To prosper, organizational psychology should … overcome methodological barriers to progress

JEFFREY R. EDWARDS*

Kenan-Flagler Business School, University of North Carolina, Chapel Hill, North Carolina, U.S.A.

Summary   Progress in organizational psychology (OP) research depends on the rigor and quality of the methods we use. This paper identifies ten methodological barriers to progress and offers suggestions for overcoming the barriers, in part or whole. The barriers address how we derive hypotheses from theories, the nature and scope of the questions we pursue in our studies, the ways we address causality, the manner in which we draw samples and measure constructs, and how we conduct statistical tests and draw inferences from our research. The paper concludes with recommendations for integrating research methods into our ongoing development goals as scholars and framing methods as tools that help us achieve shared objectives in our field. Copyright © 2008 John Wiley & Sons, Ltd.

Introduction

Scientific progress rests on the interplay between theory and empiricism, whereby theories are used to derive hypotheses that are investigated empirically, and the results of these investigations are used to corroborate or refute aspects of the theory and perhaps stimulate new ideas for theory development (Runkel & McGrath, 1972). In principle, this interplay is a smooth process in which knowledge progressively accumulates over time, although in practice the process often proceeds in fits and starts (Kuhn, 1970; Pfeffer, 1993). In either case, scientific progress depends on the quality of methods used in empirical research, as these methods form the epistemological bridge that links theory to data and constitute the tools used to evaluate evidence for and against theories.

Because methods are essential to scientific progress, they are worth examining in their own right. In this paper, I consider the methods used in organizational psychology (OP) research, pointing out methodological barriers to progress and offering suggestions to help researchers confront and, ideally, overcome these barriers. Although my focus is on OP research, the barriers discussed here are also relevant to research in allied fields, such as organizational behavior and human resource management. Many of the barriers have complex philosophical and methodological underpinnings, and some have triggered debates that continue to rage. Given this context, my objectives are restricted to articulating

* Correspondence to: Jeffrey R. Edwards, Kenan-Flagler Business School, University of North Carolina, Chapel Hill, NC 27599-3490, U.S.A. E-mail: jredwards@unc.edu

the barriers chosen for discussion, giving suggestions as warranted, and pointing to relevant sources for elaboration. I should add that I do not intend to malign the field, nor have I singled out researchers whose work illustrates a particular barrier, as I believe we all encounter these barriers in our work. My goal is to be constructive rather than critical, and I approach the barriers with guarded optimism that many of them can be ameliorated, if not surmounted.

Methodological Barriers to Progress

For reasons of scope, the following discussion is limited to ten methodological barriers. These barriers were chosen to span different stages of the research process and to reflect issues that have attracted attention in the methodological literature and regularly surface in empirical research. The barriers concern how theories are translated into hypotheses, how conditions for causality are built into research design, how samples are drawn and measures are collected and evaluated, and how hypotheses are tested and statistical inferences are drawn. Although the problems created by each barrier are considered, the overall discussion of the barriers is framed in terms of solutions rather than problems, as my intent is to encourage rather than discourage attempts to contend with the barriers. I conclude with some observations concerning the future of methodology in our field.

Expose theories to risky tests

In OP research, theories are usually translated into directional hypotheses that predict a positive or negative monotonic relationship, such that an increase in one variable will be associated with an increase or decrease in another variable. Directional hypotheses are certainly stronger than nondirectional hypotheses, which merely predict that two variables relate without specifying whether the relationship is positive or negative. Nonetheless, directional hypotheses put theories at little risk, because the range of parameter values that constitute support for the hypothesis is quite large (Meehl, 1967). For example, assume a theory predicts a positive correlation between organizational commitment and intent to stay, and this hypothesis is tested using a sample of 170 cases and a p-value of .05. Under these conditions, any correlation greater than .15 would be taken as support for the hypothesis, limited only by values sufficiently high to question the discriminant validity of the measures involved. Hypotheses that accommodate such a broad range of values confront low hurdles of acceptance, and as a result, the theories that such hypotheses are intended to test bear little risk of falsification.

The scarcity of tests that challenge theories is compounded by the overrepresentation in the published literature of studies that yield statistically significant results. This imbalance can be traced to the tendency of authors to withhold studies that produce null results, creating what has been called the "file drawer problem" (Rosenthal, 1979), and publication biases that favor statistically significant results over null results (Hubbard, 1997; Mahoney, 1985; Sterling, Rosenbaum, & Weinkam, 1995). Both of these factors weed out results that contradict what theories predict, exaggerating the apparent support for theories among studies that are ultimately published.

We can expose OP theories to riskier tests by adopting several strategies. First, we can attempt to translate theories into predictions that are more precise than directional relationships. Clearly, OP theories cannot be expected to generate point predictions, as in hard sciences such as physics. However, this limitation does not rule out the possibility of predicting a range of values (Meehl, 1990).

For instance, hypotheses could predict whether the expected relationship should fall within ranges that represent small, medium, or large effects, with boundaries of these ranges derived from meta-analyses of relevant substantive research. Alternately, directional predictions could specify the functional form of the relationship, such as whether a positive relationship increases linearly, diminishes at higher levels, or turns downward at some critical point. Directional predictions are also made more precise by specifying moderator variables that influence the predicted relationship, altering its magnitude and perhaps changing its sign.

Second, hypothesis can be made more precise by unpacking relationships embedded in measures intended to represent complex theoretical entities. For example, concepts such as congruence, diversity, and heterogeneity are often operationalized using indices that impose constraints on relationships involving measures that constitute the index (Edwards, 1994). These constraints effectively represent hypotheses that are encased within the index, shielded from potential falsification. These hypotheses can be tested by determining whether the constraints are supported or refuted by the data, exposing the theory to risks it would otherwise evade.

Third, as authors, reviewers, and journal editors, we should adopt a more open stance to studies with results that run counter to what theories predict. This recommendation does not mean that we should uncritically accept null results, particularly when they come from studies with weak designs, unreliable measures, or low statistical power (Cortina & Folger, 1998). Rather, we should evaluate all studies against rigorous methodological standards, and studies that meet these standards should not be considered unworthy of publication because their results contradict prior theory. Instead, contradictory results should be interpreted as falsifying evidence that refines or perhaps rejects a theory. Such results are valuable for advancing our science and do not merit the reaction that the study did not "work out." Studies that give valid answers to research questions are worthwhile regardless of whether the answers are affirmative or negative.

Conduct strong inference studies

Typically, studies in OP research are framed around a single theory or a set of theories that yield similar predictions. Studies based on a single theory usually derive hypotheses from that theory and infer support for the theory when empirical results conform to predictions (e.g., parameter estimates differ from zero in the expected direction). Studies that draw from multiple theories often use them to strengthen the arguments in favor of a hypothesis, selecting theories that are consistent with one another and lead to the same prediction. Evidence that conforms to the hypothesis is taken as support for any or all of the theories used to derive the hypothesis.

Studies designed to confirm one or more theories can be contrasted with those that pit theories against one another, with the goal of achieving *strong inference* (Platt, 1964). Strong inference consists of deriving alternative hypotheses and designing studies that can produce multiple outcomes, each of which rules out one or more of the alternative hypotheses in part or whole. Studies that rely on strong inference expose theories to risky tests and accelerate scientific advancement, given that results that favor one hypothesis necessarily preclude other hypotheses. Platt (1964) likens the progress yielded by strong inference to climbing a tree, with each study representing a fork that prompts us to follow one branch to the exclusion of others. Strong inference also fosters the discipline of thinking in terms of multiple hypotheses and asking ourselves what evidence would disprove each hypothesis in favor of another (Popper, 1959).

Strong inference research can be approached in various ways. One approach, as outlined by Platt (1964), is to deliberately seek theories that yield contradictory hypotheses and devise studies that provide critical tests. Although simple in principle, this approach requires us to overcome the tendency

to advocate a particular theory and instead adopt an agnostic stance, such that the outcomes of critical tests are viewed as resolutions to conflicting ideas as opposed to threats to our prior beliefs or preferred theory. Examples of this approach can be found in research on social networks and team effectiveness (Balkundi & Harrison, 2006), turnover and organizational performance (Shaw, Gupta, & Delery, 2005), learning negotiation skills (Nadler, Thompson, & Van Boven, 2003), the role of gender in performance evaluations (Aguinis & Adams, 1998), and the person–environment fit approach to stress (Edwards, 1996).

Another approach is to conduct joint studies with researchers who hold opposing views, leading to what Kahneman (2003) calls *adversarial collaboration*. This approach requires the involved parties to place priority on resolving the dispute at hand and accept the possibility that their espoused position might be rejected. Adversarial collaboration can be facilitated by including a third-party colleague who serves as a mediator and helps ensure that the opposing views are compared fairly (Cooper & Richardson, 1986). Adversarial collaborations have been staged in research on goal-setting (Latham, Erez, & Locke, 1988) and conjunctive effects in judgment (Mellers, Hertwig, & Kahneman, 2001).

Address conditions for causality

Relationships among constructs that constitute OP theories are usually stated in causal terms, such that one construct directly or indirectly affects another construct. By implication, hypotheses derived from OP theories refer to causal relationships, and tests of hypotheses are incomplete unless the causal assumptions underlying the hypotheses are evaluated empirically. Although conditions for causality continue to develop and stimulate debate (Collins, Hall, & Paul, 2004; Pearl, 2000; Salmon, 1998; Sosa & Tooley, 1993; Spirtes, Glymour, & Scheines, 2000), organizational and social scientists generally agree that causality requires that the cause and effect are correlated, the cause precedes the effect in time, and alternative explanations for the presumed causal relationship (e.g., spurious correlation) can be ruled out (Cook & Campbell, 1979; James, Mulaik, & Brett, 1982).

The foregoing conditions for causality are rarely satisfied in OP research. Most OP studies rely on cross-sectional designs (Medsker, Williams, & Holahan, 1994; Mitchell, 1985; Scandura & Williams, 2000), which cannot establish that the cause occurred before the effect and are susceptible to numerous interpretations other than the causal effect of interest (Cook & Campbell, 1979). Cross-sectional studies also tend to produce biased estimates of the relation between the cause and effect, given that the correlation between measures collected at the same time generally differs from the correlation between measures separated by the time interval that corresponds to the causal effect (Maxwell & Cole, 2007; Mitchell & James, 2001). This bias can be ameliorated with longitudinal designs, provided the cause and effect are measured with the appropriate time lag and the model is otherwise correctly specified (James et al., 1982; Mitchell & James, 2001). Regrettably, longitudinal studies in OP research often measure the cause before the effect but omit prior measures of the effect and subsequent measures of the cause (Mitchell & James, 2001). Such designs are weak for establishing temporal precedence because restricting measures to the presumed causal order does not rule out other causal orders that could operate during the course of the study.

To illustrate the limitations of longitudinal studies that merely measure the cause before the effect, consider the design shown in Figure 1a, where X_1 and Y_2 are the presumed cause and effect measured at times 1 and 2, respectively. Compare this design to that shown in Figure 1b, where X and Y are measured at both times. The design in Figure 1b can detect a reverse causal effect of Y_1 on X_2, which could equal or exceed the effect of X_1 on Y_2. Now consider the design in Figure 1c, which adds the lagged measures X_0 and Y_0. This design is equipped to determine whether the path relating X_1 to Y_2 is spurious due to Y_0 as a common cause of X_1 and Y_2 through Y_1. These examples show that merely measuring X before Y
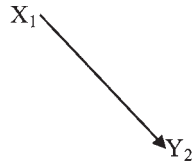
a. X and Y measured on two separate occasions.
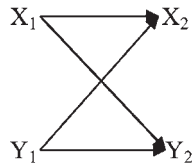
b. X and Y both measured on two occasions
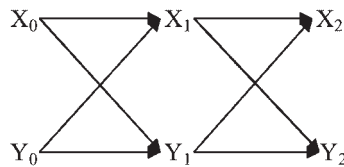
c. X and Y both measured on three occasions

Figure 1. Models depicting a cause and an effect measured at various time intervals

offers little assurance that X is a cause of Y. Longitudinal designs are susceptible to other threats to internal validity, such as history, maturation, attrition, and selection, which further undermine the interpretation of the relationship between X and Y as causal (Cook & Campbell, 1979). These threats to validity are largely avoided when longitudinal designs are accompanied by random assignment of cases to levels of X, features that characterize true experiments.

Causal inferences afforded by OP research can be strengthened in several ways. Perhaps the most obvious approach is to conduct experimental research, the merits of which have been discussed for decades (Cook, Campbell, & Perrachio, 1990; Fromkin & Streufert, 1976; Ilgen, 1986; Weick, 1965). Despite its advantages, experimentation is apparently on the decline in OP research (Scandura & Williams, 2000), which underscores the importance of determining the degree to which nonexperimental studies justify causal inferences (Cook & Campbell, 1979). Our thinking on these matters can be sharpened by clarifying what we mean by causality. One useful perspective on causality frames it in counterfactual terms, such that when we think about causality, we are imagining what would have been the outcome if a case (e.g., person, group) had been at a different level on a causal variable (Collins et al., 2004; Holland, 1986; Rubin, 1978). The counterfactual perspective highlights a fundamental dilemma in causal inference, which is that we can *never* observe the same case at different

levels of a causal variable at a given time. In practice, we attempt to address this dilemma by observing different cases across levels of the causal variable, as when cases are assigned to treatment and control groups or occupy different levels of a measured variable. In general, cases can differ for reasons other than their standing on the causal variable, which undermines our ability to ascribe variation in an outcome solely to the causal variable, as opposed to other variables that correlate with the causal variable and the outcome. This type of confounding is avoided when cases are randomly assigned to levels of the causal variable, which justifies the assumption that cases are equal in all respects other than the level of the causal variable. This confounding can also be ameliorated by matching cases on variables that represent alternative explanations for the causal effect or statistically controlling for such variables (Cook & Campbell, 1979; Rubin, 1974). Clearly, these procedures are inferior to randomization, but they should be employed when randomization is impossible or impractical. Combining these procedures with longitudinal panel designs that incorporate appropriate time lags can strengthen our ability to draw causal inferences from nonexperimental studies and would improve upon cross-sectional and longitudinal designs typical of OP research.

Link samples to populations

As a field, OP research relies heavily on convenience samples, whose members are studied primarily because they are accessible (Mitchell, 1985). The widespread use of convenience samples reflects practical constraints that inhibit other sampling strategies, most notably random sampling (Henry, 1990). Random sampling requires the researcher to specify a population from which cases will be selected and to which inferences will be drawn. Specifying the population brings into focus basic issues relevant to most studies such as whether the effects predicted by the theory under investigation vary across groups or cultures, how well the obtained sample corresponds to the target population, and to whom the results of the study are intended to generalize. Indeed, the use of inferential statistics rests on the premise that quantities computed from sample data (e.g., group means, regression coefficients) are intended not to describe the sample itself, but instead to estimate parameters in some population. Naturally, this exercise is meaningful only if we have a particular population in mind.

In OP research, the populations we effectively target with inferential statistics are rarely identified in explicit terms. In laboratory research, we typically rely on college students and acknowledge that the obtained results might not generalize to working adults, implying that the target population might be broadly defined as ''people who work.'' In survey research, we generally administer questionnaires to members of one or more organizations and subsequently compare respondents to nonrespondents by analyzing mean differences on descriptive variables (e.g., demographics, job type, and years of experience). These analyses document the similarity between the initial sample and final sample, but they say nothing about the similarity between either sample and some broader population that presumably represents the target of statistical inference. Moreover, mean differences on descriptive variables do not capture differences in relationships among substantive variables. These relationships are the focus of the studies we conduct, and concerns about generalizability should address whether these relationships differ between the final sample and the target population, not whether means on descriptive variables differ between respondents and nonrespondents. Perhaps we analyze differences on descriptive variables because the data are available, but these analyses answer the wrong question if our goal is to establish the generalizability of our substantive results.

I am not recommending that we stop comparing respondents to nonrespondents on available measures. If nothing more, such comparisons help us understand who was in the final sample and how they are compared to members of the initial sample and perhaps members of some broader organizational setting. Rather, we should recognize that such comparisons yield limited information,

and we should seek additional procedures that can help us assess whether our results generalize beyond the data in hand and our statistical inferences are legitimate. For instance, if we find differences on descriptive variables and suspect that such differences could translate into differences in relationships among substantive variables, then we can differentially weight cases in the final sample to parallel the distribution of the descriptive variables in the initial sample (Winship & Radbill, 1994). Comparing results from the weighted and unweighted analyses would indicate whether differences on the descriptive variables are relevant to the substantive conclusions of the study. Weights could also be used that mimic the distributions of descriptive variables in the target population, perhaps based on census figures or survey data from other sources. Alternately, variables that might differentiate the final sample from the population could be measured and tested in the final sample as moderators of key substantive relationships (Mitchell, 1985). The distributions of the moderator variables in the final sample might be more restricted than in the target population, so results from the moderator analyses should be taken as suggestive rather than definitive. Other procedures for addressing sampling biases can be applied to bolster population inferences drawn from the sample used for analysis (Rosenbaum, 1995; Wainer, 1986).

Scrutinize self-reports

Throughout its history, OP research has relied heavily on self-reports as measures of substantive constructs (Mitchell, 1985; Podsakoff & Organ, 1986; Sackett & Larson, 1990; Spector, 1994). The widespread use of self-reports is understandable, given that much OP research concerns the perceptions, attitudes, and feelings of people in organizations, and a natural way to assess such constructs is to ask people about them. Nonetheless, our faith in self-reports is based on the premise that respondents interpret our questions as intended, know and can retrieve the information we seek, and integrate and translate the information into a suitable response (Tourangeau, Rips, & Rasinski, 2000). These psychological processes are influenced by a host of factors that can markedly influence the answers we obtain to our questions. Such factors include the framing and wording of individual questions, the context and order of sets of questions, the type, structure, and appearance of response alternatives, the conversational logic applied by respondents, the mood and engagement of the respondent at the time of measurement, the stated purpose of the study, and even the perceived credibility of the researcher (Schuman & Presser, 1981; Schwarz, 1999; Sudman, Bradburn, & Schwarz, 1996; Tourangeau & Rasinski, 1988; Tourangeau et al., 2000).

   Without question, factors that influence self-reports have been acknowledged in the OP literature (Feldman & Lynch, 1988; Harrison & McLaughlin, 1993, 1996; Harrison, McLaughlin, & Coalter, 1996; Podsakoff & Organ, 1986). For the most part, those among us who use self-reports admit their limitations, and we sometimes attempt to collect data from sources other than the focal respondent and encourage others to do the same. For some constructs, we might obtain measures that avoid factors that influence self-reports, as when absenteeism is measured using attendance records. However, measures recommended as alternatives to self-reports are often nothing more than self-reports from other sources, such as senior managers, supervisors, peers, or subordinates. Such measures might help address concerns about common method variance, as discussed below, but they depend on the same psychological processes that influence self-reports obtained from the focal respondent. Furthermore, many of the constructs we wish to measure reside in the mind of the focal respondent, and for such constructs, self-reports are arguably the best source of data (Spector, 1994). For instance, if we want to assess how satisfied a respondent is with his or her job, it is doubtful that anyone other than the respondent could give a legitimate response. Hence, we are faced with a conundrum, in that we are

compelled to use self-reports, yet we know that they can be significantly swayed by factors other than the constructs we wish to assess.

In light of this conundrum, perhaps the best recourse is to enhance our understanding of the cognitive processes that generate self-reports (Stone, Turkkan, Bachrach, Jobe, Kurtzman, & Cain, 2000; Sudman et al., 1996; Tourangeau et al., 2000). For instance, we can collect concurrent verbal protocols in which respondents are asked to think aloud while they read, process, and answer questions (DeMaio & Rothgeb, 1996; Ericsson & Simon, 1993; Harrison et al., 1996). These protocols could be gathered when measures are pretested, supplementing the statistics used to evaluate measures with qualitative information that sheds light on how questions are interpreted and answers are generated. This approach requires that people can describe their ongoing mental processes (Nisbett & Wilson, 1977), and procedures that facilitate this task are available (Ericsson & Simon, 1993; White, 1980). We can also adopt measurement strategies in which respondents do not explicitly answer questions but instead make choices or engage in behaviors that implicitly reveal answers to questions. Possibilities include conditional reasoning tests (James, 1998; LeBreton, Barksdale, Robin, & James, 2007), response latencies (Haines & Summer, 2006), clinical judgments (Taber, 1991), implicit association tests (Greenwald, McGhee, & Schwartz, 1998), experience sampling (Alliger & Williams, 1993; Hektner, Schmidt, & Csikszentmihalyi, 2007; Stone, Shiffman, & DeVries, 1999), and diaries (Bolger, Davis, & Rafaeli, 2003). More generally, we can view the psychological processes that influence self-reports not as a nuisance, but as part and parcel of the phenomena we investigate (Schwarz, 2007).

Reconsider common method variance

Common method variance is a perennial concern in OP research (Conway, 2002; Crampton & Wagner, 1994; Doty & Glick, 1998; Podsakoff, MacKenzie, Lee, & Podsakoff, 2003; Spector, 2006; Williams & Anderson, 1994; Williams & Brown, 1994). Common method variance refers to variation in measures of different constructs that results from sharing the same method of measurement. Although often associated with self-report measures, common method variance is relevant whenever different constructs are measured with the same method, whether that method involves reports from supervisors, scores from trained observers, clinical interviews, company records, and so forth (Conway, 2002; Spector, 2006). Sources of common method variance include response tendencies that raters apply across measures, similarities in item structure or wording that induce similar responses, the proximity of items in an instrument, and similarities in the medium, timing, or location in which measures are collected (Podsakoff et al., 2003).

Common method variance is problematic because, when measures of different constructs use the same method, the relationship between the measures can be influenced by factors other than the relationship between the constructs themselves. As such, common method variance can bias estimates of the relationships between constructs, which in turn can impact inferences drawn from empirical research. However, the nature and degree of bias created by common method variance is more complex than often believed. It is generally assumed that common method variance inflates the relationships among measures, such that the actual relationships among the underlying constructs are smaller than they appear (Spector, 2006). However, common method variance can inflate or deflate relationships among measures, depending on how method factors enter into the model that describes the measures and the size and direction of the parameters in the model (Conway, 2002; Spector, 2006; Williams & Brown, 1994). Moreover, bias due to common method variance is a matter of degree and depends on the substantive nature of the constructs assessed by the measures (Crampton & Wagner, 1994; Doty & Glick, 1998; Spector, 2006; Williams, Cote, & Buckley, 1989). Hence, although common method variance is a legitimate concern, we should not automatically dismiss studies that collect data using a

single method (Spector, 2006). Rather, we should address the effects of common method variance on a study-by-study basis to determine whether and how they might influence the substantive conclusions drawn from our research.Common method variance can be addressed through the design of a measure, the manner in which it is administered, and by using statistical procedures after the data have been collected (Podsakoff et al., 2003; Spector, 2006). For instance, a survey can be divided into subsections that are administered at separate times and locations or using different formats. Alternately, measures can be collected from sources other than the focal respondent, such as subordinates, peers, and supervisors. Various statistical approaches for addressing common method variance have also been proposed (Podsakoff et al., 2003). One approach involves the use of a general factor intended to represent the source of common method variance. Variations of this approach include controlling for a general factor score computed from the first principal component of the measures (Podsakoff & Todor, 1985), introducing a latent variable as a common method factor to which all measures are assigned (Widaman, 1985), and loading subsets of measures onto latent variables that represent different method factors, as illustrated by confirmatory factor analyses of multitrait–multimethod matrices (Marsh & Bailey, 1991; Schmitt & Stults, 1986; Williams et al., 1989). These procedures yield estimates of relationships among substantive constructs after controlling for the specified method factors and can quantify the proportion of variance in measures due to the method factors. However, controlling for a general factor can remove substantive as well as method variance (Kemery & Dunlap, 1986; Marsh, 1989), and the interpretation of the general factor is ambiguous in that it represents all sources of covariance among the measures other than their assigned substantive factors. An alternative approach is to directly measure factors thought to induce common method variance, such as social desirability, trait affectivity, or general impressions of others, and incorporate these measures into models that contain the substantive variables of interest (Conway, 1998; Williams & Anderson, 1994; Williams, Gavin, & Williams, 1996). This approach makes the meaning of method factors explicit and avoids identification and estimation problems that occur when method factors do not have their own measures (Kenny & Kashy, 1992; Millsap, 1992). Of course, this approach only takes into account the method factors that are measured, but this limitation is offset by the clarity provided by the approach and the foundation it establishes for developing and testing theories that explain the effects of method factors from a conceptual standpoint (Schmitt, 1994).

Address the effects of measurement error

The importance of measurement error has long been recognized in OP research (Campbell, 1976; DeShon, 1998; Edwards, 2003; Schmidt & Hunter, 1996; Schwab, 1980). Concerns over measurement error are justified, given that measurement error undermines the interpretation of measures of theoretical constructs and violates assumptions of commonly used statistical procedures, such as ordinary least squares (OLS) regression (Berry, 1993). In empirical OP research, measurement error is usually assessed with Cronbach's α (Cronbach, 1951), which is routinely reported in published work (Schmitt, 1996). When measures fail to reach some standard for adequate reliability, such as .70 (Nunnally, 1978), we typically exclude the measure from further analysis or warn our readers that results for the measure should be interpreted with caution. If measures demonstrate what we consider adequate reliability, we often proceed with analyses without further concern over the effects of measurement error on our results (Schmitt, 1996).

Handling measurement error in this manner is problematic on several counts. One set of problems arises from the widespread use of α, which relies on various assumptions that are rarely considered in practice. First, α requires that items are τ-equivalent, meaning they have equal loadings on the factor that underlies the measure (Novick & Lewis, 1967). When this assumption is violated, α

underestimates the reliability of the measure (Komaroff, 1997; Raykov, 1997; Zimmerman, Zumbo, & Lalonde, 1993). Second, α rests on the assumption that measurement errors are uncorrelated (Cronbach, 1951). When this assumption is not met, reliability estimates given by α can be biased upward or downward, depending on the signs of the correlations among measurement errors (Komaroff, 1997; Raykov, 2001; Zimmerman et al., 1993). Third, α distinguishes only two sources of variance, one attributable to the factor common to the items and another that reflects aspects unique to each item, which include item specificity and random error (Harman, 1976). Other potential sources of variance include differences in raters, timing, context, and method of measurement (Murphy & DeShon, 2000). To the extent these sources of variance underlie a measure, reliability estimates based on α are incorrect (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991).

Other problems result from disregarding the effects of measurement error when some conventional threshold of reliability is reached, such as the .70 criterion commonly applied in OP research. The damage caused by measurement error does not hinge on whether reliability exceeds or falls short of .70, but rather is a matter of degree such that reliabilities that exceed conventional thresholds can signify enough measurement error to alter substantive conclusions. For example, assume two constructs are correlated .40 and measures of each construct exhibit reliabilities of .75, levels that generally would be considered acceptable. Applying the standard formula for attenuation due to measurement error (e.g., Nunnally, 1978), the correlation between the measures would equal .40 $(.75^{1/2}.75^{1/2}) = .30$, a reduction of 25 per cent. If the measures represented predictor and criterion variables, the corresponding R^2 would drop from .16 to .09, a reduction of nearly 50 per cent. These differences are sufficiently large to change the results of statistical tests and alter substantive conclusions.

The effects of measurement error become more complicated in the multivariate case, as when two or more variables measured with error are used as predictors in regression analysis. To illustrate, consider a regression equation in which X_1 and X_2 are predictors of Y. Assuming X_1 and X_2 are measured without error, the standardized regression coefficients for X_1 and X_2 can be written as follows (Cohen, Cohen, Aiken, & West, 2003):

$$\beta_1 = \frac{r_{Y1} - r_{Y2}r_{12}}{1 - r_{12}^2} \tag{1}$$

$$\beta_2 = \frac{r_{Y2} - r_{Y1}r_{12}}{1 - r_{12}^2} \tag{2}$$

where β_1 and β_2 are standardized regression coefficients for X_1 and X_2, r_{Y_1} and r_{Y_2} are the correlations of Y with X_1 and X_2, and r_{12} is the correlation between X_1 and X_2. To make matters concrete, assume X_1 and X_2 correlate .40 and .30 with Y and are correlated .50 with one another. Applying Equations 1 and 2 yields $\beta_1 = .33$ and $\beta_2 = .13$. If X_1 and X_2 are measured with error, Equations 1 and 2 become:

$$\beta_1 = \frac{\sqrt{r_{11}}(r_{22}r_{Y1} - r_{Y2}r_{12})}{r_{11}r_{22} - r_{12}^2} \tag{3}$$

$$\beta_2 = \frac{\sqrt{r_{22}}(r_{11}r_{Y2} - r_{Y1}r_{12})}{r_{11}r_{22} - r_{12}^2} \tag{4}$$

where r_{11} and r_{22} are reliability estimates of X_1 and X_2. Note that, if r_{11} and r_{22} both equal unity, meaning X_1 and X_2 are perfectly reliable, then Equations 3 and 4 reduce to Equations 1 and 2. Also note that r_{11} and r_{22} both appear in the numerators and denominators of Equations 3 and 4. As a result, measurement error in X_1 and X_2 can decrease or increase β_1 and β_2, depending on the amount of measurement error and the correlations among X_1, X_2, and Y. For instance, if X_1 and X_2 both have reliabilities of .75 and the correlations among X_1, X_2, and Y are as before, then β_1 increases from .33 to

.42, whereas β_2 decreases from .13 to .07. Again, changes such as these can alter the substantive conclusions drawn from a study. Moreover, the fact that β_1 and β_2 can increase or decrease means that measurement error does not make results more conservative, as is sometimes asserted, but can make results lenient or conservative. The effects of measurement become more complicated as the number of predictors increases, and the resulting bias can be appreciable even when reliabilities reach levels normally considered adequate.

To address these problems, we should expand the tools we use to estimate measurement error and, whenever possible, use statistical methods that take measurement error into account. For instance, the assumption of τ-equivalence that underlies α is relaxed by coefficient ω (Heise & Bohrnstedt, 1970; Smith, 1974), which reduces to α when item loadings happen to be equal (Greene & Carmines, 1980). ω can be readily computed using results from confirmatory factor analysis (Edwards, 2003; Jöreskog, 1971). Correlated measurement errors can be accommodated by adjusting formulas for α and ω (Komaroff, 1997) or by incorporating correlated errors into confirmatory factor models that generate parameters used to compute α or ω, provided the number of correlations added does not render the model unidentified. Sources of measurement error other than those addressed by α and ω can be accommodated by generalizability theory (Cronbach et al., 1972), which is an extension of classical measurement theory upon which α and ω are based. Generalizability theory can be applied by estimating variance components using analysis of variance procedures or by incorporating factors that represent different sources of variance in a confirmatory factor analysis (DeShon, 1998; Marcoulides, 1996).

In statistical analyses, measurement error can be taken into account by adjusting correlations or covariances for measurement error using corrections for disattenuation (Nunnally, 1978) or by using structural equation modeling with latent variables (Bollen, 1989; Kline, 2004; Loehlin, 2004). Parameter estimates yielded by these approaches are often similar, but when disattenuated correlations or covariances are used in conventional statistical procedures, such as OLS regression, the reported standard errors are incorrect because they do not take into account the sampling variability of the reliability estimates used in the disattenuation computations. This problem is avoided by structural equation modeling with latent variables, provided the usual assumptions of this procedure are met (Bollen, 1989). Structural equation modeling is well suited to models that specify linear relationships among continuous variables, and extensions are available for handling models with categorical variables (Muthén & Muthén, 2001). However, methods for testing nonlinear and interactive relationships are complicated and remain difficult to implement (Cortina, Chen, & Dunlap, 2001; Jöreskog, 1998; Li, Harmer, Duncan, Duncan, Acock, & Boles, 1998), and many published presentations contain omissions or errors that hinder their utility. Fortunately, methodological developments concerning nonlinear and interactive effects in structural equation modeling are underway, and these methods should become more accessible and practical in the near future (Edwards & Kim, 2002).

Incorporate control variables into theoretical models

OP studies often include control variables in statistical analyses (Becker, 2005; Breaugh, 2006). Typically, a control variable represents some factor thought to influence the relationships among the substantive variables under investigation. For example, when testing the relationship between a predictor and outcome using regression analysis, a control variable might be added to the regression equation to determine whether the coefficient on the predictor is reduced when the control variable is taken into account. If the coefficient becomes nonsignificant, the relationship between the predictor and outcome is often declared spurious, and the theoretical and practical worth of the predictor is

questioned. If the coefficient remains significant, we become more confident in the predictor, and our confidence increases as the predictor survives each additional control variable added to the equation (Gordon, 1968; Meehl, 1971).

Control variables are useful in empirical research, as they can reduce bias due to omitted variables (James, 1980; Mauro, 1990) and address alternative explanations that undermine causal inference (Cook & Campbell, 1979; Rubin, 1974). However, control variables introduce various complications that are often overlooked in OP research. First, control variables change the meaning of the substantive variables under study. From a statistical standpoint, including a control variable along with a predictor in a regression equation is equivalent to replacing the predictor with the residual from a separate regression in which the predictor is regressed on the control variable (Breaugh, 2006). Hence, substantive conclusions should refer not to the predictor *in toto*, but instead to that part of the predictor that is unrelated to the control variable. For example, when the relationship between procedural justice and job satisfaction is examined controlling for distributive, interactional, and interpersonal justice (Colquitt, Conlon, Wesson, Porter, & Ng, 2001), the relationship should be attributed not to procedural justice as it is usually defined, but instead to the part of procedural justice that is unrelated to distributive, interactional, and interpersonal justice. Based on population correlations reported by Colquitt et al. (2001), these three forms of justice explain over half the variance in procedural justice. The residual variance in procedural justice could be interpreted in various ways, but it is certainly not the same as procedural justice itself.

Second, when control variables are measured with error, estimates of relationships for substantive variables can be biased even when their measures are free from error. To illustrate, assume a multiple regression equation contains two predictors, X_1 and X_2, which are framed as a control variable and a substantive variable, respectively. If X_2 is measured without error, its standardized regression coefficient can be written as:

$$\beta_2 = \frac{(r_{11}r_{Y2} - r_{Y1}r_{12})}{r_{11} - r_{12}^2} \qquad (5)$$

Equation 5 is a simplified version of Equation 4 in which r_{22}, the reliability of X_2, is set to unity. As shown by Equation 5, the reliability of X_1 (i.e., r_{11}) influences the estimate of β_2 even when the measure of X_2 is perfectly reliable. For instance, if X_2 is a dummy variable that distinguishes treatment and control groups and X_1 is a covariate, the estimate of the treatment effect is biased to the extent the covariate is measured with error (Cook & Campbell, 1979; Lord, 1950). The effects of measurement error become more complicated as additional control variables are used and substantive variables are measured with error. These effects are likely to be overlooked because the reliabilities of control variables are generally given less attention than those of substantive variables (Becker, 2005).

Finally, the nature of the relationship between a substantive and control variable has important implications for the conclusions drawn from a study. To illustrate, consider the causal models in Figure 2, which depict three different relationships between a control variable X_1 and a substantive variable X_2. In Figure 2a, the control and substantive variables are treated as correlated predictors, such that their effects on the outcome Y are represented by paths a and b, respectively. The correlation between the variables, indicated by path c, is attributed to forces outside the model and is excluded from causal inferences about the effects of X_1 and X_2 on Y. This model corresponds to statistical control using regression analysis, in which the unique contributions of the control and substantive variables are estimated and the correlation between the variables is disregarded. In Figure 2b, the control variable causes the substantive variable. As before, the effect of the substantive variable is path b, but the effect of the control variable is the quantity a+ bc (Alwin & Hauser, 1975). In Figure 2c, the substantive variable causes the control variable, in which case the effect of the control variable is path a and the
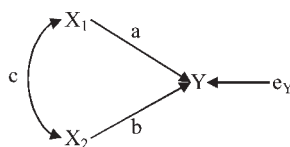
a. Control variable (X_1) correlated with substantive variable (X_2)
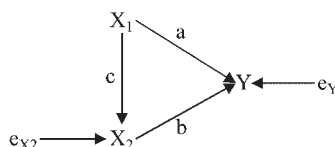
b. Control variable (X_1) cause of substantive variable (X_2)
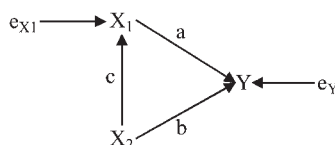
c. Control variable (X_1) caused by substantive variable (X_2)

Figure 2. Models depicting three different relationships between a control variable and a substantive variable

effect of the substantive variable is the quantity $b + ac$. Hence, when the effect of a substantive variable is tested using path b, as is usually the case, the test is correct for Models 2a and 2b but incorrect for Model 2c, given that the term ac is disregarded (Christenfeld, Sloan, Carroll, & Greenland, 2004).

Substantive and control variables might relate to outcomes in other ways that would dictate different analyses and conclusions. For instance, referring to Model 2b, if the effect of the control variable on the outcome is fully mediated by the substantive variable, then path a should be omitted, and the control variable should not appear in the equation relating the substantive variable to the outcome (James, Muliak, & Brett, 2006; Meehl, 1971). Alternately, the control variable might function as a moderator, such that the notion of control is not whether a substantive variable predicts the outcome beyond the control variable, but whether the relationship between the substantive variable and outcome differs across levels of the control variable (Hull, Tedlie, & Lehn, 1992). In this case, the control variable should be supplemented by its product with the substantive variable (Aiken & West, 1991). As the number of control and substantive variables increases, the ways in which they might combine become more complex.

The foregoing issues lead to several recommendations for OP researchers. In particular, when control variables are used, substantive variables should be viewed as the portions of the original variables that are unrelated to the control variables. Control variables do not merely render tests of substantive variables more stringent. Rather, they change the meaning of effects ascribed to substantive

variables. This perspective should be integrated into the development of hypotheses and the conclusions drawn from studies (Becker, 2005). In addition, measures of control variables should be subjected to the same standards of reliability and construct validity as measures of substantive variables. As shown earlier, measurement error in control variables can bias results for substantive variables, and this source of bias should be taken into account when control variables are analyzed. Finally, control variables should be formally incorporated into theoretical models, making explicit the presumed causal structure by which control variables relate to one another and to the substantive variables under study (Becker, 2005; Meehl, 1971). Doing so will yield more appropriate tests of substantive variables and increase the theoretical rigor of research. After all, one researcher's control variable is another researcher's substantive variable, and putting both types of variables through the same conceptual and empirical scrutiny can facilitate advancement within the broader scope of the field.

Examine alternative models

In OP research, we usually limit our analyses to tests that correspond directly to the hypotheses we develop. As a general rule, we use our hypotheses to dictate the analyses we conduct and avoid analyses that deviate from the relationships we predict. Such analyses are usually deemed exploratory at best and atheoretical at worst, and those that appear in published work are often relegated to footnotes or labeled as supplemental, implying they are tangential to the main purpose of the study. For the most part, analyses we call supplemental are conducted to probe reasons why hypotheses were not supported or to defend results taken as evidence in favor of hypotheses. When results seem to provide clear support for hypotheses, we see little reason to conduct further analyses, and the conceptual framework underlying the hypotheses is viewed as corroborated.

Focusing analyses on hypothesized relationships has the virtue of parsimony and avoids the analytical mayhem that could result from testing all possible relationships among a set of variables. Nonetheless, limiting our analyses to relationships we predict renders tests of our conceptual models incomplete. In many OP studies, the hypotheses tested involve some subset of the possible relationships among the variables under investigation. Each omitted relationship constitutes an implicit or explicit prediction that the variables involved are *not* related. These predictions should be subject to the same rules of evidence we apply to predictions that relationships differ from zero (Cortina & Folger, 1998; Frick, 1995). Moreover, a hypothesis that predicts a relationship between two variables usually implies that the relationship is linear rather than curvilinear (Cummings, 1982; Hage, 1980), and a hypothesis that predicts a multiplicative interaction implies the absence of quadratic effects (Ganzach, 1997; MacCallum & Mar, 1995). Again, these hypotheses are not fully tested unless alternative functional forms are analyzed. If we establish that the hypothesized relationships have the predicted functional form and rule out nonhypothesized relationships, we still confront the fact that the relationships among our variables can be explained equally well by alternative models with different causal structures (Lee & Hershberger, 1990; MacCallum, Wegener, Uchino, & Fabrigar, 1993; Stelzl, 1986).

We can test our conceptual models more thoroughly by conducting analyses that probe the assumptions that underlie our hypotheses and considering alternative models that might explain the relationships among the variables under study. Relationships predicted to be zero can be tested individually or simultaneously using procedures for assessing model fit in regression analysis (Specht, 1975) or structural equation modeling (Anderson & Gerbing, 1988; James et al., 1982; Medsker et al., 1994). Because these tests represent attempts to confirm the null hypothesis, the study should provide adequate power for nonzero relationships to emerge (Cortina & Folger, 1998; Frick, 1995). Alternative functional forms can be analyzed using higher-order terms and by examining scatterplots of residuals

(Cohen et al., 2003; Pedhazur, 1997). In principle, tests of relationships and functional forms other than those hypothesized are exploratory, and results that deviate from predictions should be considered tentative, pending cross-validation (Anderson & Gerbing, 1988; Cudeck & Browne, 1983; MacCallum, Roznowski, Mar, & Reith, 1994; Mosier, 1951; Snee, 1977). Models with different causal structures might be ruled out based on prior research or could prompt initial design decisions that increase confidence in the proposed causal structure. Unfortunately, there are no analytical criteria for choosing among models that specify different causal orders between variables, regardless of whether the models are statistically equivalent (Lee & Hershberger, 1990; MacCallum et al., 1993; Stelzl, 1986) or impose different restrictions on the relationships among variables (Duncan, 1975). Establishing causal order is a matter of research design, not statistical analysis, and must be addressed before data are collected (Cook & Campbell, 1979; James et al., 1982).

Adopt alternatives to null hypothesis significance tests

The use of null hypothesis significance tests is widespread in OP research (Bonett & Wright, 2007; Hubbard, Parsa, & Luthy, 1997). This practice follows naturally from our tendency to pose hypotheses that predict a nonnull effect of unstated magnitude, which are then tested by determining whether a test statistic differs from zero at some conventional probability level, such as $p < .05$. Statistics with progressively smaller probability levels are adorned with increasing numbers of asterisks, perhaps accompanied by descriptions that distinguish findings that are ''significant'' in an ordinary sense from those that are ''highly significant'' or ''marginally significant.'' Findings declared nonsignificant are omitted from substantive interpretation and are sometimes treated as if the corresponding parameter equals zero, as when a nonsignificant predictor is dropped from a regression analysis and the equation is reestimated. Findings that achieve significance are taken as support for hypotheses and corroboration of the theory from which they were derived.

For decades, null hypothesis significance testing has been criticized on various grounds (Bakan, 1966; Cohen, 1994; Lykken, 1968; Nickerson, 2000; Rozeboom, 1960; Schmidt, 1996). For instance, in most studies, an effect that precisely equals zero can be ruled out *a priori,* which undermines the value of testing whether an observed effect differs from zero (Cohen, 1994; Lykken, 1968; Meehl, 1978). In addition, researchers often misinterpret the *p*-value of a null hypothesis significance test as the probability that the null hypothesis is false given the obtained effect, when in fact it represents the probability of an equal or larger effect if the null hypothesis were true (Cohen, 1994; Nickerson, 2000). Moreover, contrary to popular belief, failure to reject the null hypothesis does not justify the conclusion that the null hypothesis is true (Kluger & Tikochinsky, 2001). To complicate matters, effects classified as significant and nonsignificant are often themselves not significantly different from each another, as when the *p*-values for two tests are .04 and .06 (Gelman & Stern, 2006). Finally, the calibration of *p*-values using terms such as ''marginally significant'' and ''highly significant'' is inappropriate, because a *p*-value does not represent the magnitude of the associated effect (Harcum, 1989).

Various alternatives to null hypothesis significance tests have been proposed. One common recommendation is to report confidence intervals for parameter estimates (Cohen, 1994; Cumming & Finch, 2001; Schmidt, 1996; Tryon, 2001). Confidence intervals capture the degree of uncertainty associated with a parameter estimate and can be used to compare the estimate against any population value, including zero or other values of theoretical or practical interest. Parameter estimates can also be translated into effect sizes (Breaugh, 2003; Carver, 1978; Folger, 1989; Rosenthal, Rosnow, & Rubin, 2000), which are expressed in standard metrics that facilitate comparison and integration across studies. Other recommendations underscore the value of basing inferences about population parameters on replication (Carver, 1978; Hubbard & Ryan, 2000; Krueger, 2001; Lykken, 1968) and

meta-analysis (Hunter & Schmidt, 2004; Schmidt, 1996). Although many researchers argue that these procedures should replace null hypothesis significance tests, others maintain that such tests can be useful in some circumstances (Abelson, 1997; Chow, 1996; Frick, 1996; Hagen, 1997; Harris, 1997; Mulaik, Raju, & Harshman, 1997) and can complement the information provided by proposed alternatives (Howard, Maxwell, & Fleming, 2000). Alternatives to null hypothesis significance tests have been slow to disseminate (Krueger, 2001), but they should be taken seriously, as they have the capacity to dramatically change the way we evaluate theories and accumulate knowledge in our field (Schmidt, 1996).

Rethinking How We Think About Methods

The methodological barriers considered here have impeded progress in OP research, and all of us have something to gain by addressing these barriers in our own work. To facilitate this process, I would like to offer some suggestions concerning how we think about methods. Most of us learn about methods in graduate school, where we take courses that cover research design, measurement, statistics, and related topics. Over the years, my discussions with students and colleagues have indicated that few people find these topics intrinsically interesting, and some find them downright aversive, the educational equivalent of swallowing castor oil. After we graduate, we continue to rely on the methods we initially learned, perhaps because we become better at them with practice, fall into habits that are eventually hard to break, or see little reason to take another dosage of methodological medicine, particularly if the first taste was not that pleasant. Occasionally, a reviewer will raise objections to our work that force us to scratch some unfamiliar methodological ground, or a student will apply a novel statistic that requires us to become at least conversant with the new tool. Beyond that, keeping up with methodological developments might be seen as a distraction from our substantive research, and given the constraints on our time and energy, we limit our exposure to the research methods literature. This pattern is surely not universal, but I believe it is quite common, and it is partly responsible for the methodological barriers we confront.

What can be done to encourage us to stay abreast of methodological developments and willingly apply them to our research? I would first suggest a change in attitude about methods, such that we view them not as hurdles to clear or tangents to our primary work, but instead as the skills of our craft as researchers. As noted earlier, research methods are intimately involved in the interplay between theory and empiricism by which our field makes progress (Runkel & McGrath, 1972), and if progress is our ultimate goal, then it is in our collective interest to hone our methodological skills. Moreover, by sharpening our methodological skills, we can increase the rigor of our conceptual thinking, see new ways to answer theoretical questions, and perhaps identify questions that would not have occurred to us otherwise (Blalock, 1969; Pearl, 2000). Certainly, we have different aptitudes and tastes for methodological topics, and not everyone will consume the research methods literature for its own sake, let alone conduct primary research on methodological topics. Nonetheless, I firmly believe that good theoretical and methodological thinking go hand in hand, and developing our expertise in both domains creates synergies that make us better as researchers.

I would also encourage to avoid using methodological issues as clubs to beat one another or forums to stage arguments. I confess that, at times during my career, I have used a critical tone in my methodological writing, and I have willingly stepped into debates partly for the thrill of the exchange (Bedeian, Day, Edwards, Tisak, & Smith, 1994). I have since learned that we undermine our ability to convince others to adopt new methods when we have previously held up their work as bad practice, and

although debates might have the entertainment value of a boxing match, they often do little more than create confusion and give researchers sources to cite that appear to support one side of what remains an unresolved issue. Ultimately, we all conduct research that has flaws (McGrath, 1982), and as a field, we are best served by working together to help one another improve the methodological quality and rigor of our research.

Looking Ahead

In this paper, I have identified some methodological barriers to progress in OP research, described the problems these barriers cause, and offered some suggestions for dealing with the barriers, accompanied by references to sources that address the issues at hand in much greater detail that I have here. I hope my discussion of these barriers gives us reason to pause and think about how they apply to our own work and what we can do to address them. I have attempted to frame the barriers and their solutions in a positive light, such that working to overcome the barriers is not about repairing research that is broken, but about bringing our research to higher methodological levels, such that the progress of our field is more assured.

Author biography

Jeffrey R. Edwards is the Belk Distinguished Professor of Organizational Behavior and Strategy at the Kenan-Flagler Business School at the University of North Carolina at Chapel Hill. He received his Ph.D. from Carnegie Mellon University and has held positions at the University of Virginia and the University of Michigan. His research focuses on person-environment fit in organizations, stress, coping, and well-being, the work-nonwork interface, and methodological issues in organizational research. He is a Fellow of the Academy of Management, the American Psychological Association, the Society of Industrial and Organizational Psychology, and the Center for the Advancement of Research Methods and Analysis (CARMA) and has been elected to the Society of Organizational Behavior.

References

Abelson, R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science*, *8*, 12–15.

Aguinis, H., & Adams, S. K. R. (1998). Social-role versus structural models of gender and influence use in organizations: A strong inference approach. *Group & Organization Management*, *23*, 414–446.

Aiken, L. A., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Alliger, G. M., & Williams, K. J. (1993). Using signal-contingent experience sampling methodology to study work in the field: A discussion and illustration examining task perceptions and mood. *Personnel Psychology*, *46*, 525–549.

Alwin, D. F., & Hauser, R. M. (1975). The decomposition of effects in path analysis. *American Sociological Review*, *40*, 37–47.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two step approach. *Psychological Bulletin*, *103*, 411–423.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*, 423–437.

Balkundi, P., & Harrison, D. A. (2006). Ties, leaders, and time in teams: Strong inference about network structure's effects on team viability and performance. *Academy of Management Journal*, *49*, 49–68.

Becker, T. E. (2005). Potential problems in the statistical control of variables in organizational research: A qualitative analysis with recommendations. *Organizational Research Methods*, *8*, 274–289.

Bedeian, A. G., Day, D. V., Edwards, J. R., Tisak, J., & Smith, C. S. (1994). Difference scores: Rationale, formulation, and interpretation. *Journal of Management*, *20*, 673–698.

Blalock, H. M. (1969). *Theory construction: From verbal to mathematical formulations*. New York: Prentice-Hall.

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology*, *54*, 579–616.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bonett, D. G., & Wright, T. A. (2007). Comments and recommendations regarding the hypothesis testing controversy. *Journal of Organizational Behavior*, *28*, 647–659.

Breaugh, J. A. (2003). Effect size estimation: Factors to consider and mistakes to avoid. *Journal of Management*, *29*, 79–97.

Breaugh, J. A. (2006). Rethinking the control of nuisance variables in theory testing. *Journal of Business and Psychology*, *20*, 429–443.

Campbell, J. P. (1976). Psychometric theory. In M. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 185–222). Chicago: Rand McNally.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Education Review*, *48*, 378–399.

Chow, S. L. (1996). *Statistical significance: Rationale, validity and utility*. Thousand Oaks, CA: Sage.

Christenfeld, N. J. S., Sloan, R. P., Carroll, D., & Greenland, S. (2004). Risk factors, confounding, and the illusion of statistical control. *Psychosomatic Medicine*, *66*, 868–875.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Collins, J. Hall, N. & Paul L. A. (Eds.). (2004). *Causation and counterfactuals*. Cambridge, MA: MIT Press.

Colquitt, J. A., Conlon, D. E., Wesson, M. J., Porter, C. O. L. H., & Ng, K. Y. (2001). Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. *Journal of Applied Psychology*, *86*, 425–445.

Conway, J. M. (1998). Estimation and uses of the proportion of method variance in multitrait-multimethod data. *Organizational Research Methods*, *1*, 209–222.

Conway, J. M. (2002). Method variance and method bias in industrial and organizational psychology. In S. G. Rogelberg (Ed.), *Handbook of research methods in industrial and organizational psychology* (pp. 344–365). Malden, MA: Blackwell.

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

Cook, T. D., Campbell, D. T., & Peracchio, L. (1990). Quasi experimentation. In M. D. Dunnette, & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 1, pp. 491–576). Palo Alto, CA: Consulting Psychologists Press.

Cooper, W. H., & Richardson, A. J. (1986). Unfair comparisons. *Journal of Applied Psychology*, *71*, 179–184.

Cortina, J. M., Chen, G., & Dunlap, W. P. (2001). Testing interaction effects in LISREL: Examination and illustration of available procedures. *Organizational Research Methods*, *4*, 324–360.

Cortina, J. M., & Folger, R. G. (1998). When is it acceptable to accept a null hypothesis: No way, Jose? *Organizational Research Methods*, *1*, 334–350.

Crampton, S. M., & Wagner, J. A., III. (1994). Percept-percept inflation in microorganizational research: An investigation of prevalence and effect. *Journal of Applied Psychology*, *79*, 67–76.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Cronbach, L. J., Gleser, G., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research*, *18*, 147–167.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*, 532–574.

Cummings, L. L. (1982). Organizational behavior. *Annual Review of Psychology*, *33*, 541–579.

DeMaio, T. J., & Rothgeb, J. M. (1996). Cognitive interviewing techniques: In the lab and in the field. In N. Schwarz, & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 177–196). San Francisco: Jossey-Bass.

DeShon, R. (1998). A cautionary note on measurement error corrections in structural equation modeling. *Psychological Methods*, *3*, 412–423.

Doty, D. H., & Glick, W. H. (1998). Common methods bias: Does common methods variance really bias results? *Organizational Research Methods*, *1*, 374–406.

Duncan, O. D. (1975). *Introduction to structural equation models*. New York: Academic Press.

Edwards, J. R. (1994). The study of congruence in organizational behavior research: Critique and a proposed alternative. *Organizational Behavior and Human Decision Processes*, *58*, 51–100 (erratum, 58, 323–325).

Edwards, J. R. (1996). An examination of competing versions of the person-environment fit approach to stress. *Academy of Management Journal*, *39*, 292–339.

Edwards, J. R. (2003). Construct validation in organizational behavior research. In J. Greenberg (Ed.), *Organizational behavior: The state of the science* (2nd ed., pp. 327–371). Mahwah, NJ: Erlbaum.

Edwards, J. R., & Kim, T. Y. (2002). Moderation in structural equation modeling: Specification, estimation, and interpretation using quadratic structural equations. Paper presented at the Annual Meeting of the Society for Industrial and Organizational Psychology, Toronto, Ontario.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Feldman, J. M., & Lynch, J. G., Jr. (1988). Self-generated validity and other effects of measurement of belief, attitude, intention, and behavior. *Journal of Applied Psychology*, *73*, 421–435.

Folger, R. (1989). Significance tests and the duplicity of binary decisions. *Psychological Bulletin*, *106*, 155–160.

Frick, R. W. (1995). Accepting the null hypothesis. *Memory and Cognition*, *23*, 132–138.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, *1*, 379–390.

Fromkin, H. L., & Streufert, S. (1976). Laboratory experimentation. In M. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 415–466). Chicago: Rand McNally.

Ganzach, Y. (1997). Misleading interaction and curvilinear terms. *Psychological Methods*, *2*, 235–247.

Gelman, A., & Stern, H. (2006). The difference between ''significant'' and ''not significant'' is not itself statistically significant. *The American Statistician*, *60*, 328–331.

Gordon, R. A. (1968). Issues in multiple regression. *American Journal of Sociology*, *73*, 592–616.

Greene, V. L., & Carmines, E. G. (1980). Assessing the reliability of linear composites. In K. F. Schuessler (Ed.), *Sociological methodology* (pp. 160–175). San Francisco: Jossey-Bass.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*, 1464–1480.

Haines, E. L., & Sumner, K. E. (2006). Implicit measurement of attitudes, stereotypes, and self-concepts in organizations: Teaching old dogmas new tricks. *Organizational Research Methods 9*, 536–553.

Hage, J. (1980). *Theories of organizations*. New York: Wiley.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist*, *52*, 15–24.

Harcum, E. R. (1989). The highly inappropriate calibrations of statistical significance. *American Psychologist*, *44*, 964.

Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). Chicago: University of Chicago Press.

Harris, R. J. (1997). Significance tests have their place. *Psychological Science*, *8*, 8–11.

Harrison, D. A., & McLaughlin, M. E. (1993). Cognitive processes in self-report responses: Tests of item context effects in work attitude measures. *Journal of Applied Psychology*, *78*, 129–140.

Harrison, D. A., & McLaughlin, M. E. (1996). Structural properties and psychometric qualities of organizational self-reports: Field tests of connections predicted by cognitive theory. *Journal of Management*, *22*, 313–338.

Harrison, D. A., McLaughlin, M. E., & Coalter, T. M. (1996). Context, cognition, and common method variance: Psychometric and verbal protocol evidence. *Organizational Behavior and Human Decision Process*, *68*, 246–261.

Heise, D. R., & Bohrnstedt, G. W. (1970). Validity, invalidity, and reliability. In E. F. Borgatta, & G. W. Bohrnstedt (Eds.), *Sociological methodology* (pp. 104–129). San Francisco: Jossey-Bass.

Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (2007). *Experience sampling method: Measuring the quality of everyday life*. Thousand Oaks, CA: Sage.

Henry, G. T. (1990). *Practical sampling*. Newbury Park, CA: Sage.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945–960.

Howard, G. S., Maxwell, S. E., & Fleming, K. J. (2000). The proof of the pudding: An illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. *Psychological Methods*, *5*, 315–332.

Hubbard, R. (1997). Publication bias against null results. *Psychological Reports*, *80*, 337–338.

Hubbard, R., Parsa, R. A., & Luthy, M. R. (1997). The spread of statistical significance testing in psychology: The case of the *Journal of Applied Psychology*, 1917–1994. *Theory & Psychology*, *7*, 545–554.

Hubbard, R., & Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology—And its future prospects. *Educational and Psychological Measurement*, *60*, 661–681.

Hull, J. G., Tedlie, J. C., & Lehn, D. A. (1992). Moderator variables in personality research: The problem of controlling for plausible alternatives. *Personality and Social Psychology Bulletin*, *18*, 115–117.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis* (2nd ed.). Thousand Oaks, CA: Sage.

Ilgen, D. R. (1986). Laboratory research: A question of when, not if. In E. A. Locke (Ed.), *Generalizing from laboratory to field settings* (pp. 257–267). Lexington, MA: D. C. Heath and Company.

James, L. R. (1980). The unmeasured variable problem in path analysis. *Journal of Applied Psychology*, *65*, 415–421.

James, L. R. (1998). Measurement of personality via conditional reasoning. *Organizational Research Methods*, *1*, 131–163.

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal analysis: Assumptions, models and data*. Beverly Hills, CA: Sage.

James, L. R., Mulaik, S. A., & Brett, J. M. (2006). A tale of two methods. *Organizational Research Methods*, *9*, 233–244.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*, 109–133.

Jöreskog, K. G. (1998). Interaction and nonlinear modeling: Issues and approaches. In R. E. Schumacker, & G. A. Marcoulides (Eds.), *Interaction and nonlinear effects in structural equation modeling* (pp. 239–250). Mahwah, NJ: Erlbaum.

Kahneman, D. (2003). Experiences of collaborative research. *American Psychologist*, *58*, 723–730.

Kemery, E. R., & Dunlap, W. P. (1986). Partialling factor scores does not control method variance: A rejoinder to Podsakoff and Todor. *Journal of Management*, *12*, 525–530.

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin*, *112*, 165–172.

Kline, R. B. (2004). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.

Kluger, A. N., & Tikochinsky, J. (2001). The error of accepting the ''theoretical'' null hypothesis: The rise, fall and resurrection of common sense hypotheses in psychology. *Psychological Bulletin*, *127*, 408–423.

Komaroff, E. (1997). Effect of simultaneous violations of essential tau-equivalence and uncorrelated error on coefficient alpha. *Applied Psychological Measurement*, *21*, 337–348.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist*, *56*, 16–26.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.

Latham, G. P., Erez, M., & Locke, E. A. (1988). Resolving scientific disputes by the joint design of crucial experiments by the antagonists: Application to the Erez-Latham dispute regarding participation in goal setting. *Journal of Applied Psychology*, *73*, 753–772.

LeBreton, J. M., Barksdale, C. D., Robin, J., & James, L. R. (2007). Measurement issues associated with conditional reasoning tests: Indirect measurement and test faking. *Journal of Applied Psychology*, *92*, 1–16.

Lee, S., & Hershberger, S. (1990). A simple rule for generating equivalent models in covariance structure modeling. *Multivariate Behavioral Research*, *25*, 313–334.

Li, F., Harmer, P., Duncan, T. E., Duncan, S. C., Acock, A., & Boles, S. (1998). Approaches to testing interaction effects using structural equation modeling methodology. *Multivariate Behavioral Research*, *33*, 1–39.

Loehlin, J. C. (2004). *Latent variable models: An introduction to factor, path, and structural analysis* (4th ed.). Hillsdale, NJ: Erlbaum.

Lord, F. M. (1950). Large-sample covariance analysis when the control variable is fallible. *Journal of the American Statistical Association*, *55*, 307–321.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*, 151–159.

MacCallum, R. C., & Mar, C. M. (1995). Distinguishing between moderator and quadratic effects in multiple regression. *Psychological Bulletin*, *118*, 405–421.

MacCallum, R., Roznowski, M., Mar, C. M., & Reith, J. V. (1994). Alternative strategies for cross-validation of covariance structure models. *Multivariate Behavioral Research*, *29*, 1–32.

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin*, *114*, 185–199.

Mahoney, M. J. (1985). Open exchange and epistemic progress. *American Psychologist*, *40*, 29–39.

Marcoulides, G. A. (1996). Estimating variance components in generalizability theory: The covariance structure analysis approach. *Structural Equation Modeling*, *3*, 290–299.

Marsh, H. W. (1989). Confirmatory factor analyses of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement*, *13*, 335–361.

Marsh, H. W., & Bailey, M. (1991). Confirmatory factor analyses of multitrait-multimethod data: A comparison of alternative models. *Applied Psychological Measurement*, *15*, 47–70.

Mauro, R. (1990). Understanding L.O.V.E. (left out variables error): A method for estimating the effects of omitted variables. *Psychological Bulletin*, *108*, 314–329.

Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. *Psychological Methods*, *12*, 23–44.

McGrath, J. E. (1982). Dilemmatics: A study of research choices and dilemmas. In J. E. McGrath, J. Martin, & R. A. Kulka (Eds.), *Judgment calls in research* (pp. 69–102). Beverly Hills, CA: Sage.

Medsker, G. J., Williams, L. J., & Holahan, P. J. (1994). A review of current practices for evaluating causal models in organizational behavior and human resources management research. *Journal of Management*, *20*, 439–464.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103–115.

Meehl, P. E. (1971). High school yearbooks: A reply to Schwarz. *Journal of Abnormal Psychology*, *77*, 143–148.

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.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*, 108–141.

Mellers, M., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, *12*, 269–275.

Millsap, R. E. (1992). Sufficient conditions for rotational uniqueness in the additive MTMM model. *British Journal of Mathematical and Statistical Psychology*, *45*, 125–138.

Mitchell, T. R. (1985). An evaluation of the validity of correlational research conducted in organizations. *Academy of Management Review*, *10*, 192–205.

Mitchell, T. R., & James, L. R. (2001). Building better theory: Time and the specification of when things happen. *Academy of Management Review*, *26*, 530–547.

Mosier, C. I. (1951). AT Symposium: The need and means of cross-validation. *Educational and Psychological Measurement*, *11*, 5–11.

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and place for significance testing. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests*? (pp. 65–116). Hillsdale, NJ: Erlbaum.

Murphy, K. R., & DeShon, R. (2000). Progress in psychometrics: Can industrial and organizational psychology catch up? *Personnel Psychology*, *53*, 913–924.

Muthén, L. K., & Muthén, B. O. (2001). *Mplus user's guide*. Los Angeles, CA: Muthen & Muthen.

Nadler, J., Thompson, L., & Van Boven, L. (2003). Learning negotiation skills: Four models of knowledge creation and transfer. *Management Science*, *49*, 529–540.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of old and continuing controversy. *Psychological Methods*, *5*, 241–301.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231–259.

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measures. *Psychometrika*, *32*, 1–13.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research* (3rd ed.). New York: Holt.

Platt, J. R. (1964). Strong inference. *Science*, *146*, 347–353.

Pfeffer, J. (1993). Barriers to the advance of organizational science: Paradigm development as a dependent variable. *Academy of Management Review*, *18*, 599–620.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*, 879–903.

Podsakoff, P. M., & Organ, D. W. (1986). Self-reports in organizational research: Problems and prospects. *Journal of Management*, *12*, 531–544.

Podsakoff, P. M., & Todor, W. D. (1985). Relationships between leader reward and punishment behavior and group processes and productivity. *Journal of Management*, *11*, 55–73.

Popper, K. R. (1959). *The logic of scientific discovery*. New York: Basic Books.

Raykov, T. (1997). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research*, *32*, 329–353.

Raykov, T. (2001). Bias of coefficient alpha for fixed congeneric measures with correlated errors. *Applied Psychological Measurement*, *25*, 69–76.

Rosenbaum, P. R. (1995). *Observational studies*. New York: Springer-Verlag.

Rosenthal, R. (1979). The ''file drawer problem'' and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. Oxford: Cambridge University Press.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*, 416–428.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, *6*, 34–58.

Runkel, P. J., & McGrath, J. E. (1972). *Research on human behavior: A systematic guide to method*. New York: Holt.

Sackett, P. R., & Larson, J. R. (1990). Research strategies and tactics in industrial and organizational psychology. In M. D. Dunnette, & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 1, pp. 419–490). Palo Alto, CA: Consulting Psychologists Press.

Salmon, W. C. (1998). *Causality and explanation*. Oxford: Oxford University Press.

Scandura, T. A., & Williams, E. A. (2000). Research methodology in management: Current practices, trends, and implications for future research. *Academy of Management Journal*, *43*, 1248–1264.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*, 115–129.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, *1*, 199–223.

Schmitt, N. (1994). Method bias: The importance of theory and measurement. *Journal of Organizational Behavior*, *15*, 393–398.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*, 350–353.

Schmitt, N., & Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement*, *10*, 1–22.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. New York: Academic Press.

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

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, *54*, 93–105.

Schwarz, N. (2007). Cognitive aspects of survey methodology. *Applied Cognitive Psychology*, *21*, 277–287.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Shaw, J. D., Gupta, N., & Delery, J. E. (2005). Alternative conceptualizations of the relationship between voluntary turnover and organizational performance. *Academy of Management Journal*, *48*, 50–68.

Smith, K. W. (1974). On estimating the reliability of composite indexes through factor analysis. *Sociological Methods and Research*, *2*, 485–510.

Snee, R. D. (1977). Validation of regression models: Methods and examples. *Technometrics*, *19*, 415–428.

Sosa, E. & Tooley M. (Eds.). (1993). *Causation*. Oxford: Oxford University Press.

Specht, D. A. (1975). On the evaluation of causal models. *Social Science Research*, *4*, 113–133.

Spector, P. E. (1994). Using self-report questionnaires in OB research: A comment on the use of a controversial method. *Journal of Organizational Behavior*, *15*, 385–392.

Spector, P. E. (2006). Method variance in organizational research: Truth or urban legend? *Organizational Research Methods*, *9*, 221–232.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search* (2nd ed.). Cambridge, MA: MIT Press.

Stelzl, I. (1986). Changing a causal hypothesis without changing the fit: Some rules for generating equivalent path models. *Multivariate Behavioral Research*, *21*, 309–332.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, *49*, 108–112.

Stone, A. A., Shiffman, S. S., & DeVries, M. W. (1999). *Ecological momentary assessment* In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well-being: The foundations of hedonic psychology* (pp. 353–373). New York: Russell Sage Foundation.

Stone, A. A., Turkkan, J. S., Bachrach, C. A., Jobe, J. B., Kurtzman, H. S., & Cain V. S. (Eds.). (2000). *The science of self-report: Implications for research and practice*. Mahwah, NJ: Erlbaum.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass.

Taber, T. D. (1991). Triangulating job attitudes with interpretive and positivist measurement methods. *Personnel Psychology*, *44*, 577–600.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, *103*, 299–314.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. New York: Cambridge University Press.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, *6*, 371–386.

Wainer, H. (1986). *Drawing inferences from self-selected samples*. New York: Springer-Verlag.

Weick, K. (1965). Laboratory experiments with organizations. In J. March (Ed.), *Handbook of organizations* (pp. 194–260). Chicago, IL: Rand McNally.

White, P. (1980). Limitations on verbal reports of internal events: A refutation of Nisbett and Wilson and of Bem. *Psychological Review*, *87*, 105–112.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement*, *9*, 1–26.

Williams, L. J., & Anderson, S. E. (1994). An alternative approach to method effects by using latent-variable models: Applications in organizational behavior research. *Journal of Applied Psychology*, *79*, 323–332.

Williams, L. J., & Brown, B. K. (1994). Method variance in organizational behavior and human resources research: Effects on correlations, path coefficients, and hypothesis testing. *Organizational Behavior and Human Decision Process*, *57*, 185–209.

Williams, L. J., Cote, J. A., & Buckley, M. R. (1989). Lack of method variance in self-reported affect and perceptions at work: Reality or artifact? *Journal of Applied Psychology*, *74*, 462–468.

Williams, L. J., Gavin, M. B., & Williams, M. L. (1996). Measurement and nonmeasurement processes with negative affectivity and employee attitudes. *Journal of Applied Psychology*, *81*, 88–101.

Winship, C., & Radbill, L. (1994). Sampling weights and regression analysis. *Sociological Methods and Research*, *23*, 230–257.

Zimmerman, D. W., Zumbo, B. D., & Lalonde, C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. *Educational and Psychological Measurement*, *53*, 33–49.