Self-Report: Psychology's Four-Letter Word

GERALD J. HAEFFEL and GEORGE S. HOWARD
University of Notre Dame

Self-report continues to be one of the most widely used measurement strategies in psychology despite longstanding concerns about its validity and scientific rigor. In this article, the merits of self-report are examined from a philosophy of science perspective. A framework is also provided for evaluating self-report measures. Specifically, four issues are presented that can be used as a decision aid when making choices about measurement.

Self-report is among the most widely used measurement tools in psychology. It is also among the most criticized. In a typical critique, self-report is described as a major limitation, and it is made clear that more objective measures are needed to substantiate the results of the study. Noting the limitations of self-report is not, by itself, a problem. On the contrary, it is important to acknowledge the limitations of any measurement strategy; it may also be useful to include multiple measurement strategies. What is troubling is that researchers do not typically describe the limitations of self-report or how those limitations might affect the measurement of the theoretical constructs. Researchers rarely provide a rationale for why "objective measures" would be superior to self-report. The assumption seems to be that everyone already knows that self-report is inherently flawed (and that other measures would provide more valid data). Indeed, as Howard (1994) states, "It seems as if self-report bashing might be an article of faith of some Scientific Apostle's Creed—I believe in good science, the empirical determination of theory choice, the control of extraneous variables, and the fallibility of self-report measures" (p. 399).

Concerns about the validity and rigor of self-report measures are not new. The debate about self-report has been raging for more than 30 years and can be traced back to Nisbett and Wilson's (1977) seminal

Summer 2010, Vol. 123, No. 2 pp. 181–188 •

article, "Telling More Than We Can Know: Verbal Reports on Mental Processes." In their review, Nisbett and Wilson persuasively (and accurately) document how humans can be grossly inaccurate when reporting on their own cognitive processes. These criticisms appear to have only grown stronger over the years, particularly in light of recent measurement advancements in neuroimaging (e.g., functional magnetic resonance imaging), molecular genetics (e.g., DNA microarrays), and software for designing and conducting experiments (e.g., E-Prime; Schneider, Eschman, & Zuccolotto, 2002). A brief examination of the literature supports this notion. A comparison of journal articles published recently (2008 and 2009) with those published 10 years ago (1998 and 1999) illustrates how views of self-report have changed since the "decade of the brain." We examined four issues (two recent and two from 10 years ago) in each of the following journals: *Psychological Science, Journal of Abnormal Psychology, Journal of Personality and Social Psychology,* and *Child Development.* We found that 10 years ago (99 articles examined), 60% of published articles used self-report measures. However, this percentage was driven largely by the clinical and social psychology journals; 80% of articles in these journals used self-report measures. Of the articles using self-report, approximately 10% listed the use of self-report as a limitation. About 90% of these critiques noted that self-report was susceptible to demand characteristics and potential biases.

Similar results were found for studies published in recent journal issues. Approximately 60% of the studies reviewed (129 articles examined) used self-report measures. Again, this was driven in large part by work in clinical and social psychology, where approximately 90% of the research studies used self-report. Of the articles using self-report, approximately 20% listed self-report as a limitation. This is twice the percentage of that found for articles published 10 years ago. More importantly, the limitations of self-report are described differently. Ten years ago researchers focused on potential biases in self-report (e.g., demand characteristics). In contrast, today's researchers (about 70%) emphasize the need for future work to use behavioral or biological measures. Surprisingly, researchers tend to undermine their own results by arguing that future work should use behavioral or biological measures in order to validate their results. There is no explanation (or data) for why these alternative strategies are more valid than self-report. It is important to note that studies using behavioral or biological measures did not recommend the use of self-report measures to validate the results.

The view that behavioral or biological measures are more objective or scientific than self-report can also be found in research on measurement evaluation. It is easy to find review articles in which dozens of validity coefficients are offered as evidence of the invalidity of self-reports because of their lack of agreement measured with behavioral or biological measures (cf. Hook & Rosenshine, 1979; Rodin & Rodin, 1972). In such cases one never (at least in these authors' experience) finds the opposite situation, in which self-report measures are assumed to be the criterion and some non–self-report measure is validated by comparison with them. Because this background assumption seems to be almost universally present, it is interesting to speculate about the development of this bias.

We hypothesized that psychologists' early experience with research methods and measurement theories from undergraduate and graduate courses might contribute to the bias against self-report. We tested this hypothesis by examining 24 textbooks that are used in such courses. We investigated classic texts (by authors such as Anastasi, Cronbach, Cook, and Campbell) and texts written in the twenty-first century. Each book's index was searched for terms such as *criterion validity, construct validity, concurrent validity,* and *predictive validity*. Each lead from the index was followed up, and most revealed only a verbal description of the concept in question. In 15 instances the texts' authors also gave a concrete example of the concept, ranging over an array of constructs (e.g., assertiveness, self-esteem, anxiety, spelling, driving ability, job selection, intelligence). The next research decision involves which measure (the self-report or the non–self-report) will be selected as the criterion and which will be validated by reference to that criterion. In all 15 cases, the self-report was chosen as the measure to be validated. These early educational experiences may send an implicit message that self-report is less valid than other types of measures.

The data suggest that many researchers view self-report as inferior to other measurement strategies. Thus, it is unclear why researchers continue to use self-report. Perhaps it is because of tradition (i.e., inertia), ease of use, or cost effectiveness. Given the continued use of self-report, it is necessary to examine whether this measurement tool is detrimental to scientific progress. To this end, we consider the merits of self-report from a philosophy of science perspective. Next, a framework is provided for evaluating self-report measures. Specifically, four issues are presented that can be used as a decision aid when making choices about whether to use self-report.

Philosophy of Science: Is Self-Report Inherently Unscientific?
Donald Campbell (1974) once made a very insightful observation:

> All common-sense and scientific knowledge is presumptive. In any setting in which we seem to gain new knowledge, we do so at the expense of many presumptions, untestable—to say nothing of unconfirmable—in that situation. While the appropriateness of some presumptions can be probed singly or in small sets, this can only be done by assuming the correctness of the great bulk of other presumptions. Single presumptions or small subsets can in turn be probed, but the total set of presumptions [which] is not of demonstrable validity, is radically underjustified. (p. 5)

Campbell does not view common-sense knowledge as imperfect and scientific knowledge as perfect but rather sees them as stationed on a continuum anchored at one end by total skepticism or solipsism, in which we give up knowing or science, and anchored at the other end by total credulity. Ordinary knowing and science lie between these extremes and somehow combine a capacity for focused distrust and revision with a belief in the common body of knowledge claims (Campbell, 1974). It follows from this imperfect yet improvable view of scientific knowledge that the cumulative revision of scientific knowledge becomes possible through a process of "trusting (tentatively at least) the great bulk of current scientific and common-sense belief ('knowledge') and using it to discredit and revise one aspect of scientific belief. The ratio of the doubted to the trusted is always a very small fraction" (Campbell, 1974, p. 5). As the von Neurath metaphor, reported by Quine (1953), suggests, we are like sailors who repair a rotting ship while at sea. We place trust in the great bulk of timbers while a particularly weak timber is replaced. Each of the formerly trusted timbers may in turn be replaced. However, the ratio of timbers being replaced to those trusted as sound must, at any one time, be small.

Scientific progress is made by trusting the bulk of current knowledge in the form of implicit assumptions in our research efforts. For example, we trust that randomization produces its intended effect, that subjects will truthfully report their behavior, and that the theoretical variables of interest are reflected in the specific operational definitions used. These acts of faith are made in order to systematically examine the effects of a single or small set of variables, usually the independent or predictor variables. The results suggest the degree to which our prior beliefs concerning the independent variable or variables should be maintained or modified. The corpus of scientific knowledge changes and improves as new evidence supports or alters our beliefs.

However, one can consider the currently accepted body of research methods, strategies, and practices as a body of scientific knowledge about how one might go about seeking veridical knowledge (see Proctor & Capaldi, 2001, for an in-depth discussion of this position). Rather than viewing current practices as a given, we might subdivide them into a collection of assumptive stances that we have come to trust implicitly in order to obtain scientific knowledge. However, our trust in these assumptions need not remain implicit (Laudan, 1996). Each in its turn can be systematically doubted by making the current accepted practice one level of an independent variable in a study, while some alternative practice serves as another level of the independent variable. The dependent variable consists of some index of the validity or adequacy of the data obtained from each approach. If the incumbent procedure prevails, we can return it to the trust portion of our future studies with more explicit and justifiable confidence in its adequacy. Conversely, if the alternative approach proves superior, it can be incorporated into future research. This latter example is analogous to having replaced a particularly rotted timber in a ship while at sea in the von Neurath metaphor.

In practice, one never demonstrates that one methodological approach is always superior to another. An elaboration and extension of a parable by astronomer

Eddington might draw this point into sharp relief. Eddington tells of a scientist who wanted to catalog the fish in the sea (the research question). He took a net of 2-in. mesh (a research method) and cast it into the sea repeatedly. After carefully cataloging his findings, he concluded that there were no fish in the sea smaller than 2 in. In this apocryphal example, the scientist's trust in the adequacy of his method was somewhat misplaced and led the researcher to draw an inaccurate conclusion. However, if someone had doubted the adequacy of the netting procedure and performed an investigation specifically to test its adequacy relative to some specific alternative procedure, the misinterpretation might have been recognized. For example, our researcher might have considered an alternative research method: damming a small inlet of the sea, draining the water, and examining the bodies of the fish left behind. As fish smaller than 2 in. were found, the limitations of the netting procedure would become apparent. However, one would not be surprised to find that the largest fish obtained via the damming approach was substantially smaller than was obtained with the netting approach, signaling another potential problem. Therefore, research testing the adequacy of research methods does not prove which technique is better but provides evidence for the potential strengths and limitations of each. From this information, researchers can determine when one of two approaches, or both, should be the method of choice.

The purpose of the aforementioned analogies is to demonstrate that measurement techniques are not inherently scientific or unscientific. The validity of a measure depends on the context in which it is used; it is determined by evaluating whether the measure adequately assesses the construct of interest. Measures are simply tools that are used to test theories (Cronbach & Meehl, 1955). Indeed, science is characterized by its theories, not its measurement tools. According to Popper (1959), the defining feature of science is falsification. If a theory is falsifiable, it is by definition scientific. Popper's definition of science does not depend on the presence or absence of a particular measurement technique (e.g., self-report); the focus is solely on having falsifiable theories (see also Meehl, 1978).

From a philosophy of science perspective, self-report measures are not inherently flawed, unscientific, or inferior to more objective measures. They are just one of many possible measurement tools in a researcher's toolbox. However, this does not mean that they are infallible. As Eddington's fish example reminds us, all measures have strengths and limitations. Thus, it is important to carefully evaluate whether the construct to be measured matches the strengths of a particular measurement technique. We now present four issues to consider when determining whether to use a self-report measure.

Issue 1: Theory

Measurement decisions must be theory driven. Theories are needed to define the constructs to be measured and delineate how those constructs behave. This theoretical framework provides a basis for determining the construct validity of a measure and, in turn, a method for making measurement choices. Construct validity is concerned with the relationship between constructs within what Cronbach and Meehl (1955) called the nomological net. The nomological net is the system of hypothesized relationships that constitute a theory. A measure is said to have construct validity if it behaves as one would expect the construct it measures to behave according to the theory.

Construct validity can be a valuable metric for making measurement decisions. Consider the following example: A researcher is interested in identifying a measure of cognitive risk for depression. A quick review of the literature suggests a number of empirically supported measurement options such as the Cognitive Style Questionnaire (CSQ; Haeffel et al., 2008), Dysfunctional Attitude Scale (DAS; Weissman & Beck, 1978), and Response Styles Questionnaire (RSQ; Nolen-Hoeksema & Morrow, 1991). The question becomes how to choose between these measures, which purportedly all assess cognitive risk for depression. The answer, according to a construct validation approach, lies in the theory being tested. The theory specifies how a cognitive risk factor should behave (i.e., its nomological net). For example, a cognitive theory might state that cognitive risk should interact with stress to predict depression, predict event specific inferences for stressful life events, be elevated in females, be a specific predictor of depression and not other disorders such as alcoholism, and be stable over time. Using the nomological net as a framework, it is possible to choose the optimal measure—the one that best matches the predictions put forth by the theory.

This example underscores the need for well-articulated theories in psychology. It is difficult to evaluate the validity of our measurement tools without good theories. Unfortunately, good theories are the exception rather than the rule in psychology. Indeed, some researchers (e.g., Willner, 1985) have argued that most "theories" in psychology are not theories at all because they lack specificity. Psychological theories tend to identify an association between two constructs (e.g., dopamine and schizophrenia) but rarely specify mechanisms that can explain the association. Meehl (1978) argued that it is the lack of well-articulated theories in psychology that has led to its slow scientific progress. This means that the first step in making a measurement decision may be to strengthen the existing theory (e.g., increase its specificity). Once a well-articulated theory is in place, then the search for a measurement tool to test the theory can begin.

Issue 2: When Is Self-Report Valid?

Nisbett and Wilson's (1977) seminal article on the validity of self-report continues to have a tremendous influence on the field and is often the only citation needed when critiquing the validity of self-report measures. The Nisbett and Wilson article reviewed a large body of evidence demonstrating that people have little insight into their cognitive processes and are susceptible to demand characteristics. It is these limitations that most researchers associate with self-report measures. What many do not realize, however, is that the Nisbett and Wilson review was not an overarching condemnation of self-report. Nisbett and Wilson clearly acknowledged that there is at least some information on which humans can validly report:

> The individual knows a host of personal historical facts; he knows the focus of his attention at any given point in time; he knows what his current sensations are and has what almost all psychologists and philosophers would assert to be "knowledge" at least quantitatively superior to that of observers concerning his emotions, evaluations, and plans. (p. 255)

As this quotation highlights, it is critical to differentiate between cognitive processes and cognitive content. Although humans may not be capable of accurately reporting on inner processes, they are able to validly answer questions about a variety of constructs including their moods, attributions, plans, attitudes, and beliefs. In fact, Ericsson and Simon (1980) provide an extensive review illustrating that humans are capable of reporting on all data that can be brought into short-term memory.

Self-report may also be a valid measure of behavioral outcomes. Surprisingly, there is evidence that self-report may even be superior to behavioral measures in estimating behavioral outcomes. This idea flies in the face of both intuition and presumably hundreds of studies showing low correlations between self-report and behavioral measures of the same construct. These low correlations are typically interpreted as evidence against the validity of self-report. However, this conclusion is based on faulty logic because it assumes that the behavioral measure is a perfect measure of behavior (i.e., a criterion variable). This may not be the case, and therefore it is necessary to adopt a critical multiplist orientation wherein not two but rather six (for example) measures of the construct of interest are obtained. If measure #1 is a self-report and measure #2 is a behavioral measure, each of their validities should be ascertained by correlating them with an independent, multiply operationalized index of the construct made up of measures #3 through #6 (e.g., a role play measure, an expert's rating, an in vivo index, and a significant-others report). Dozens of such studies have been carried out (Cole, Howard, & Maxwell, 1981; Cole, Lazarick, & Howard, 1987; Gabbard, Howard, & Dunfee, 1986; Howard, Conway, & Maxwell, 1985; Howard, Maxwell, Wiener, Boynton, & Rooney, 1980), and surprisingly, the construct validity coefficients of the self-report measures were almost always superior to those of the behavioral measures. Thus, it appears that the thousands of low correlations between self-reports and behavioral measures should have told us more about the weakness of behavioral measures than about the weakness of self-reports.

Taken together, research demonstrates that self-report is a valuable (and valid) measure of cognitive products, plans, emotions, attitudes, and other constructs that are perceptual in nature. It can also be a valid indicator of behavior (Greco & Baenninger, 1989; Howard, 1994; Howard et al., 1980). The fact that self-report is cost-effective and easy to administer is an additional (rather than a primary) reason to use this measurement strategy.

Issue 3: Not All Self-Report Is Created Equal

Like any measurement tool, self-report has limitations that need to be considered before one adopts its use. Self-report is susceptible to demand characteristics and may not be valid for assessing cognitive processes. Therefore, it is necessary to evaluate the degree to which a given self-report measure may be vulnerable to these limitations.

Self-report measures differ in their level of transparency. This is important because demand characteristics are more likely to occur for measures that are transparent than those that are not. It may be possible to determine transparency by eyeballing the items on the measure, but a more empirical approach is recommended. For example, Schulman, Seligman, and Amsterdam (1987) conducted two studies to test the transparency of their cognitive risk measure, the Attributional Style Questionnaire. Participants were randomized to one of three conditions: an incentive group in which participants were told they would receive $100 if they could obtain the "best overall score," an incentive and coaching group in which participants received the $100 incentive and received coaching on what the test measured, and a no-incentive group. Results of the two studies showed that there were no significant differences between the three groups. These results provide evidence that this self-report measure is uninfluenced even when demand characteristics are changed (i.e., motivation to fake optimistic responses).

Another issue to consider when choosing a self-report measure is the degree of insight required from the participant. For example, as discussed earlier, the CSQ and DAS are both measures of cognitive risk for depression. However, these measures differ in the level of insight participants need to complete the questionnaire accurately. The CSQ, compared with the DAS, requires a much lower level of insight on the part of the participant. The CSQ simply asks questions about the implications of specific hypothetical events and does not directly ask whether the participant thinks he or she has a "cognitive risk" for depression. In essence, participants provide a cognitive sample on the CSQ that is thought to reveal their general cognitive style. Conversely, the DAS *does* require the participant to possess a high level of self-awareness. The DAS directly asks participants to make global judgments about themselves. For example, an item on the DAS reads, "My value as a person depends greatly on what others think of me." To rate accurately how much he or she agrees or disagrees with this statement, the participant must have insight into his or her own concept of self-worth. Moreover, as this example illustrates, items in the DAS are very general statements (unlike the specific events on the CSQ). Answering questions about general beliefs rather than specific events requires a greater degree of self-awareness on the part of the participant. Given that humans tend to be inaccurate about cognitive processes and judgments requiring insight, the measure that requires less self-awareness is often the best option.

As these examples illustrate, all self-report measures are not equal. These measures can differ in their level of transparency and the degree of insight required of the participant. When possible, self-report should avoid items that require a high level of insight. Similarly, if demand characteristics are likely, then the transparency of a measure is another issue to consider. Keeping these issues in mind may help to minimize some of the limitations of self-report.

Issue 4: When Validity May Not Matter

There may be instances in which the predictive power of a measure is more important than its validity. In these cases, it does not matter if the participant is grossly inaccurate in his or her self-report as long as it predicts meaningful outcomes. For example, a clinician may not care if a patient's self-reported level of social support (or stress) is valid as long it predicts therapeutic outcomes. In these cases, the patient's perception of his or her situation is more important than the objective reality of the situation (Cotton, 1980).

Self-report measures tend to be excellent predictors of moods, emotions, and psychopathology. Indeed, self-report measures often outperform other measurement techniques in terms of predictive power. Although biological factors are considered important in the etiology of both depression and schizophrenia, measures of these factors (e.g., functional magnetic resonance imaging, electroencephalography) rarely outperform self-report measures of psychological constructs in predicting future onset of psychopathology. For example, self-report measures of cognitive risk (e.g., CSQ and DAS) and social anhedonia (e.g., Chapman Scales; Chapman, Chapman, & Raulin, 1976) can be administered to nonclinical samples with no history of mental illness and used to predict first onsets of depression and schizophrenia spectrum disorders, respectively (Haeffel et al., 2008; Kwapil, 1998). We are unaware of a biological measure that has demonstrated this degree of predictive power.

In summary, there may be situations, particularly in applied settings, in which measurement is not tied to theory testing. In these cases, self-report can be a useful measurement tool because of its ability to predict important outcomes. Indeed, a person's inaccurate perception may often be a more robust predictor of future behavior than his or her objective reality. However, it is important to note that there is some danger in using a measure for purely predictive purposes. It is possible that data collected with these measures could be misused in theory evaluation because of their limited validity.

Conclusion

Researchers continue to use self-report measures at a high rate despite rising levels of criticism. Interestingly, the tone of the critiques has changed over the last decade. Today, researchers are much more likely to provide an unsophisticated critique that dismisses self-report as not being as valid as behavioral or biological measures (one out of seven studies using self-report measures contains this type of criticism). These critiques are presented in a theoretical vacuum and rarely explain why alternative measures would be better options than self-report.

To address these critiques, we examined self-report from a philosophy of science perspective to determine whether its use is detrimental to scientific progress. We concluded that self-report measures are not inherently inferior to behavioral or biological measures. This is because the validity of a measure depends on the theoretical context in which it is used; measures are simply tools that are used to test theories (Cronbach & Meehl, 1955). The integrity of the tool depends on how well it does its job (i.e., does it measure the construct of interest?). Research suggests that self-report is well suited for assessing a number of theoretical constructs including cognitive products (e.g., attributions, plans, attitudes, and beliefs), emotions, and moods. Moreover, self-report may be a valid indicator of behavior.

Self-report can be a valid and useful tool in psychological research. It does not have to be a four-letter word. We believe that it is time for scientists to define themselves by their theories rather than the measurement tools they use. When evaluating self-report measures, we need to move beyond our knee-jerk reactions and think critically about whether the tool exhibits theoretical and measurement fidelity. It is by building more specific theories, and not by eliminating self-reports, that we may accelerate the slow progress of soft psychology (Meehl, 1978).

NOTE

Address correspondence about this article to Gerald J. Haeffel, 108 Haggar Hall, University of Notre Dame, Notre Dame, IN 46656 (e-mail: ghaeffel@nd.edu).

REFERENCES

Campbell, D. T. (1974). *Qualitative knowing in action research*. Kurt Lewin Award Address, American Psychological Association Convention, New Orleans.

Chapman, L. J., Chapman, J. P., & Raulin, M. L. (1976). Scales for physical and social anhedonia. *Journal of Abnormal Psychology, 85,* 374–382.

Cole, D. A., Howard, G. S., & Maxwell, S. E. (1981). The effects of mono versus multiple operationalization in construct validation efforts. *Journal of Consulting and Clinical Psychology, 49,* 395–405.

Cole, D. A., Lazarick, D. M., & Howard, G. S. (1987). Construct validity and the relation between depression and social skill. *Journal of Counseling Psychology, 34,* 315–321.

Cotton, J. L. (1980). Verbal reports on mental processes: Ignoring data for the sake of the theory. *Personality and Social Psychology Bulletin, 6,* 278.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87,* 215–251.

Gabbard, C. E., Howard, G. S., & Dunfee, E. J. (1986). Reliability, sensitivity to measuring change, and construct validity of therapist adaptability. *Journal of Counseling Psychology, 33,* 377–386.

Greco, M., & Baenninger, R. (1989). Self-report as a valid measure of yawning in the laboratory. *Bulletin of Psychonomic Society, 27,* 75–76.

Haeffel, G. J., Gibb, B. E., Abramson, L. Y., Alloy, L. B., Metalsky, G. I., Joiner, T., et al. (2008). Measuring cognitive vulnerability to depression: Development and validation of the Cognitive Style Questionnaire. *Clinical Psychology Review, 28,* 824–836.

Hook, C. M., & Rosenshine, B. V. (1979). Accuracy of teacher reports of their classroom behavior. *Review of Educational Research, 49,* 1–12.
Howard, G. S. (1994). Why do people say nasty things about self-reports? *Journal of Organizational Behavior, 15,* 399–404.
Howard, G. S., Conway, C. G., & Maxwell, S. E. (1985). Construct validity of measures of college teaching effectiveness. *Journal of Educational Psychology, 77,* 187–196.
Howard, G. S., Maxwell, S. E., Wiener, R. L., Boynton, K. S., & Rooney, W. M. (1980). Is a behavioral measure the best estimate of behavioral parameters? Perhaps not. *Applied Psychological Measurement, 4,* 293–311.
Kwapil, T. R. (1998). Social anhedonia as a predictor of the development of schizophrenia-spectrum disorders. *Journal of Abnormal, 107,* 558–565.
Laudan, L. (1996). *Beyond positivism and relativism: Theory, method, and evidence*. Boulder, CO: Westview.
Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.
Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231–259.
Nolen-Hoeksema, S., & Morrow, J. (1991). A prospective study of depression and posttraumatic stress symptoms after a natural disaster: The 1989 Loma Prieta earthquake. *Journal of Personality and Social Psychology, 61,* 115–121.
Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson.
Proctor, R. W., & Capaldi, E. J. (2001). Empirical evaluation and justification of methodologies in psychological science. *Psychological Bulletin, 127,* 759–772.
Quine, W. V. O. (1953). *From a logical point of view.* Cambridge, MA: Harvard University Press.
Rodin, M., & Rodin, B. (1972). Student evaluations of teachers. *Science, 177,* 1164–1166.
Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime user's guide*. Pittsburgh: Psychology Software Tools Inc.
Schulman, P., Seligman, M. E., & Amsterdam, D. (1987). The attributional style questionnaire is not transparent. *Behaviour Research and Therapy, 25,* 391–395.
Weissman, A., & Beck, A. T. (1978, November). *Development and validation of the Dysfunctional Attitudes Scale: A preliminary analysis*. Paper presented at the meeting of the American Educational Research Association, Toronto, Ontario, Canada.
Willner, P. (1985). On the nature of psychological explanations. In P. Willner (Ed.), *Depression: A psychobiological synthesis* (pp. 529–541). New York: Wiley.