscript review toward journal publication. *Description* is the most straightforward task and includes providing details of the study. Even though this is an obvious requirement of the report, basic details often are omitted in published articles (e.g., gender and race of the participants, means, and standard deviation; see Shapiro & Shapiro, 1983; Weiss & Weisz, 1990). *Explanation* is slightly more complex insofar as this task refers to presenting the rationale of several facets of the study. The justification, decision-making process, and the connections between the decisions and the goals of the study move well beyond description. There are numerous decision points in any given study, most of which can be questioned. The author is obliged to make the case to explain why the specific options elected are well suited to the hypotheses or the goals of the study. Finally, *contextualization* moves one step further away from description of the details of the study and addresses how the study fits in the context of other studies and in the knowledge base more generally. This latter facet of

[1] Preparation of manuscripts for publication can be discussed from the perspective of authors and the perspective of reviewers (i.e., those persons who evaluate the manuscript for publication). This article emphasizes the perspective of authors and the task of preparing an article for publication. The review process raises its own issues, which this article does not address. Excellent readings are available to prepare the author for the journal review process (Kafka, *The Trial*, The Myth of Sisyphus, and Dante's *Inferno*).

article preparation reflects such lofty notions as scholarship and perspective, because the author places the descriptive and explanatory material into a broader context.

The extent to which description, explanation, and contextualization are accomplished increases the likelihood that the report will be viewed as a publishable article and facilitates integration of the report into the knowledge base. Guidelines follow that emphasize these tasks in the preparation and evaluation of research reports. The guidelines focus on the logic to the study; the interrelations of the different sections; the rationale for specific procedures and analyses; and the strengths, limitations, and place of the study in the knowledge base. It may be helpful to convey how these components can be addressed by focusing on the main sections of manuscripts that are prepared for journal publication.

Main Sections of the Article

Abstract. At first glance, the abstract certainly may not seem to be an important section or core feature of the article. Yet, two features of the abstract make this section quite critical. First, the abstract is likely to be read by many more people than is the article. The abstract probably will be entered into various databases that are available internationally. Consequently, this is the only information that most readers will have about the study. Second, for reviewers of the manuscript and readers of the journal article, the abstract sometimes is the first impression of what the author studied and found. Ambiguity, illogic, and fuzziness here are ominous. Thus, the abstract is sometimes the only impression or first impression one may have about the study. What is said is critically important.

Obviously, the purpose of the abstract is to provide a relatively brief statement of purpose, methods, findings, and conclusions of the study. Critical methodological descriptors pertain to the participants and their characteristics, experimental and control groups or conditions, design, and major findings. Often space is quite limited; indeed a word limit (e.g., 100- or 120-word maximum) may be placed on the abstract by the journals. It is useful to make substantive statements about the characteristics of the study and the findings rather than to provide general and minimally informative comments. Similarly, vacuous statements (e.g., "Implications of the results are discussed" or "Future directions for research are suggested") should be replaced with comments about the findings or one or two specific implications and research directions (e.g., "The findings raise the prospect that there is a Big One rather than a Big Five set of personality characteristics").

Introduction. The introduction is designed to convey the overall rationale and objective of the research. The task of the author is to convey in a clear and concise fashion why this particular study is needed and the current questions, void, or deficiency the study is designed to address. The section should not review the literature in a study-by-study fashion, but rather convey issues and evaluative comments that set the stage for the study that is to follow. The task of contextualization is critically important in this section. Placing the study in the context of what is and is not known and conveying the essential next step in research in the field require mastery of the pertinent literatures and reasonable communication skills. Saying that the study is important (without systematically establishing the context) or noting that no one else has studied this phenomenon often are viewed as feeble attempts to circumvent the contextualization of the study.

Limitations of previous work and how those limitations can be overcome may be important to consider. These statements build the critical transition from an existing literature to the present study and establish the rationale for design improvements or additions in relation to those studies. Alternatively or in addition, the study may build along new dimensions to advance the theory, hypotheses, and constructs to a broader range of domains of performance, samples, settings, and so on. The rationale for the specific study must be very clearly established. If a new measure is being presented, then the need for the measure and how it supplements or improves on existing measures, if any are available, are important to include. If a frequently used measure is presented, the rationale needs to be firmly established what precisely this study will add.

In general, the introduction will move from the very general to the specific. The very general refers to the opening of the introduction, which conveys the area of research, general topic, and significance of a problem. For example, if an article is on the assessment of alcohol abuse or marital bliss (or their interrelation), a brief opening statement noting the current state of the topic and its implications outside of the context of measurement is very helpful. Although reviewers are likely to be specialists in the assessment domain, many potential readers would profit from clarification of the broader context.

The introduction does not usually permit authors to convey all of the information they wish to present. In fact, the limit is usually two to four manuscript pages. A reasonable use of this space involves brief paragraphs or implicit sections that describe the nature of the problem, the current status of the literature, the extension that this study is designed to provide, and how the methods to be used are warranted. To the extent that the author conveys a grasp of the issues in the area and can identify the lacunae that the study is designed to fill greatly improves the quality of the report and the chances of acceptance for journal publication.

Method. This section of the article encompasses several points related to who was studied, why, how, and so on. The section not only describes critical procedures, but also provides the rationale for methodological decisions. Initially, the research participants (or subjects) are described, including several basic descriptors (e.g., age, genders, ethnicity, education, occupation, and income). From a method and design standpoint, information beyond basic descriptors can be helpful to encompass factors that plausibly could affect generality or replication of the results or that might influence comparison of the data with information obtained from normative or standardization samples.

The rationale for the sample should be provided. Why was this sample included and how is it appropriate to the substantive area and question of interest? In some cases, the sample is obviously relevant because participants have the characteristic or disorder of interest (e.g., parents accused of child abuse) or are in a setting of interest (e.g., nursing home residents). In other cases, samples are included merely because they are available (college students or a clinic population recruited for some other purpose than the study). Such samples of convenience often count against the investigator. If characteristics of the sample

are potentially objectionable in relation to the goals of the study, the rationale may require full elaboration to convey why the sample was included and how features of the sample may or may not be relevant to the conclusions the author wishes to draw. A sample of convenience is not invariably a problem for drawing valid inferences. Yet, invariably, a thoughtful discussion will be required regarding its use. More generally, participant selection, recruitment, screening, and other features warrant comment. The issue for the author and reviewer is whether features of the participant selection process could restrict the conclusions in some unique fashion or, worse, in some way represent a poor test of the hypotheses.

Assessment studies may be experimental studies in which groups vary in whether they receive an intervention or experimental manipulation. More commonly, assessment studies focus on intact groups without a particular manipulation. The studies form groups based on subject selection criteria (e.g., one type of patient vs. another, men vs. women) for analyses. The rationale for selecting the sample is obviously important. If the sample is divided into subgroups, it is as critical to convey how the groups will provide a test of the hypotheses and to show that characteristics incidental to the hypotheses do not differ or do not obscure interpretation of the results (see Kazdin, 1992). Also, the selection procedure and any risks of misclassification based on the operational criteria used (e.g., false positives and negatives) warrant comment. Reliability of the assessment procedures used to select cases, especially when human judgment is required, is very important because of the direct implications for interpretation and replication of the findings. A common example for which this arises in clinical research is in invoking psychiatric diagnoses using interview techniques.

Several measures are usually included in the study. Why the constructs were selected for study should be clarified in the introduction. The specific measures and why they were selected to operationalize the constructs should be presented in the method section. Information about the psychometric characteristics of the measures is often summarized. This information relates directly to the credibility of the results. Apart from individual assessment devices, the rationale for including or omitting areas that might be regarded as crucial (e.g., multiple measures, informants, and settings) deserves comment. The principle here is similar to other sections, namely, the rationale for the author's decisions ought to be explicit.

Occasionally, ambiguous statements may enter into descriptions of measures. For example, measures may be referred to as "reliable" or "valid" in previous research, as part of the rationale for use in the present study. There are, of course, many different types of reliability and validity. It is important to identify those characteristics of the measure found in prior research that are relevant to the present research. For example, high internal consistency (reliability) in a prior study may not be a strong argument for use of the measure in a longitudinal design in which the author hopes for test–retest reliability. Even previous data on test–retest reliability (e.g., over 2 weeks) may not provide a sound basis for test–retest reliability over annual intervals. The information conveys the suitability of the measure for the study and the rationale of the author for selecting the measure in light of available strategies.

Results. It is important to convey why specific analyses were selected and how a particular test or comparison addresses the hypotheses or purposes presented earlier in the article. It is often the case that analyses are reported in a rote fashion in which, for example, the main effects are presented first, followed by the interactions for each measure. The author presents the analyses in very much the same way as the computer printout that provided multiple runs of the data. Similarly, if several dependent measures are available, a particular set of analyses is automatically run (e.g., omnibus tests of multivariate analyses of variance followed by univariate analyses of variance for individual measures). These are not the ways to present the data.

In the presentation of the results, it is important to convey why specific tests were selected and how these tests serve the specific goals of the study. Knowledge of statistics is critical for selecting the analysis to address the hypotheses of interest and conditions met by the data. The tests ought to relate to the hypotheses, predictions, or expectations outlined at the beginning of the article (Wampold, Davis, & Good, 1990). Presumably, the original hypotheses were presented in a special (nonradom) order, based on importance or level of specificity. It is very useful to retain this order when the statistics are presented to test these hypotheses. As a general rule, it is important to emphasize the hypotheses or relations of interest in the results; the statistics are only tools in the service of these hypotheses.

It is often useful to begin the results by presenting basic descriptors of the data (e.g., means and standard deviations for each group or condition) so the readers have access to the numbers themselves. If there are patterns in the descriptors, it is useful to point them out. Almost-significant results might be noted here to err on the side of conservatism regarding group equivalence on some domain that might affect interpretation of the results, particularly if power (or sample size) was weak to detect such differences.

The main body of the results presents tests of the hypotheses or predictions. Organization of the results (subheadings) or brief statements of hypotheses before the specific analyses are often helpful to prompt the author to clarify how the statistical test relates to the substantive questions. As a step towards that goal, the rationale for the statistical tests chosen or the variations within a particular type of test ought to be noted. For example, within factor analyses or multiple regression, the options selected (e.g., method of extracting factors, rotation, and method of entering variables) should be described along with the rationale of why these particular options are appropriate. The rationales are important as a general rule, but may take on even greater urgency because of the easy use of software programs than can run the analyses. Default criteria on many software programs are not necessarily related to the author's conceptualization of the data, that is, the hypotheses. (Such information is referred to as "default criteria" because if the results do not come out with thoughtless analyses, it is partially "de fault of the criteria de investigator used.") Statistical decisions, whether or not explicit, often bear conceptual implications regarding the phenomena under investigation and the relations of variables to each other and to other variables.

Several additional or ancillary analyses may be presented to elaborate the primary hypotheses. For example, one might be able to reduce the plausibility that certain biases may have accounted for group differences based on supplementary or ancillary data analyses. Ancillary analyses may be more exploratory

and diffuse than tests of the primary hypotheses. Manifold variables can be selected for these analyses (e.g., gender, race, and height differences) that are not necessarily conceptually interesting in relation to the goals of the study. The author may wish to present data and data analyses that were unexpected, were not of initial interest, and were not the focus of the study. The rationale for these excursions and the limitations of interpretation are worth noting. From the standpoint of the reviewer and reader, the results should make clear what the main hypotheses were, how the analyses provide appropriate and pointed tests, and what conclusions can be reached as a result. In addition, thoughtful excursions (i.e., with the rationale guiding the reader) in the analyses are usually an advantage.

Discussion. The discussion consists of the conclusions and interpretations of the study and hence is the final resting place of all issues and concerns. Typically, the discussion includes an overview of the major findings, integration or relation of these findings to theory and prior research, limitations and ambiguities and their implications for interpretation, and future directions. The extent that this can be accomplished in a brief space (e.g., two to five manuscript pages) is to the author's advantage.

Description and interpretation of the findings may raise a tension between what the author wishes to say about the findings and their meaning versus what can be said in light of how the study was designed and evaluated. Thus, the discussion shows the reader the interplay of the introduction, method, and results sections. For example, the author might draw conclusions that are not quite appropriate given the method and findings. The discussion conveys flaws, problems, or questionable methodological decisions within the design that were not previously evident. However, they are flaws only in relation to the introduction and discussion. That is, the reader of the article can now recognize that if these are the types of statements the author wishes to make, the present study (design, measures, and sample) is not well suited for making them. The slight mismatch of interpretative statements in the discussion and the methodology is a common, albeit tacit basis for not considering a study as well conceived and well executed. A slightly different study may be required to support the specific statements the author makes in the discussion; alternatively, the discussion might be more circumscribed in the statements that are made.

It is usually to the author's credit to examine potential sources of ambiguity given that he or she is in an excellent position because of familiarity with procedures and expertise to understand the area. A candid, nondefensive appraisal of the study is very helpful. Here, too, contextualization may be helpful because limitations of a study are also related to prior research, trade-offs inherent in the exigencies of design and execution, what other studies have and have not accomplished, and whether a finding is robust across different methods of investigation. Although it is to the author's credit to acknowledge limitations of the study, there are limits on the extent to which reviewers grant a pardon for true confessions. At some point, the flaw is sufficient to preclude publication, whether or not is acknowledged by the author. At other points, acknowledging potential limitations conveys critical understanding of the issues and directs the field to future work. This latter use of acknowledgement augments the contribution of the study and the likelihood of favorable evaluation by readers.

Finally, it is useful in the discussion to contextualize the results by continuing the story line that began in the introduction. With the present findings, what puzzle piece has been added to the knowledge base, what new questions or ambiguities were raised, what other substantive areas might be relevant for this line of research, and what new studies are needed? From the standpoint of contextualization, the new studies referred to here are not merely those that overcome methodological limitations of the present study, but rather those that focus on the substantive foci of the next steps for research.

Guiding Questions

The section-by-section discussion of the content of an article is designed to convey the flow or logic of the study and the interplay of description, explanation, and contextualization. The study ought to have a thematic line throughout, and all sections ought to reflect that thematic line in a logical way. The thematic line consists of the substantive issues guiding the hypotheses and the decisions of the investigator (e.g., with regard to procedures and analyses) that are used to elaborate these hypotheses.

Another way to consider the tasks of preparing a report is to consider the many questions the article ought to answer. These are questions for the authors to ask themselves or, on the other hand, questions reviewers and consumers of the research are likely to want to ask. Table 1 presents questions that warrant consideration. They are presented according to the different sections of a manuscript. The questions emphasize the descriptive information, as well as the rationale for procedures, decisions, and practices in the design and execution. Needless to say, assessment studies can vary widely in their purpose, design, and methods of evaluation, so the questions are not necessarily appropriate to each study nor are they necessarily exhaustive. The set of questions is useful as a way of checking to see that many important facets of the study have not been overlooked.

General Comments

Preparation of an article often is viewed as a task of describing what was done. With this in mind, authors often are frustrated at the reactions of reviewers. In reading the reactions of reviewers, the authors usually recognize and acknowledge the value of providing more details that are required (e.g., further information about the participants or procedure). However, when the requests pertain to explanation and contextualization, authors are more likely to be baffled or defensive. This reaction may be reasonable because graduate training devotes much less attention to these facets of preparing research reports than to description. Also, reviewers' comments and editorial decision letters may not be explicit about the need for explanation and contextualization. For example, some of the more general reactions of reviewers are often reflected in comments such as "Nothing in the manuscript is new," "I fail to see the importance of the study," or "This study has already been done in a much better way by others."[2] In fact, such characterizations may be true. Alternatively, the comments could also reflect the

[2] I am grateful to my dissertation committee for permitting me to quote their comments at my oral exam. In keeping with the spirit embodied in their use of pseudonyms in signing the dissertation, they wish not to be acknowledged by name here.

Table 1
Major Questions to Guide Journal Article Preparation

Abstract
What were the main purposes of the study?
Who was studied (sample, sample size, special characteristics)?
How were participants selected?
To what conditions, if any, were participants exposed?
What type of design was used?
What were the main findings and conclusions?
Introduction
What is the background and context for the study?
What in current theory, research, or clinical work makes this study useful, important, or of interest?
What is different or special about the study in focus, methods, or design to address a need in the area?
Is the rationale clear regarding the constructs to be assessed?
What specifically were the purposes, predictions, or hypotheses?
Method
Participants
Who were the participants and how many of them were there in this study?
Why was this sample selected in light of the research goals?
How was this sample obtained, recruited, and selected?
What are the participant and demographic characteristics of the sample (e.g., gender, age, ethnicity, race, socioeconomic status)?
What if any inclusion and exclusion criteria were invoked (i.e., selection rules to obtain participants)?
How many of those participants eligible or recruited actually were selected and participated in the study?
Was informed consent solicited? How and from whom, if special populations were used?
Design
What is the design (e.g., longitudinal, cross-sectional) and how does the design relate to the goals of the study?
How were participants assigned to groups or conditions?
How many groups were included in the design?
How were the groups similar and different in how they were treated in the study?
Why were these groups critical to address the questions of interest?
Assessment
What were the constructs of interest and how were they measured?
What are the relevant reliability and validity data from previous research (and from the present study) that support the use of these measures for the present purposes?
Were multiple measures and methods used to assess the constructs?
Are response sets or styles relevant to the use and interpretation of the measures?
How was the assessment conducted? By whom (as assessors/observers)? In what order were the measures administered?
If judges (raters) were used in any facet of assessment, what is the reliability (inter- or intrajudge consistency) in rendering their judgments/ratings?
Procedures
Where was the study conducted (setting)?
What materials, equipment, or apparatuses were used in the study?
What was the chronological sequence of events to which participants were exposed?
What intervals elapsed between different aspects of the study (e.g., assessment occasions)?
What procedural checks were completed to avert potential sources of bias in implementation of the manipulation and assessments?
What checks were made to ensure that the conditions were carried out as intended?
What other information does the reader need to know to understand how participants were treated and what conditions were provided?
Results
What were the primary measures and data on which the predictions depend?
What are the scores on the measures of interest for the different groups and sample as a whole (e.g., measures of central tendency and variability)?
How do the scores compare with those of other study, normative, or standardization samples?
Are groups of interest within the study similar on measures and variables that could interfere with interpretation of the hypotheses?
What analyses were used and how specifically did these address the original hypotheses and purposes?
Were the assumptions of the data analyses met?
If multiple tests were used, what means were provided to control error rates?
If more than one group was delineated, were they similar on variables that might otherwise explain the results (e.g., diagnosis, age)?
Were data missing due to incomplete measures (not filled out completely by the participants) or due to loss of participants? If so, how were these handled in the data analyses?
Are there ancillary analyses that might further inform the primary analyses or exploratory analyses that might stimulate further work?
Discussion
What were the major findings of the study?
How do these findings add to research and how do they support, refute, or inform current theory?
What alternative interpretations can be placed on the data?
What limitations or qualifiers must be placed on the study given methodology and design issues?
What research follows from the study to move the field forward?

Note. Further discussion of questions that guide the preparation of journal articles can be obtained in additional sources (Kazdin, 1992; Maher, *1978*). Concrete guidelines on the format for preparing articles are provided by the American Psychological Association (1994).

extent to which the author has failed to contextualize the study to obviate these kinds of reactions.

The lesson for preparing and evaluating research reports is clear. Describing a study does not eo ipso establish its contribution to the field, no matter how strongly the author feels that the study is a first. Also, the methodological options for studying a particular question are enormous in terms of possible samples, constructs and measures, and data-analytic methods. The reasons for electing the particular set of options the author has chosen deserve elaboration.

In some cases, the author selects options because they were used in prior research. This criterion alone may be weak, because objections levied at the present study may also be appropriate to some of the prior work as well. The author will feel unjustly criticized for a more general flaw in the literature. Yet, arguing for a key methodological decision solely because "others have done this in the past" provides a very weak rationale, unless the purpose of the study is to address the value of the option as a goal of the study. Also, it may be that new evidence has emerged that makes the past practice more questionable in the present. For example, investigators may rely on retrospective assessment to obtain lifetime data regarding symptoms or early characteristics of family life, a seemingly reasonable assessment approach. Evidence suggests, however, that such retrospective information is very weak, inaccurate, and barely above chance when compared with the same information obtained prospectively (e.g., Henry, Moffitt, Caspi, Langley, & Silva, 1994; Robins et al., 1985). As evidence accumulates over time to make this point clear and as the domain of false memories becomes more well studied, the use of retrospective assessment methods is likely to be less acceptable among reviewers. In short, over time, the standards and permissible methods may change.

In general, it is beneficial to the author and to the field to convey the thought processes underlying methodological and design decision. This information will greatly influence the extent to which the research effort is appreciated and viewed as enhancing knowledge. Yet, it is useful to convey that decisions were thoughtful and that they represent reasonable choices among the alternatives for answering the questions that guide the study. The contextual issues are no less important. As authors, we often expect the latent Nobel Prize caliber of the study to be self-evident. It is better to be very clear about how and where the study fits in the literature, what it adds, and what questions and research the study prompts.

Common Interpretive Issues in Evaluating Assessment Studies

In conducting studies and preparing reports of assessment studies, a number of issues can be identified to which authors and readers are often sensitive. These issues have to do with the goals, interpretation, and generality of the results of studies. I highlight three issues here: test validation, the relations of constructs to measures, and sampling. Each of these is a weighty topic in its own right and will be considered in other articles in this issue. In this article, they are addressed in relation to interpretation and reporting of research findings.

Interpreting Correlations Among Test Scores

Text validation is a complex and ongoing process involving many stages and types of demonstrations. As part of that process, evidence often focuses on the extent to which a measure of interest (e.g., a newly developed measure) is correlated with other measures. Interpreting seemingly simple correlations between measures requires attention to multiple considerations.

Convergent validation. *Convergent validity* refers to the extent to which a measure is correlated with other measures that are designed to assess the same or related constructs (Campbell & Fiske, 1959). There are different ways in which convergent validity can be shown, such as demonstrating that a given measure correlates with related measures at a given point in time (e.g., concurrent validity) and that groups selected on some related criterion (e.g., history of being abused vs. no such history) differ on the measure, as expected (e.g., criterion or known-groups validity).[3] In convergent validity, the investigator may be interested in showing that a new measure of a construct correlates with other measures of that same construct or that the new measure correlates with measures of related constructs. With convergent validity, some level of agreement between measures is sought.

In one scenario, the investigator may wish to correlate a measure (e.g., depression) with measures of related constructs (e.g., negative cognitions and anxiety). In this case, the investigator may search for correlations that are in the moderate range (e.g., $r = .40–.60$) to be able to say that measure of interest was correlated in the positive direction, as predicted, with the other (criterion) measures. Very high correlations raise the prospect that the measure is assessing the "same" construct or adds no new information. In cases in which the investigator has developed a new measure, the correlations of that measure will be with other measures of the same construct. In this case, high correlations may be sought to show that the new measure in fact does assess the construct of interest.

Interpretation of convergent validation data requires caution. To begin with, the positive, moderate-to-high correlation between two measures could well be due to shared trait variance in the construct domains, as predicted between the two measures. For example, two characteristics (e.g., emotionality and anxiety) might overlap because of their common psychological, biological, or developmental underpinnings. This is usually what the investigator has in mind by searching for convergent validity. However, other interpretations are often as parsimonious or even more so. For example, shared method variance may be a viable alternative interpretation for the positive correlation. *Shared method variance* refers to similarity or identity in the procedure or format of assessment (e.g., both measures are self-report or both are paper-and-pencil measures). For example, if two measures are completed by the same informant, their common method variance might contribute to the magnitude of the correlation. The correlations reflect the shared method variance, rather than, or in addition to, the shared construct variance.

[3] There are of course many different types of validity, and often individual types are referred to inconsistently. For a discussion of different types of validity and their different uses, the reader is referred to other sources (Kline, 1986; Wainer & Braun, 1988).

The correlation between two measures that is taken to be evidence for validity also could be due to shared items in the measures. For example, studies occasionally evaluate the interrelations (correlations) among measures of depression, self-esteem, hopelessness, and negative cognitive processes. Measures of these constructs often overlap slightly, so that items in one particular scale have items that very closely resemble items in another scale (e.g., how one views or feels about oneself). Item overlap is not an inherent problem because conceptualizations of the two domains may entail common features (i.e., shared trait variance). However, in an effort of scale validation, it may provide little comfort to note that the two domains (e.g., hopelessness and negative cognitive processes) are moderately to highly correlated "as predicted." When there is item overlap, the correlation combines reliability (alternative form or test–retest) with validity (concurrent and predictive).

Low correlations between two measures that are predicted to correlate moderately to highly warrant comment. In this case, the magnitude of the correlation is much lower than the investigator expected and is considered not to support the validity of the measure that is being evaluated. Three considerations warrant mention here and perhaps analysis in the investigation. First, the absolute magnitude of the correlation between two measures is limited by the reliability of the individual measures. The low correlation may then underestimate the extent to which the reliable portion of variance within each measure is correlated. Second, it is possible that the sample and its scores on one or both of the measures represent a restricted range. The correlation between two measures, even if high in the population across the full range of scores, may be low in light of the restricted range. Third, it is quite possible that key moderators within the sample account for the low correlation. For example, it is possible that the correlation is high (and positive) for one subsample (men) and low (and negative) for another subsample. When these samples are treated as a single group, the correlation may be low or zero, and nonsignificant. A difficulty is scavenging for these moderators in a post hoc fashion. However, in an attempt to understand the relations between measures, it is useful to compute within-subsample correlations on key moderators such as gender, ethnicity, and patient status (patient vs. community) where relations between the measures are very likely to differ. Of course, the study is vastly superior when an influence moderating the relations between measures is theoretically derived and predicted.

Discriminant validity. Disciminant validity refers to the extent to which measures not expected to correlate or not to correlate very highly in fact show this expected pattern.[4] By itself, discriminant validity may resemble support for the null hypothesis; namely, no relation exists between two measures. Yet, the meaning of discriminant validity derives from the context in which it is demonstrated. That context is a set of measures, some of which are predicted to relate to the measure of interest (convergent validity) and others predicted to relate less well or not at all (discriminant validity). Convergent and discriminant validity operate together insofar as they contribute to construct validity (i.e., identifying what the construct is and is not like). A difficulty in many validational studies is attention only to convergent validity.

With discriminant validity, one looks for little or no relation between two or more measures. As with convergent validity, discriminant validity also raises interpretive issues. Two measures may have no conceptual connection or relation but still show significant and moderate-to-high correlation because of common method variance. If method variance plays a significant role, as is often the case when different informants are used, then all the measures completed by the same informant may show a similar level of correlation. In such a case, discriminant validity may be difficult to demonstrate.

Discriminant validity raises another issue for test validation. There is an amazing array of measures and constructs in the field of psychology, with new measures being developed regularly. The question in relation to discriminant validity is whether the measures are all different and whether they reflect different or sufficiently different constructs. The problem has been recognized for some time. For example, in validating a new test, Campbell (1960) recommended that the measure be correlated with measures of social desirability, intelligence, and acquiescence and other response sets. A minimal criterion for discriminant validation, Campbell proposed, is to show that the new measure cannot be accounted for by these other constructs. These other constructs, and no doubt additional ones, have been shown to have a pervasive influence across several domains, and their own construct validity is relatively well developed. It is likely that they contribute to and occasionally account for other new measures.

Few studies have adhered to Campbell's (1960) advice, albeit the recommendations remain quite sound. For example, a recent study validating the Sense of Coherence Scale showed that performance on the scale has a low and nonsignificant correlation with intelligence ($r = .11$) but a small-to-moderate correlation ($r = .39$) with social desirability (Frenz, Carey, & Jorgensen, 1993). Of course, convergent and discriminant validity depend on multiple sources of influence rather than two correlations. Even so, as the authors noted, the correlation with social desirability requires some explanation and conceptual elaboration.

General comments. Convergent and discriminant validity raise fundamental issues about validation efforts because they require specification of the nature of the construct and then tests to identify the connections and boundary conditions of the measure. Also, the two types of validity draw attention to patterns of correlations among measures in a given study and the basis of the correlation. The importance of separating or examining the influence of shared method factors that contribute to this correlation pattern motivated the recommendation to use multitrait and multimethod matrices in test validation (Campbell & Fiske, 1959). In general, demonstration of convergent and discriminant validity and evaluation of the impact of common method variance are critical to test validation. In the design and reporting of assessment studies, interpretation of the results very much depends on what can and cannot be said about the measure. The interpretation is greatly facilitated by

[4] Discriminant validity is used here in the sense originally proposed by Campbell and Fiske (1959). Occasionally, discriminant validity is used to refer to cases in which a measure can differentiate groups (e.g., Trull, 1991). The different meanings of the term and the derivation of related terms such as *discriminate, discriminative,* and *divergent validity* reflect a well-known paradox of the field, namely, that there is little reliability in discussing validity.

providing evidence for both convergent and discriminant validity.

Constructs and Measures

Assessment studies often vary in the extent to which they reflect interests in constructs or underlying characteristics of the measures and in specific assessment devices themselves. These emphases are a matter of degree, but worth distinguishing to convey the point and its implications for preparing and interpreting research reports. Usually researchers develop measures because they are interested in constructs (e.g., temperament, depression, or neuroticism). Even in cases in which measures are guided by immediately practical goals (e.g., screening and selection), there is an interest in the bases for the scale (i.e., the underlying constructs).

The focus on constructs is important to underscore. The emphasis on constructs draws attention to the need for multiple measures. Obviously, a self-report measure is important, but it is an incomplete sample of the construct. Perhaps less obvious is the fact that direct samples of behavior also are limited, because they are only a sample of the conditions as specified at a given time under the circumstances of the observations. Sometimes investigators do not wish to go beyond the measure or at least too much beyond the measure in relation to the inferences they draw. Self-report data on surveys (e.g., what people say about a social issue or political candidate or what therapists say they do in therapy with their clients) and direct observations of behavior (e.g., how parents interact with their children at home) may be the assessment focus. Even in these instances, the measure is used to represent broader domains (e.g., what people feel, think, or do) beyond the confines of the operational measure. In other words, the measure may still be a way of talking about a broader set of referents that is of interest besides test performance. Anytime an investigator wishes to say more than the specific items or contents of the measure, constructs are of interest.

Any one measure, however well established, samples only a part or facet of the construct of interest. This is the inherent nature of operational definitions. In preparing reports of assessment studies, the investigator ought to convey what constructs are underlying the study and present different assessment devices in relation to the sampling from the construct domain. A weakness of many studies is using a single measure to assess a central construct of interest. A single measure can sample a construct, but a demonstration is much better when multiple measures represent that construct.

The focus on constructs also draws attention to the interrelation among different constructs. Although a researcher may wish to validate a given measure and evaluate his or her operational definition, he or she also wants to progress up the ladder of abstraction to understand how the construct behaves and how the construct relates to other constructs. These are not separate lines of work, because an excellent strategy for validating a measure is to examine the measure in the context of other measures of that construct and measures of other constructs. For example, a recent study examined the construct psychological stress by administering 27 self-report measures and identifying a model to account for the measures using latent-variable analyses (Scheier & Newcomb, 1993). Nine latent factors were identified through confirmatory factor analyses (e.g., emotional distress, self-derogation, purpose in life, hostility, anxiety, and others). Of special interest is that the study permitted evaluation of several scales to each other as well as to the latent variable and the relation of latent variables (as second-order factors) to each other. This level of analysis provides important information about individual measures and contributes to the understanding of different but related domains of functioning and their interrelations to each other. At this higher level of abstraction, one can move from assessment to understanding the underpinnings of the constructs or domains of functioning (e.g., in development), their course, and the many ways in which they may be manifested.

Although all assessment studies might be said to reflect interest in constructs, clearly many focus more concretely at a lower level of abstraction. This is evident in studies that focus on the development of a particular scale, as reflected in evaluation of psychometric properties on which the scale depends. Efforts to elaborate basic features of the scale are critically important. Later in the development of the scale, one looks to a measure to serve new purposes or to sort individuals in ways that elaborate one's understanding of the construct. It is still risky to rely on a single measure of a construct no matter how well that validational research has been. Thus, studies using an IQ test or an objective personality inventory still raise issues if only one test is used, as highlighted later. For a given purpose (e.g., prediction), a particular measure may do very well. Ultimately, the goal is understanding in addition to prediction, and that requires greater concern with the construct and multiple measures that capture different facets of the construct.

In designing studies that emphasize particular measures, it is important to draw on theory and analyses of the underlying constructs as much as possible. From the standpoint of psychology, interest usually extends to the theory, construct, and clinical phenomena that the measure was designed to elaborate. Also, research that is based on a single assessment device occasionally is met with ambivalence. The ambivalence often results from the view that a study of one measure is technical in nature, crassly empirical, and theoretically bereft. The focus on a single measure without addressing the broader construct in different ways is a basis for these concerns. And, at the level of interpretation of the results, the reliance on one measure, however well standardized, may be viewed as a limitation.

At the same time, there is a widespread recognition that the field needs valid, standardized, and well-understood measures. Programs of research that do the necessary groundwork are often relied on when selecting a measure or when justifying its use in a study or grant proposal. When preparing articles on assessment devices, it is important to be sensitive to the implications that the study has for understanding human functioning in general, in addition to understanding how this particular measure operates. Relating the results of assessment studies to conceptual issues, rather than merely characterizing a single measure, can greatly enhance a manuscript and the reactions of consumers regarding the contribution.

Sample Characteristics and Assessment Results

Sampling can refer to many issues related to the participants, conditions of the investigation, and other domains to which one

wishes to generalize (Brunswik, 1955). In assessment studies, a special feature of sampling warrants comment because of its relevance for evaluating research reports. The issue pertains to the structure and meaning of a measure with respect to different population characteristics. Occasionally, the ways in which studies are framed suggest that the characteristics of a scale inhere in the measure in some fixed way, free from the sample to which the scale was applied.

It is quite possible that the measure and indeed the constructs that the measure assesses behave differently across samples, as a function of gender, age, race, and ethnicity (e.g., McDermott, 1995). Such differences have important implications for test standardization and interpretation beyond the scope of the present discussion. Sensitivity to such potential differences and evaluation of such differences in the design of research can be very helpful. Ideally, an assessment study will permit analyses of the influence of one or more sample characteristics that plausibly could influence conclusions about the measure. For example, in a recent evaluation of scales to study motives for drinking alcohol, analyses showed that the factor model that fit the measure was invariant across male and female, Black and White, and older and younger adolescents (Cooper, 1994). The inclusion of multiple samples and a sufficient sample size to permit these subsample analyses ($N > 2{,}000$) enabled the research to make a significant contribution to assessment and scale structure. From the study, it was learned that the structure of the measure is robust across samples. Apart from scale characteristics, the generality of the model may have important implications for adolescent functioning in general.

A more common research approach is to sift through separate studies, each representing an attempt to replicate the factor structure with a slightly different population (e.g., Derogatis & Cleary, 1977; Schwarzwald, Weisenberg, & Solomon, 1991; Takeuchi, Kuo, Kim, & Leaf, 1989). Such research often shows that the central features of the measure differ with different samples. One difficulty lies in bringing order to these sample differences, in large part because they are not tied to theoretical hypotheses about characteristics of the samples that might explain the differences (Betancourt & Lopez, 1993). Also, from the standpoint of subsequent research, guidelines for using the measure are difficult to cull from the available studies.

Evaluating assessment devices among samples with different characteristics is important. However, one critically important step before evaluating these assessment devices is the replication of the scale results with separate samples from the same population. Some studies include large standardization samples and hence provide within-sample replication opportunities. More common among assessment studies is the evaluation of the measure with smaller samples. It is important to replicate findings on the structure of the scale or the model used to account for the factors within the scale. Even when separate samples are drawn from the same population, the findings regarding scale characteristics may not be replicated (e.g., Parker, Endler, & Bagby, 1993). Evaluation of multiple samples is very important in guiding use of the measure in subsequent research.

Sampling extends beyond issues related to participants. Sampling refers to drawing from the range of characteristics or domains to which one wishes to generalize (Brunswik, 1955). In relation to assessment studies, the use of multiple measures to assess a construct is based in part on sampling considerations. Conclusions should not be limited to a single operation (measure or type of measure). There may be irrelevancies associated with any single measure that influences the obtained relation between the constructs of interest. A study is strengthened to the extent that it samples across different assessment methods and different sources of information.

The familiar finding of using multiple measures of a given construct is that the measures often reflect different conclusions. For example, two measures of family functioning may show that they are not very highly related to each other. One measure may show great differences between families selected because of a criterion variable, whereas the other measure may not. These results are often viewed as mixed or as partial support for an original hypothesis. The investigator usually has to prepare a good reason why different measures of seemingly similar constructs show different results. However, the study is stronger for the demonstration when compared with a study that did not operationalize family functioning in these different ways. An issue for the field is to make much further conceptual progress in handling different findings that follow from different methods of assessment.

Conclusion

Preparing reports for publication involves describing, explaining, and contextualizing the study. The descriptive feature of the study is essential for the usual goals such as facilitating interpretation and permitting replication of the procedures, at least in principle. However, the tasks of explaining the study by providing a well thought-out statement of the decisions and contextualizing the study by placing the demonstration into the field more generally are the challenges. The value of a study is derived from the author's ability to make the case that the study contributes to the literature, addresses an important issue, and generates important answers and questions.

In this article, I discussed some of the ways in which authors can make such a case when preparing a research article.[5] Generally, the task is to convey the theme or story line, bringing all of the sections of the study in line with that, and keeping irrelevancies to a minimum. In the context of assessment studies, three issues were highlighted because they affect many studies and their interpretation. These include interpretation of correlations between measures, the relation of constructs and measures, and sampling. Each issue was discussed from the standpoint of ways of strengthening research. Test validation, development of assessment methods from constructs, and sampling raise multiple substantive and methodological issues that affect both the planning and reporting of research. Many of the articles that follow elaborate on these issues.

[5] In closing, it is important to convey that recommendations in this article regarding manuscript preparation and journal publication derive from my experiences as an editor rather than as an author. As an author, the picture has not always been as pretty. For example, over the course of my career, such as it is, two journals went out of business within a few months after a manuscript of mine was accepted for publication and forwarded to production. Although this could be a coincidence in the career of one author, in this case the result was significant ($p < .05$), using a chi round test and correcting for continuity, sphericity, and leptokurtosis.

References

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: Author.

Betancourt, H., & Lopez, S. R. (1993). The study of culture, ethnicity, and race in American psychology. *American Psychologist, 48,* 629–637.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62,* 193–217.

Campbell, D. T. (1960). Recommendations for APA test standards regarding construct, trait, and discriminant validity. *American Psychologist, 15,* 546–553.

Campbell, D. T., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Cooper, M. L. (1994). Motivations for alcohol use among adolescents: Development and validation of a four-factor model. *Psychological Assessment, 6,* 117–128.

Derogatis, L. R., & Cleary, P. A. (1977). Factorial invariance across gender for the primary symptom dimensions of the SCL-90. *British Journal of Social and Clinical Psychology, 16,* 347–356.

Frenz, A. W., Carey, M. P., & Jorgensen, R. S. (1993). Psychometric evaluation of Antonovsky's Sense of Coherence Scale. *Psychological Assessment, 5,* 145–153.

Henry, B., Moffitt, T. E., Caspi, A., Langley, J., & Silva, P. A. (1994). On the "remembrance of things past": A longitudinal evaluation of the retrospective method. *Psychological Assessment, 6,* 92–101.

Kazdin, A. E. (1992). *Research design in clinical psychology* (2nd ed). Needham Heights, MA: Allyn & Bacon.

Kline, P. (1986). *A handbook of test construction: Introduction to psychometric design.* London: Methuen.

Maher, B. A. (1978). A reader's, writer's, and reviewer's guide to assessing research reports in clinical psychology. *Journal of Consulting and Clinical Psychology, 46,* 835–838.

McDermott, P. A. (1995). Sex, race, class, and other demographics as explanations for children's ability and adjustment: A national appraisal. *Journal of School Psychology, 33,* 75–91.

Parker, J. D. A., Endler, N. S., & Bagby, R. M. (1993). If it changes, it might be unstable: Examining the factor structure of the Ways of Coping Questionnaire. *Psychological Assessment, 5,* 361–368.

Robins, L. N., Schoenberg, S. P., Holmes, S. J., Ratcliff, K. S., Benham, A., & Works, J. (1985). Early home environment and retrospective recall. *American Journal of Orthopsychiatry, 55,* 27–41.

Scheier, L. M., & Newcomb, M. D. (1993), Multiple dimensions of affective and cognitive disturbance: Latent-variable models in a community sample. *Psychological Assessment, 5,* 230–234.

Schwarzwald, J., Weisenberg, M., & Solomon, Z. (1991). Factor invariance of SCL-90-R: The case of combat stress reaction. *Psychological Assessment, 3,* 385–390.

Shapiro, D. A., & Shapiro, D. (1983). Comparative therapy outcome research: Methodological implications of meta-analysis. *Journal of Consulting and Clinical Psychology, 51,* 42–53.

Takeuchi, D. T., Kuo, H., Kim, K., & Leaf, P. J. (1989). Psychiatric symptom dimensions among Asian Americans and native Hawaiians: An analysis of the symptom checklist. *Journal of Community Psychology, 17,* 319–329.

Trull, T. J. (1991). Discriminant validity of the MMPI–Borderline Personality Disorder scale. *Psychological Assessment, 3,* 232–238.

Wainer, H., & Braun, H. I. (Eds.). (1988). *Test validity.* Hilldale, NJ: Erlbaum.

Wampold, B. E., Davis, B., & Good, R. H., III (1990). Hypothesis validity of clinical research. *Journal of Consulting and Clinical Psychology, 58,* 360–367.

Weiss, B., & Weisz, J. R. (1990). The impact of methodological factors on child psychotherapy outcome research: A meta-analysis for researchers. *Journal of Abnormal Child Psychology, 18,* 639–670.

Received February 27, 1995
Revision received March 27, 1995
Accepted March 27, 1995 ■