Organizational Research Methods
15(2) 199-228

DOI: 10.1177/1094428111428356
http://orm.sagepub.com

How Can Significance Tests Be Deinstitutionalized?

Marc Orlitzky[1]

Abstract

The purpose of this article is to propose possible solutions to the methodological problem of null hypothesis significance testing (NHST), which is framed as deeply embedded in the institutional structure of the social and organizational sciences. The core argument is that, for the deinstitutionalization of statistical significance tests, minor methodological changes within an unreformed epistemology will be as unhelpful as emotive exaggerations of the ill effects of NHST. Instead, several institutional-epistemological reforms affecting cultural-cognitive, normative, and regulative processes and structures in the social sciences are necessary and proposed in this article. In the conclusion, the suggested research reforms, ranging from greater emphasis on inductive and abductive reasoning to statistical modeling and Bayesian epistemology, are classified according to their practical importance and the time horizon expected for their implementation. Individual-level change in researchers' use of NHST is unlikely if it is not facilitated by these broader epistemological changes.

Keywords

philosophy of science, quantitative research, historical social science, meta-analysis, structural equation modeling

Over the past 70 years, null hypothesis significance testing (NHST), which is a dichotomous statistical inference technique for evaluating research hypotheses by assessing the probability of the observed data given that the null hypothesis[1] is true (J. Cohen, 1994), has frequently been criticized on methodological grounds (summarized by, e.g., Anderson, Burnham, & Thompson, 2000; Kaufman, 1998; Rozeboom, 1997; Schmidt, 1996; Ziliak & McCloskey, 2008). At the same time, alternative quantitative methods have made NHST largely redundant (e.g., Diaconis & Efron, 1983; Gatsonis et al., 2001; Johnson, 1999; Kline, 2004; Schwab & Starbuck, 2009). Yet, despite this controversy, NHST continues to be widely used in organization studies (Seth, Carlson, Hatfield, & Lan, 2009). In this context, *organization studies* (OS) refers to all management-related journals and disciplines, including but not limited to organizational behavior, strategy, human resource management, and organization theory. The question now is why organizational researchers have remained so committed to NHST and what viable alternatives to NHST exist.

[1] Penn State University, Altoona, PA, USA

Corresponding Author:
Marc Orlitzky, Penn State University Altoona, 214 Hawthorn, 3000 Ivyside Drive, Altoona, PA 16601
E-mail: moo3@psu.edu

In light of the longevity of NHST, the present article has two objectives. First, and more narrowly, this article summarizes the main problems associated with NHST. The continuing widespread use of NHST, despite all the rational, methodological critiques of this theory-testing procedure, suggests that NHST has become a firmly entrenched ritual in the quantitative social sciences (Gigerenzer, Krauss, & Vitouch, 2004). Second, and more important, this article integrates the literature on the philosophy of science with insights from neoinstitutional theory in order to recommend strategies and change efforts helpful for the deinstitutionalization of NHST. *Deinstitutionalization* refers to the erosion or discontinuity of an institutionalized activity or practice (Oliver, 1992). Similar to other analyses of OS activity through an institutional lens (e.g., Abrahamson, 1996), the procedure of NHST is reinterpreted as an institution-wide manifestation of social myth and ceremony (see also Carver, 1978; Malgady, 2000). Without such an institutional and sociological perspective, problems affecting entire social systems, such as scientific communities, remain intractable (Colbert, 2004; Daft & Weick, 1984). Overall, the argument of this article is based on the premise that, without an in-depth analysis of the philosophical issues presented by the epistemology and the sociology of OS knowledge, methodological change is unlikely (Meehl, 1997). Hence, the purpose of this article is to highlight the improbability of fundamental individual-level change in the use of NHST as long as there is no institution-wide consensus about the necessity of broader epistemological reforms.

So far, the persistent critiques of NHST have been underpinned by the assumption that the *individual* researchers' actions are the crux of the problem (e.g., Gigerenzer, 1993; Schmidt, 1996; Schwab & Starbuck, 2009). Expressed differently (in the words of a reviewer commenting on a prior version of this article), "Doing good research is primarily based on the skills and judgment of individual researchers." Conversely, in their normative judgments about "bad" research, observers often blame problematic outcomes on *individual* researchers' misjudgments and mistakes.[2] To be sure, these microlevel assumptions about error-prone decision making by individual OS researchers are correct to some extent, but also incomplete. They fail to take into account the broader structural forces affecting the nature and quality of extant research practice. For this reason, methodological individualism (in both senses of the term[3]) may miss macrolevel opportunities for institutional reform by ignoring the collective-level processes and structure of the social sciences (including OS).

Turning OS knowledge about the preconditions of institutional legitimacy and processes of deinstitutionalization into prescriptions for possible reforms of quantitative research, this article proceeds as follows. First, it briefly reviews the conceptual and methodological shortcomings of NHST. Second, several epistemological reforms are proposed to deinstitutionalize NHST. Such institutional reforms have not yet been initiated in OS, arguably because macrolevel insights from neoinstitutional theory and the sociology of science have not yet been applied to the NHST controversy.

Problems With NHST—A Brief Overview

Over the years, researchers have identified numerous problems associated with NHST (e.g., J. Cohen, 1990, 1994; Guttman, 1985; Harlow, Mulaik, & Steiger, 1997; Kline, 2004; Morrison & Henkel, 1970; Nickerson, 2000; Schmidt, 1996; Ziliak & McCloskey, 2008). Because of these widely cited critiques, it is already fairly well known how these problems may lead to erroneous empirical conclusions. So, the review of the methodological problems associated with the use of NHST will be relatively brief in this article. To gain a more comprehensive and detailed understanding of the problems, quantitative OS researchers are encouraged to consult the more detailed methodological critiques of NHST cited above as well as the articles by Rodgers (2010), Seth et al. (2009), and Schwab and Starbuck (2009).

Several problems associated with NHST are enduring misapplications or misinterpretations of the technique; only three of them will be introduced in the following paragraphs. First, NHST does not tell us what we really want to know (Kline, 2004). What we want to know is whether the null

hypothesis is true given the data. In contrast, NHST indicates the (inverse) probability that the data could have been obtained if the null hypothesis were true (J. Cohen, 1994). Expressed more formally, it is a fallacy to believe that obtaining data in a region whose conditional probability under a given hypothesis is low implies that the conditioning hypothesis itself is improbable (Falk & Greenbaum, 1995; Gigerenzer, 1993). J. Cohen (1994) argues that, because of this fallacy, NHST lulls quantitative researchers into a false sense of epistemic certainty by leaving them with the "illusion of attaining improbability" (p. 998). Most important, from a theoretical perspective, statistically significant findings cannot be considered support for the alternative hypothesis, which is usually the research hypothesis of interest (Meehl, 1990).

Second, except for a strictly delimited set of circumstances (Cortina & Folger, 1998), the failure to reject the null hypothesis does not provide support for the null hypothesis, either. In almost all cases, failing to reject the null implies inconclusive results. Yet, empirical research shows that most researchers—even those well trained in statistics—falsely equate acceptance of the null hypothesis and failure to reject the null hypothesis (see Armstrong, 2007; Falk & Greenbaum, 1995; Haller & Krauss, 2002; Oakes, 1986).

Third, social scientists are generally most interested in identifying practically important, rather than statistically significant, effect sizes (McCloskey & Ziliak, 1996; Thompson, 2002). However, NHST is unable to address this issue of substantive significance (Bakan, 1966; J. Cohen, 1994; Gujarati, 1988; Schmidt, 1996). The probability level p in any particular significance test is, contrary to widely held beliefs (see Seth et al., 2009; Ziliak & McCloskey, 2008), no indicator of the practical importance of a finding.

Another major problem inherent in NHST is revealed when the underlying deductive logic of significance testing is examined more closely: NHST is logically invalid. Consider first the valid syllogism of *modus tollens:*

A1: If P, then Q.
A2: Not Q.
C (conclusion validly derived from premises A1 and A2): Hence, not P.

However, when this syllogism is probabilistic, as it is in the context of NHST, it is logically invalid (J. Cohen, 1994; Hacking, 1965; Royall, 1997; Sober, 2005):

A1: If the null hypothesis H_0 is true, then we will be unlikely to observe data pattern D_a.
A2: Data pattern D_a was observed. (That is, rejected Q of the null hypothesis based on observations.)
C: Hence, the null hypothesis H_0 is probably false and alternative hypothesis H_a is probably true. (That is, rejecting P with confidence level p.)

This probabilistic reformulation of *modus tollens* allows for the possibility of C to be false even if premises A1 and A2 are true (Hofmann, 2002), thus violating formal deductive logic (which posits that C *must* be true when A1 and A2 are.)

Beyond its methodological and logical problems, NHST has arguably had a detrimental impact on the quality of OS (see, e.g., Kline, 2004; Schmidt, 1992, 1996; Ziliak & McCloskey, 2008). The main negative consequences of NHST are that it may impede the growth of knowledge, discourage study replication, and mechanize researcher decision making in OS (Orlitzky, 2011b). These and other harmful outcomes can no longer be ignored, as already acknowledged in the context of strategy and management research (Schwab & Starbuck, 2009; Seth et al., 2009). At a minimum, when a given method becomes controversial, serious questions about knowledge growth emerge because the

accumulation of knowledge necessarily depends on the quality of methods that communities of researchers use to discover, explain, or predict phenomena.

Despite ongoing debates about the legitimacy of this statistical technique, NHST is endemic in OS. The widespread use of NHST is partly due to the remarkable increase in quantitative studies, whose proportion has approximately doubled between the mid-1950s and mid-1990s (Van Maanen, 1998). In addition, even within quantitative studies in top business journals, the rate of NHST increased considerably between 1974 and 1989 (Hubbard & Armstrong, 1992). More specifically, if we assume that quantitative articles constitute 83% of all empirical articles in premier management journals and if we further assume that 95% of quantitative organizational researchers rely on NHST to test their hypotheses (Kline, 2004, p. 9), then 79% of all empirical articles published in academic management journals use the NHST in one form or another (e.g., in the context of chi-square tests, *t* tests, *F* tests, or goodness-of-fit tests).[4] This estimate of 79% is a conservative estimate because, according to other calculations, NHST is used in 94% of all articles in the *Journal of Applied Psychology* (Hubbard, Parsa, & Luthy, 1997). Similarly, in economics, reliance on NHST has actually increased rather than decreased after McCloskey and Ziliak's (1996) critique of the prevalence of NHST in the *American Economic Review* (Ziliak & McCloskey, 2008).

In management journals, reliance on NHST is even more common. For example, Schwab and Starbuck (2009) estimated that 100% of research published by *Academy of Management (AMJ)* in 2008 used NHST. For this article, I reviewed all *AMJ* studies published in 2010 and reached the same conclusion. Albeit, my own review of *AMJ* highlights an interesting finding: In the interpretation of results, researchers who conducted path analyses, while consistently using NHST, tended to eschew an emphasis on *p* levels in favor of a focus on effect size magnitudes. This observation mirrors Rodgers's (2010) conclusions. Overall, however, these estimates suggest that, consistent with reviews in psychology (Krueger, 2001; Nickerson, 2000), it is difficult to find another theory-testing procedure more widely used and abused in OS journals today (see also Seth et al., 2009). Hence, as is true in economics, quantitative studies in OS routinely apply NHST.

Even though "null hypothesis testing should be dead" (Rindskopf, 1997, p. 319) after statisticians' strident attacks on NHST, the premier publication outlets in the social sciences do not seem to have implemented any substantive reforms (Fidler, Cumming, Burgman, & Thomason, 2004; Finch, Cumming, & Thomason, 2001). If NHST is "surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students" (Rozeboom, 1997, p. 335), what is at issue now is not so much the institutionalization of NHST (on the history of NHST, see Ziliak & McCloskey, 2008) but the failure of its deinstitutionalization in the conduct of the organizational and, more broadly, social sciences.

Toward a Deinstitutionalization of the Significance Test

Understanding the institutional persistence of NHST requires an analytic approach that integrates insights from the philosophy of science and neoinstitutional theory and, therefore, is broader than the rather narrow methodological perspective that predominated in prior critiques of NHST. For one, NHST also needs to be understood from an *epistemological* perspective, which is important work initiated by Meehl (1997) and Rozeboom (1997). In addition, a broader *institutional* understanding can provide important insights. Because scientific activity encompasses more than the aggregation of individual researchers' work (Longino, 1990), a social systems view could be useful to capture institutional forces and surface the broader structural sources of problems that frequently go unnoticed (Pearce, 2004). In a nutshell, every science, including OS, is a social process, which may be affected by the same psychological, historical, and institutional dynamics as other social processes (Abrahamson, 1996; Barber, 1971; Schofer, 2004; Tsoukas & Knudsen, 2003). This view of science as a social or institutional arena is consistent with Kuhn's (1996) *Structure of Scientific Revolutions,*

a seminal account of scientific activity, as well as many other studies of science (e.g., Barnes, Bloor, & Henry, 1996; Bloor, 1991; Brown & Malone, 2004; Knorr-Cetina, 1999; Latour & Woolgar, 1986; Lynch, 1993; MacKenzie, 1981; Merton, 1957; Popper, 1962, p. 217; Schaffer, 1986). Regarding NHST as an institution is instructive because this conceptualization may be helpful for solving the riddle of NHST's longevity despite the sustained controversy surrounding it.

By now, NHST has become such a routine procedure that it has been likened to a social ritual (e.g., by Gigerenzer, 1998; Labovitz, 1972). Specifically, NHST is a rule system, or institution, with which quantitative researchers have come to classify hypotheses as true or false. Institutions are defined as "socially constructed, routine-reproduced . . . program or rule systems" (Jepperson, 1991, p. 149). An institutional view of NHST is consistent with W. R. Scott's (1994) well-known typology, which conceptualizes institutions as composed of the following elements:

1. "meaning systems and related behavior patterns [i.e., the procedure of statistical significance testing itself], which contain"
2. "symbolic elements, including representational, constitutive and normative components [*t* tests, *F* tests, chi-square tests, etc.], that are"
3. "enforced by regulatory processes [review procedures of journals and book publishers]" (p. 56).

In other words, institutions rest on three pillars: *cultural-cognitive, normative,* and *regulative* (W. R. Scott, 2008). Pointing to developments and forces within each of these three subsystems, a set of different proposals can be developed that may be effective for the deinstitutionalization of NHST and, therefore, build on and extend sociological theories of deinstitutionalization. If the reader and broader OS community agreed with my negative assessment of NHST, adopting these proposals would be beneficial for accelerating the abandonment of NHST. If, however, the reader agreed with the defenders of NHST (e.g., Abelson, 1997; Hagen, 1997; Harris, 1997; Hoover & Siegler, 2008; Mulaik, Raju, & Harshman, 1997; Wainer, 1999), the proposals herein could serve as a platform from which to prevent its deinstitutionalization and, thus, preserve the methodological status quo. This article proceeds by analyzing NHST in terms of these three elements and developing strategies for dismantling or adjusting the nature of these elements within OS.

Weaning Researchers Off Their Cultural-Cognitive Devotion to NHST

Critique of the epistemological status quo. NHST is the key statistical technique that puts into practice hypothetico-deductivism (Chow, 1998b), the scientific inference procedure based on Popper's falsifiability criterion and advocated by, among others, Hempel (1965) and Stinchcombe.[5] According to Stinchcombe (1968), scientific inference starts with general (i.e., theoretical) knowledge claims in which one class of phenomena is linked to another. Then, "by logical deduction and by operational definitions of the concepts, an empirical statement" is derived (Stinchcombe, 1968, p. 16). Stinchcombe's *empirical statements* are commonly called *hypotheses,* "statements that . . . predict a relationship between two or more indicators and . . . *can be true or false* [emphasis added]" (B. P. Cohen, 1989, p. 141). The centrality of this falsifiability criterion (see also Popper, 1969, 1972) implies that an objective, replicable classification procedure is needed for the sorting of hypotheses into *true* and *false* statements. For dichotomous (true/false) researcher decisions, NHST is conventionally considered a natural sorting mechanism from the perspective of hypothetico-deductivism (Abelson, 1997; Chow, 1998a, 2000).

So, as long as the OS community is devoted to hypothetico-deductivism, NHST will likely persist. Therefore, any cultural-cognitive reform effort to deinstitutionalize NHST should probably also be concerned with broader epistemological change in OS. At first, such a cultural-cognitive change might seem revolutionary, but an examination of the philosophy of science literature reveals it is not.

Instead, what is radical is the prevailing mainstream embrace of deductive reasoning in OS. For instance, one type of deductivism, restrictive deductivism, unreasonably implies that theories can be rejected when a single null hypothesis (with the alternative hypothesis being central to that theory) cannot be refuted (Chalmers, 1999; B. P. Cohen, 1989; Goldstone, 2004, pp. 47-48). In other words, it postulates that the testing of a single hypothesis can determine the fate of an entire theory. However, such a belief in crucial experiments or the one "perfect study" (Hunter & Schmidt, 2004, p. 17), although common, is clearly unwarranted (B. P. Cohen, 1989; J. Cohen, 1965).

A more sophisticated epistemological framework, holistic deductivism (Burawoy, 1990; Lakatos, 1970), shifts the emphasis to entire research programs in which a set of fundamental axioms are unquestioned (e.g., "All human behavior is caused by self-interest"). Around this theoretical core, auxiliary hypotheses form a protective belt to account for anomalies that might contradict postulates in the core (e.g., observations of altruism and self-sacrifice). Holistic deductivism implies that the goal of any (social, organizational, or other) science is to produce progressive rather than degenerating research programs:

> In a *progressive* program the new belts of theory expand the empirical content of the program, not only by absorbing anomalies but by making predictions, some of which are corroborated. In a *degenerating* program successive belts are only backward looking, patching up anomalies in an *ad hoc* fashion, by reducing the scope of the theory, or by simply barring counter-examples. (Burawoy, 1990, p. 778)

Because distinctions between proper, or progressive, and improper, or degenerating, protective belts are ultimately subjective (Gorski, 2004, p. 12), the entire framework of hypothetico-deductivism loses its epistemological appeal (see also Hanson, 1965). Stated succinctly, "One of [the] worst-kept secrets" in the social sciences is that hypothetico-deductivism "does not work" (Gorski, 2004, p. 28). Fortunately, other forms of scientific inference have already been advocated successfully in OS and are presented next: inductive reasoning, abduction, and statistical and mathematical modeling.

Inductive reasoning. Inductive knowledge growth turns deductivism—and the current structure of journal articles—on its head: Observations of particular data are used to infer causal relationships between phenomena or scientific laws. Championed by Aristotle and Bacon and successfully applied by, for example, Newton, Galileo, and Darwin (Locke, 2007, pp. 870-872), inductivism can, contrary to conventional assumptions in the social and organizational sciences, be shown to be a more productive mode of scientific inquiry than hypothetico-deductivism. Locke makes a convincing case for induction by pointing out that some of the most successful theories in the social sciences and OS have been developed inductively rather than deductively (Locke, 2007, pp. 873-879), including goal-setting theory, the theory voted *most important* in OS (Miner, 2003). The disdain in which induction is still held by most quantitative researchers is unjustified (Hambrick, 2007; Locke, 2007). Therefore, although induction is the name of the game among most qualitative research traditions (Creswell, 1998; Eisenhardt, 1989), it should also be taken more seriously by quantitative researchers. The shift from deductivism to inductivism would change the focus from dichotomous thinking (à la *something vs. nothing* or *true vs. false* in NHST) to a focus on the magnitude of effect sizes (Kline, 2004; Schmidt, 1996). The case for such an epistemological shift is strong, particularly for problems of high cognitive complexity (Arthur, 1994), arguably a characteristic of many research questions in OS as well as in the other social sciences. Such a change in scientific reasoning is particularly important in light of Hambrick's (2007) implication that an increase in inductive OS research could stimulate the growth of knowledge with greater practical value—an important criterion for an applied science such as OS.

Abduction. Another method of inquiry just as legitimate as inductive reasoning is abduction (Haig, 2005a; Rozeboom, 1997), which "involves reasoning from presumed effects to underlying cause" and "educated guesses constrained by relevant knowledge in the domains under investigation" (Haig, 2000, p. 293). As discussed in depth by Haig (2005b), exploratory factor analysis and other types of exploratory data analysis represent examples of abductive reasoning (see also Rozeboom, 1961; Stephenson, 1961; Tukey, 1977). In addition, computer-intensive resampling can be regarded as an important abductive methodology (Haig, 2005a; Kline, 2004).

The potentially most convincing example of abduction may be meta-analysis. This statistical technique, which quantitatively summarizes and integrates primary studies more rigorously than narrative literature reviews do, helps researchers identify the extent to which empirical regularities are robust (H. M. Cooper & Hedges, 1994; Hunt, 1997; Schmidt, 1992). One of the ways in which meta-analysis is abductive is explained by its goals: Abductive methods aid in the detection of phenomena, which are "relatively stable, recurrent, general features of the world" that researchers seek to explain (Haig, 2005a, p. 374). Contrary to widely held assumptions, "phenomena are not, in general, observable; they are abstractions wrought from the relevant data, frequently as a result of a reductive process of data analysis" (Haig, 2005a, p. 374), such as proffered, in an exemplary manner, by meta-analysis. Data, on the other hand, are "recordings or reports that are perceptually accessible; they are observable and open to public inspection" (Haig, 2005a, p. 374). In the results of many meta-analyses (particularly if they have been done carefully), this important distinction between phenomena and observed data (see also Bogen & Woodward, 1988; Woodward, 1989) is illustrated by the effect size regularities reported as corrected, or true score, correlation coefficients (ρ), as opposed to a focus on observed correlation coefficients (r_{obs}).

An example may be helpful to illustrate, more specifically, the way in which meta-analysis applies and reconstructs the following general schema of abductive reasoning:

> The surprising empirical phenomenon, P, is detected.
> But if hypothesis H were approximately true, and the relevant auxiliary knowledge, A, was invoked, then P would follow as a matter of course.
> Hence, there are grounds for judging H to be initially plausible and worthy of further pursuit. (Haig, 2005a, p. 377)

For example, an influential[6] meta-analysis (Orlitzky, Schmidt, & Rynes, 2003), which did not rely on NHST, rejected the conventional conclusion of observed irregularities in the data (e.g., Griffin & Mahon, 1997; Ullmann, 1985) and suggested the following avenues for further research:

> The surprising empirical phenomenon of regularities (of positive average ρ) between corporate social and financial performance is identified.
> If corporate social performance helps build corporate legitimacy and reputation (Mahon, 2002; Orlitzky, 2001; Waddock & Graves, 1997) and thus reduces business risk (Godfrey, 2005; Orlitzky & Benjamin, 2001), and if we can have some confidence in the reliability and validity of measures of corporate social performance (Orlitzky et al., 2003), then the observed phenomenon would follow as a matter of course.
> Hence, there are grounds for judging the risk-reputation hypothesis to be initially plausible and worthy of further pursuit.

Thus, the focus in this meta-analysis was not on deductive theory testing but instead on identifying the plausibility of different causal mechanisms, which are suggested by the phenomena behind the data (see also Orlitzky, 2006, 2008; Orlitzky & Benjamin, 2001).

To be sure, not all meta-analyses are abductive—or presented as such. Although many meta-analysts, reverting to the conventional ''some-correlation'' versus ''no-correlation'' reasoning implicit in NHST, present and frame their integrative studies as binary tests of theory,[7] this conceptualization of meta-analysis misrepresents the true aims of quantitative research syntheses (Hunter, 1998). Extracting ''a signal (the phenomenon) from a sea of noise (the data)'' (Haig, 2005a, p. 374), rather than theory testing, is ultimately the overarching purpose of meta-analysis (H. M. Cooper & Hedges, 1994; Hunt, 1997; Kline, 2004; Rosenthal, 1995; Schmidt, 1992). Meta-analytic integration of effect sizes can, most accurately, be regarded as a precursor of particular theory tests that *may* be conducted, for example, in the form of path analyses (e.g., see Hunter & Gerbing, 1982), *after* a meta-analytic research synthesis (Schmidt, 1992; Schmidt, Hunter, & Outerbridge, 1986). From the perspective of theory testing rather than phenomena detection, these multivariate causal analyses that follow up the meta-analytic integration proper are needed to avoid specification bias (Gujarati, 1988) or, more broadly, errors of the third kind (Kimball, 1957; Mitroff & Silvers, 2009). Although abduction goes beyond the description of idiosyncratic data espoused by inductivists, and at the same time eschews the amethodological formalism of hypothetico-deductivism, it has its own set of limitations, only two of which can be mentioned here. First, abduction runs the risk of generating only rudimentary theory (see, e.g., commentary on theory in Orlitzky, Siegel, & Waldman, 2011)—in contrast to the logical tightness and parsimony of theory underpinned by other epistemologies. Second, abduction introduces a complex and, arguably, rather subjective set of theory appraisal criteria (Haig, 2005a, pp. 380-382), which run counter to the clear, straightforward criterion of predictive success in hypothetico-deductivism (Friedman, 1953).

Statistical and mathematical modeling. In an interesting article, Rodgers (2010) argues that statistical and mathematical modeling circumvents the NHST problem by avoiding the simplistic and mechanistic decision making of the traditional binary NHST logic. Comparing observations to the current model rather than focusing on the nil hypothesis, model fitting (e.g., via structural equation modeling) represents a nearly perfect instantiation of Platt's (1964) important advice to all scientists, namely, that comparing the empirical validity of several alternative hypotheses (or ''models'') advances science faster than the conventional ''there-is-probably-not-nothing'' NHST reasoning (Dawes, 1991, p. 252; Hofmann, 2002, p. 70).

At the same time, though, what Rodgers (2010, p. 3) characterizes as a ''quiet methodological revolution'' does not entirely solve the problem of NHST because, within most modeling efforts, NHST is still alive and well. Properly conceived, modeling develops plausible causal processes and structures that are, for example, coherent, elegant, and parsimonious (Abelson, 1995, p. 14; Rodgers, 2010, p. 10) and compares those attributes with competing models. However, *in practice,* these comparisons retain NHST in testing parameter estimates (Rodgers, 2010, p. 10, footnote 9) and also in overall goodness-of-fit tests, many of which are based on chi-square tests and notorious for their low statistical power (see, e.g., Overall, 1980; Steele, Hurst, & Chaseling, 2008). Thus, in many studies using mathematical and statistical models, researchers may falsely conclude that a nonsignificant value of chi square indicates good model fit. Notably, such apparent ''model fit'' may be produced by a Type II error rather than by the verisimilitude of the underlying model. This is particularly problematic because, in day-to-day research practice, alternative models are far too often chance or nil models—probably because, unlike natural scientists, most social scientists are not properly trained in the epistemology of Platt's (1964) strong inference and the vital epistemological turn discussed by Rodgers (2010, pp. 7-10).

In sum, model building is, undoubtedly, an appropriate short-term first step in ameliorating the usage of NHST. However, statistical and mathematical modeling remains embedded in the epistemology of hypothetico-deductivism, which arguably gave rise to NHST in the first place (Chow, 1998b, 2000; Neyman, 1957; Ziliak & McCloskey, 2008), and does not emphasize fact finding quite

as much as the epistemologies of induction or abduction. As Rodgers (2010) perceptively pointed out, "The modeling perspective subsumes NHST" (p. 7)—it does not abandon and replace it. This arguably evolutionary (see, in particular, Rodgers, 2010, p. 10, footnote 9) rather than revolutionary approach to methodological change, though, is not only a long-term weakness but also a strength in the short run because this continuity with methodological tradition meets the standard of comprehensibility (Suchman, 1995), an important aspect in any institutional change effort, and avoids the specter of radical epistemological change, which will in turn diminish quantitative researchers' resistance to it.

Legitimacy of NHST and alternative techniques. The discussion above, which focused on the cultural-cognitive institutional pillar of epistemology, can be summarized as follows. First, the legitimacy of the epistemological foundation of NHST has come under attack. Second, alternative epistemologies have already been identified and advocated. Unfortunately, many social and organizational scientists are still unaware of the advantages of these alternative epistemologies—or reject them outright in quantitative science (see Locke, 2007; Rozeboom, 1997). Hence, another crucial step is to enhance the legitimacy of concrete data analytic techniques that are alternatives to NHST. In the institutional OS literature, legitimacy is defined as a "generalized perception or assumption that the actions of an entity are desirable, proper, or appropriate within some socially constructed system of norms, values, beliefs, and definitions" (Suchman, 1995, p. 574).

Changing Scientific Values and Norms in OS Researchers' Methods Training

The argument so far implies that, in order to effect institutional change, the normative consensus about NHST needs to be questioned and ultimately delegitimized. That is, researchers must feel compelled to abandon NHST by a sense of social obligation that is underpinned by a change in *norms,* which are the definitions of the legitimate means to pursue valued ends, and a change in *values,* which are "conceptions of the preferred or the desirable, together with the construction of standards to which existing . . . behaviors can be compared" (W. R. Scott, 2008, pp. 54-55). In this context, it is important to examine social and organizational scientists' internalized values and norms, especially because in any effort to prevent institutional change normative vocabularies are often invoked to preserve the status quo (Suddaby & Greenwood, 2005).

Hence, a major challenge for critics of NHST is to identify data analysis alternatives to NHST that are perceived as legitimate. Due to their novelty and focus on empirical uncertainty (in the form of, e.g., sampling error, measurement error, model misspecification, error of the third type, etc.), many of the alternatives to NHST currently lack legitimacy compared with the well-established NHST doctrine. Contemporary researchers have not sufficiently been trained to recognize the problems inherent in NHST and adopt norms and values more supportive of methodological alternatives. Thus, because methods training in graduate school can be considered the primary way in which aspiring researchers are socialized to the norms and values of legitimate quantitative research (Deem & Brehony, 2000; Hakala, 2009; Mendoza, 2007; Schmidt, 1996), the next section is a comprehensive discussion of the reforms in doctoral training needed for a deinstitutionalization of NHST.

Point estimates and confidence intervals. Many methods experts recommend, as an alternative to NHST, point estimates of effect sizes with confidence intervals around them as the proper expression of the empirical uncertainty of these effects (e.g., Brandstaetter, 2001; Dixon, 2003; Hofmann, 2002; Jones & Tukey, 2000; Kline, 2004; Masson & Loftus, 2003; McGrath, 1998; Meehl, 1997; Schmidt, 1996; Schwab & Starbuck, 2009; Tryon, 2001). However, current courses in research methodology, unfortunately, still focus on the calculation and reporting of the crude, binary results of NHST. A shift from NHST to point estimates and confidence intervals would de-emphasize nil-

hypothesis comparisons and reliance on *p* values in favor of a more productive focus on affirmative predictions of different models, similar to the pedagogical approach advocated by Rodgers (2010, pp. 9-10). Consistent with the epistemologies of induction and abduction, fact finding as well as proper acknowledgment of empirical uncertainty presented by sampling error (typically expressed as the standard error) and measurement error would be core and center in such methods training.

Some defenders of NHST (e.g., Abelson, 1997) present the counterargument that confidence intervals fail to respond adequately to the cognitive demands on researcher information processing. According to Abelson, researchers require categorical markers for distinguishing between important findings and trivial findings. Probability values *p* can and, according to Abelson, ought to be used to establish the credibility of an empirical finding because *p* provides an estimate of the likelihood of only chance having produced the result (Abelson, 1995). In other words, efficiency in information processing and communication requires a categorical filter of "importance" against the (chance) nil hypothesis.

This defense of NHST warrants closer scrutiny. First, in almost all theory contexts NHST is no valid marker of importance, as already mentioned (Bakan, 1966; J. Cohen, 1994; Schmidt, 1996; Seth et al., 2009; Ziliak & McCloskey, 2008). Second, because confidence intervals can be interpreted as equivalent to NHST (Bedeian, Sturman, & Streiner, 2009; Cortina & Dunlap, 1997), it is unclear how NHST passes Abelson's (1997) cognitive hurdle, but confidence intervals do not. Third, this interpretational equivalence also raises some questions about the claim that confidence intervals represent a genuine alternative to NHST. Admittedly, for every misuse of NHST, there is an analogous misuse and misinterpretation of confidence intervals (Abelson, 1997). So, like the epistemologies of mathematical and statistical modeling, data analysis based on point estimates with confidence intervals around them is an appropriate first step to be taken but does not really address the fundamental problems of NHST in the long run.

Most important, the fact that confidence intervals *can* be interpreted as equivalent to NHST does not imply they inevitably *must* be (Kline, 2004, p. 80). For example, until the 1930s, "probable errors" (i.e., 50% confidence intervals) were often reported but were not interpreted as significance tests (Schmidt, 1996, p. 124). Point estimates and confidence intervals provide more detailed and, therefore, useful information about the magnitude and variance of effect sizes—especially across studies (Hunter & Schmidt, 2004; Kline, 2004; Schmidt, 1992)—than NHST does, which invariably reduces the evidence to binary *true/false* conclusions (Reichardt & Gollob, 1997). More generally, emphasis on point estimates and confidence intervals would further increase the interpretive affinity between the physical sciences and the social sciences (Hedges, 1987). However, what is presently unclear is whether this shift to physical science values and norms, which favor information more complex than that proffered by NHST, fits the self-constructed identities of social scientists (see also Townley, 2002); it certainly does not fit all of them (see also Burrell, 1996; Tsoukas & Knudsen, 2003). Thus, the next suggestion may be more consistent with the self-understandings of many social scientists.

Shift from objective to subjective probabilities. Traditionally, objectivity is upheld as one of the highest values in science (Chalmers, 1999; Kincaid, 1996). For science to be valued internally and externally, it is usually held, "nonideological modes of observing, or 'objectivity'" must be institutionalized (Fuchs, 2001, p. 34). Because results of NHST are often (mis)interpreted as indicators of a study's replicability or importance of effect size (J. Cohen, 1994; Oakes, 1986; Schmidt, 1996), NHST becomes a signifier of, and shorthand for, scientific objectivity. The "sizeless stare of statistical significance" (Ziliak & McCloskey, 2008, p. 33) may increase the appearance that social science research is nonideological and value free and, thus, conforms to the objectivist canon. Yet, contradicting this mirage of objectivity seemingly created by NHST, many physical scientists regard NHST as unscientific (Schmidt & Hunter, 1997). Thus, future researchers must be disabused of the

idea that the objective "reality" of an important finding hinges on the outcome of a statistical significance test (J. Cohen, 1994; Ziliak & McCloskey, 2008).

To enhance the legitimacy of NHST alternatives, methods training should clearly and explicitly unmask as illusory the belief that NHST is some deus ex machina for instilling objectivity and factuality to researcher observations. Rather than embracing objectivism and denying social construction in social science research, graduate training could move toward subjectivist and intersubjectivist norms and values. Most importantly, in the long run, this would involve a shift toward a Bayesian view of probability (Trafimow, 2003). Such a change in graduate training would direct attention to the explicit acknowledgment of researchers' subjective beliefs and prior probabilities in theory validation (Cumming, 2005; Killeen, 2005b; Krueger, 2001). In contrast to the traditional view, which brackets researcher beliefs and, thus, leads to a false sense of objectivity (Gatsonis et al., 2001; Kline, 2004), Bayesian estimation explicitly specifies researchers' expectations or degrees of belief (Chater, Tenenbaum, & Yuille, 2006). It explicitly (and more effectively than the traditional frequentist statistics used in OS today) acknowledges variability in researcher perceptions of the plausibility of given hypotheses, measurement inaccuracies, and the cumulative structure of science (Matthews, 2000; Pruzek, 1997; Rindskopf, 1997). The main reason why this particular reform effort is currently quite difficult to implement is that it requires broader normative change in OS toward values that make subjectivism and intersubjectivism in quantitative research explicit and transparent. To be sure, in select areas in computer science, economics, and medicine (see, e.g., G. F. Cooper, 1990; di Bacco, d'Amore, & Scalfari, 2004; Press, 2003), Bayesian approaches have already replaced the traditional frequentist view of probability, which is the underlying logic of NHST. Also, in strategic management, several Bayesian studies have started to appear in print (e.g., J. G. Scott, 2009; Tang & Liou, 2009).

Training in Bayesian statistics raises a number of complex technical issues (see, e.g., Dalton & Dalton, 2008, pp. 133-134; Gatsonis et al., 2001; Gigerenzer & Hoffrage, 1999; Howard, Maxwell, & Fleming, 2000; Killeen, 2006; Lewis & Keren, 1999; Press, 2005; Schmidt, 2008, pp. 107-111; Steel & Kammeyer-Mueller, 2008). So, to avoid the impression that Bayesian statistics is "just another" tool set, effective pedagogy will have to start with a foundational, relatively easy-to-understand text such as, for instance, Bovens and Hartmann's (2004) *Bayesian Epistemology* or, alternatively, Howson and Urbach's (2005) *Scientific Reasoning: The Bayesian Approach*. Both books introduce students to the fundamental change in scientific discovery required for Bayesian interpretations. Then, two excellent articles introducing Bayesian inference could be assigned (i.e., Pruzek, 1997; Rindskopf, 1997). To emphasize the Bayesian view of probability as a subjective degree of belief about unknown parameters, Bayes's theorem, which provides the probability that the hypothesis is true given the data, that is, $p(H|D)$, could be introduced—with all its wide-ranging implications for epistemology and data analysis:

$$p(H|D) = \frac{p(H)p(D|H)}{p(D)}; \quad \text{where}$$

- $p(H|D)$ is the *posterior* probability of the hypothesis given the data, replacing the erroneous focus on $p(D|H)$ implicit in conventional significance testing (J. Cohen, 1994);
- $p(H)$ is the (*prior*) probability of the hypothesis before the data are collected;
- $p(D)$ is the (*prior*) probability of the data irrespective of the truth of the hypothesis; and
- $p(D|H)$ is the *conditional* probability of the data under the hypothesis, which is, of course, analogous to the p value in NHST if $H = H_0$.

Although frequentist methods are easy to identify in today's top OS journals, applications of Bayesian methods, though increasing, are unfortunately still too sparse to provide aspiring Bayesian

researchers with many exemplars (see Huff, 1999, on the importance of exemplars in research training). However, the use of some excellent methods textbooks could relatively easily circumvent this problem. The intricacies of Bayesian data analysis could be presented through the use of, for example, Press's (2003) *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications,* which garnered very favorable reviews (Zellner, 2009); Gelman, Carlin, Stern, and Rubin's (1995) *Bayesian Data Analysis;* or Zellner's (1996) *Introduction to Bayesian Inference in Econometrics*.

Triangulation. More broadly, inculcating the values of triangulation in graduate training would most likely shift emphasis from NHST to alternative and frequently complementary data-analytic techniques. Triangulation involves the use of multiple study designs, multiple data collection and analysis techniques, multiple research settings, or most generally, multiple research strategies in examining a given research question (Sackett & Larson, 1990; Scandura & Williams, 2000; Webb, Campbell, Schwartz, Sechrest, & Grove, 1981). By embracing triangulation and de-emphasizing the value and epistemic certainty of single studies (Nelson, Megill, & McCloskey, 1987, p. 8), researchers would realize the necessity of establishing the verisimilitude of findings over a long series of constructive (rather than literal or operational) replications (Lykken, 1968; Tsang & Kwan, 1999). Because, as Scandura and Williams (2000) pointed out, OS has so far insufficiently relied on triangulation and, therefore, does not reach its full potential in terms of data interpretation, it is no surprise that NHST has remained prevalent in OS. Hence, more efforts should be invested in emphasizing, in all graduate training, the value of approximating parameters incrementally. Over time, with such a normative commitment to constructive replications, a particular finding would be considered robust if and only if several follow-up studies (changing research settings or designs) showed similar magnitudes in effect size (such as *r* or *d*).

In the long run, researchers convinced of the value of triangulation would, metaphorically speaking, learn to "use multiple reference points to locate an object's exact position" (Jick, 1979, p. 602). That is, in triangulating research the focus is on maximizing the diversity of methodologies applied to examine a given theory. In contrast, meta-analysts can work with data obtained not only in constructive replications but also in literal and operational replications. Indeed, some researchers (e.g., Algera, Jansen, Roe, & Vijn, 1984) regard the results of a meta-analysis as more credible if the cross-study operational definitions and measures are as similar as possible, which is, of course, different from the intent of triangulation. Although triangulation and meta-analysis are both concerned with replicability, their approaches to external validation are slightly different: Meta-analysis focuses on effect sizes (such as *r* or *d*) and their cross-study generalizability, whereas triangulation focuses on widening the epistemological lens. For example, triangulation studies that add a qualitative or interpretive angle to a traditionally quantitative line of inquiry may demonstrate the substantive impact of researchers' epistemological priors on research conclusions. The extent to which meta-analyses can accomplish the same objective is limited (see, e.g., Orlitzky, 2011a). These subtle differences in procedures and orientations explain why triangulation and meta-analysis may contribute to the deinstitutionalization of NHST in slightly different ways.

Whether triangulation (or meta-analysis) can, in the long run, really help eliminate NHST from social science practice will primarily depend on future theory development in OS. Currently, OS is typically characterized by relatively weak theory (Hannan, Polos, & Carroll, 2007; Meehl, 1991; Pfeffer, 1993), which (among other defects) portrays reality as binary: Some causal effect either exists or does not exist. However, strong inference (Platt, 1964) requires iterative testing of alternative theoretical explanations, rather than the one-time comparison of the research hypothesis to the nil hypothesis. So, if more advanced theory cannot be used to specify different parameter values, which is a precondition, for example, for comparisons of different analytic models (Rodgers, 2010), NHST is bound to remain a bastion of theory testing in the organizational and all other social

sciences. The history of psychology seems to reaffirm this caveat. In social psychology, for example, replications have been under way for many years, yet there is certainly no waning of NHST in this field.

On the Inadequacy of Exclusive Reliance on Regulatory Reforms

Cognitive-cultural and normative institutional processes typically require regulative sanctions and incentives for maximum impact. That is, the enforcement mechanisms associated with institutions include coercive elements through the use of authority and rule setting. Especially in the context of hierarchical structures (W. R. Scott, 2008), the importance of these regulative forces should not be underestimated in any account of the persistence of an institutional practice such as NHST. However, their impact is ambivalent and uncertain, as argued below.

In the institutional context of publish-or-perish career pressures, journal editors and reviewers serve as powerful regulatory gatekeepers (Beyer, Chanove, & Fox, 1995). In a context of ontological and epistemological pluralism (Burrell, 1996; Donaldson, 1995; Pfeffer, 1993; Van Maanen, 1995), scientific value and validity generally are not naturally inherent qualities of the studies submitted to journals but are ultimately socially conveyed awards (Macdonald & Kam, 2007; Peters & Ceci, 1982). A relatively small group's (i.e., two to four anonymous reviewers' and the editor's) intersubjective judgments of appropriate methods (and theories) generally determine publishing success and, thus, professional advancement (Kuhn, 1996; Park & Gordon, 1996). Usually, the "wisdom of crowds" (Surowiecki, 2004, p. 3)—for example, in the form of citation counts—applies to the quality assessment of research only *after* a particular study has been vetted by a small group of peers. Unfortunately, though, empirical evidence suggests that, in contrast to collective judgments about research quality, the small-group peer review process seems to be fraught with errors, particularly in social and organizational science journals (Gans & Shepherd, 1994; Oswald, 2007; Peters & Ceci, 1982; Starbuck, 2005).

Currently, the consensus among the gatekeepers of OS and other social science journals seems to be that statistically nonsignificant results do not constitute valuable contributions to knowledge (Gigerenzer, 2004; Porter, 1992). Reviewers and editors often base rejections on observations of small effect sizes (Austin, Boyle, & Lualhati, 1998), which, all else equal, are less likely to be statistically significant than large effect sizes (e.g., Begg, 1994; Coursol & Wagner, 1986; McNemar, 1960). An unsurprising consequence of this convention, which equates *small* and *trivial* (J. Cohen, 1994), is that little journal space is devoted to nonsignificant findings (Cortina & Folger, 1998; Hill, 2003; Starbuck, 2007). Junior scholars, who tend to follow editors' and reviewers' dictates (Huff, 1999), are not at fault, of course. Researchers who want to appear in print and survive in a publish-or-perish environment are motivated by force, fear, and expedience to continue using NHST (Gigerenzer, 2000). As Mahoney noted in reference to NHST years ago, "Until the rules of the science game are changed, one must abide by at least some of the old rules or drop out of the game" (Mahoney, 1976, p. xiii). At present, the rules of the game are rituals enforced by many journal reviewers and editors, who assume that the mechanistic use of NHST can lend more credibility to findings because $p < .001$ results can somehow be considered more "real" or substantively important than $p < .05$ or nonsignificant results (e.g., $p = .0501$). Although this assumption is clearly false (J. Cohen, 1994; Schmidt, 1996; Schmidt & Hunter, 1997; Ziliak & McCloskey, 2008), long back-and-forth debates with reviewers and editors about the validity and scientific value of statistically nonsignificant results are often futile (Abrahamson, 2007; Starbuck, 2007).

From an institutional perspective of systemic change, regulative pressures are undoubtedly necessary because we cannot really expect an abandonment of NHST in OS if only the (prepublication) actions of individual researchers are the focus of NHST critics' advice. For the demise of NHST, journal editors must follow their NHST-averse rhetoric with concrete actions because of the

importance of shifts in power coalitions for the deinstitutionalization of social practices (Oliver, 1992). For example, although Mahoney (1976) and J. P. Campbell (1982) both privately recognized and lamented the weaknesses of NHST, during their editorships at two different psychology journals reliance on NHST continued unabated in articles in those journals (Pedhazur & Schmelkin, 1991, p. 201). So, without a change in the broader regulative insistence on NHST at OS journals, a shift toward alternatives is unlikely. How much of a difference regulative pressure can make was shown at two medical journals (*Epidemiology* and *American Journal of Public Health*) where a change in editorial regimes brought about a precipitous decline in NHST (Fidler, Cumming et al., 2004, p. 617; Shrout, 1997, p. 1). In *Epidemiology,* the shift toward greater researcher attention to effect size magnitude persisted, whereas in the *American Journal of Public Health* there was a resurgence of NHST after the respective NHST opponents had left their editorial posts at these two medical journals (Ziliak & McCloskey, 2008).

In general, OS editors and reviewers may be reluctant to implement regulative reforms because alternatives to NHST not only are currently imbued with a great deal of epistemic uncertainty but also require changes in graduate training (see the prior section on cultural-cognitive reforms). To help editors make more informed decisions about the transition to alternative statistical techniques and ultimately avoid editorial conservatism, task forces of the main scientific associations, as well as subsequent revisions of guidelines in publication manuals, could be an important first step in these regulative reform efforts (e.g., Fidler, 2002; Finch, Thomason, & Cumming, 2002). For example, the American Psychological Association (APA) created a task force to study the NHST controversy. Although some (e.g., Hunter, 1997) had hoped for a NHST ban as its outcome, the APA Task Force on Statistical Inference only went so far as to recommend the reporting of effect sizes and confidence intervals around it (Kline, 2004). The task force regarded a NHST ban as too extreme (Wilkinson, 1999), most likely because effect sizes that are not based on NHST are unavailable for quite a few research designs and statistical analyses (Fidler, 2002; Kline, 2004). The fifth edition of the APA *Publication Manual,* used by many social and organizational sciences other than psychology, adopted the task force's critical but moderate stance on NHST. The sixth edition (see APA, 2010, p. 33) preserves this conservative stance toward NHST by emphasizing the liberties and prerogatives of editors in setting journal policy regarding the statistical reporting of results.

Several years before the APA task force, in 1988, the International Committee of Medical Journal Editors (ICMJE) embarked on a more far-reaching regulatory reform effort. The ICMJE's guidelines for medical researchers were as follows:

> When possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals). Avoid relying solely on statistical hypothesis testing, such as the use of *p* values, which fail to convey important information about effect size. (Fidler, Thomason, Cumming, Finch, & Leeman, 2004, p. 120)

Over 300 medical journal editors indicated their willingness to comply with these guidelines (Ziliak & McCloskey, 2008). Nonetheless, when the ICMJE's and one important change agent's (Rothman, 1986) goals are compared to the actual research reports in medical journals, the evidence again suggests that such regulative reforms have largely failed (Fidler, Thomason et al., 2004; Ziliak & McCloskey, 2008).

Hence, the view based on the institutional perspective proposed in this article implies that simply banning NHST will, most likely, be ineffective or unrealistic. Such a ban may be ineffective because some research suggests that institutionalized processes are very difficult to delegitimize through regulatory fiat. For example, an ethnomethodological study by Zucker (1991) indicates that institutional effects are easy to create but difficult to dissolve through sanctions. When reform efforts are only regulative and lack the two aforementioned characteristics (cognitive and normative legitimacy),

they often engender considerable resistance to change (Henisz & Zelner, 2004). Furthermore, a NHST ban may be unrealistic because, first, for some complex multivariate research designs, effect sizes have not been developed yet (Fidler, 2002; Kline, 2004). Second, it would be unclear who would issue or enforce such an edict in a pluralistic and decentralized field such as OS (Van Maanen, 1995).

A neoinstitutional perspective on the NHST problem suggests that, most likely, additional incentives will have to be offered to make other, non-NHST procedures more attractive. Some of the incentives may already be present because of study-inherent rewards, such as the observed generally high research impact of meta-analyses (Judge, Cable, Colbert, & Rynes, 2007). Of course, such approach-based, rather than avoidance-based, solutions to the NHST problem will have to be accompanied by referees' constant reminders to focus on effect sizes and measures of empirical uncertainty rather than on NHST because even a cursory review of OS journals shows that NHST is still endemic in meta-analyses and mathematical models.

Furthermore, journal editors and reviewers could offer external rewards by incentivizing the use of suitable alternative approaches, which are determined by study context (Schwab, Abrahamson, Starbuck, & Fidler, 2011; Ziliak & McCloskey, 2008), *in addition to* NHST. At present, the approach that appears to be most compatible and congruent with the current epistemology of theory testing in the social and organizational sciences is, arguably, the modeling framework advocated by Rodgers (2010). Other guidelines, such as reporting of the probability of replicating an effect (p_{rep}), may be useful not only for eliminating some of the pitfalls of NHST but also for gauging the risks of pursuing a particular line of study (Killeen, 2005a). It is important to note, though, that the congruity of new techniques (such as mathematical modeling or p_{rep}) with epistemological conventions (such as the true/false dichotomy) will increase their initial accessibility. However, these congruous methods also do not present a viable long-term solution because, as noted herein and by other researchers (e.g., Doros & Geier, 2005; Wagenmakers & Gruenwald, 2006), they preserve quite a few problems inherent in NHST. In contrast, when alternatives contradict widespread epistemological premises and are difficult to comprehend (see, e.g., Cumming, 2005; Schmidt, 2008), their influence may be limited to particular journals and, thus, fail to transform an entire discipline. This realistic expectation, applicable at least to some extent to techniques such as p_{rep} and Bayesian procedures, is again consistent with the neoinstitutional approach of this article, which highlighted the social forces affecting the conduct of science. Generally, in contrast to previous methodological discussions of NHST, the critics of NHST should realize that deinstitutionalization must involve reform efforts that are not only regulative but also firmly supported by the other two institutional pillars—cultural-cognitive and normative—in order to neutralize the "contagion of legitimacy" (Zucker, 1991, p. 105) currently attached to NHST.

Conclusion

In sum, this article argued that, in any effort to delegitimize and deinstitutionalize NHST, two serious and opposite errors should be avoided. First, small piecemeal change is likely to be counterproductive in the long run because it does not lead to any lasting substantive change. When each reform effort proposed in this article is implemented in isolation rather than in conjunction with other institutional, or systemic, change efforts, NHST is bound to prevail because of the predominant characteristic of institutions to stabilize social systems and, thus, be relatively inert (Jepperson, 1991; Meyer & Rowan, 1977; W. R. Scott, 2008). From such a sociological perspective, when institutions are maintained by all three pillars (i.e., cultural-cognitive, normative, and regulative forces) bottom-up reform through a change of *individual* researchers' methodological practices is an unrealistic expectation. For example, if NHST is really such a fundamental problem in data analysis, can we really expect a solution from individual researchers embracing statistical techniques that either are

widely perceived as equivalent to NHST or draw on NHST logic at nearly every step? Or, can NHST really be eliminated without a commitment to greater precision in the theoretical specification of parameters?

Second, the opposite extreme, namely, the wholesale rejection of all NHST approaches, is equally counterproductive. For example, there is a rhetorical strategy among some NHST critics based on the argument that post-NHST organizational and social science must start from scratch:

> You can't believe anything [in NHST-based research in economics]. Not a word. It is all nonsense, which future generations of economists are going to have to do all over again. Most of what appears in the best journals of economics is unscientific rubbish. (McCloskey, 2002, p. 55)

Contrary to this argument, many research reports in OS do provide data that are useful for meta-analytic integrations, which in turn can be used as input for the development of causal models among key constructs (Hunter & Schmidt, 2004; Schmidt, 1992). As an analogy, although it is true that Einstein revolutionized physics, it is false to claim that all of Newton's earlier findings were rendered obsolete by Einstein's theories (Hawking, 1996; Wiener, 1958). So, the rhetoric employed in attacks on the illegitimacy of a scientific practice should not exaggerate the ill effects of NHST. Exaggerated rhetoric is likely to produce unnecessary and unhelpful psychological resistance among significance testers, who presently comprise the vast majority of social and organizational scientists. Inflated claims about the futility of NHST-based research will only strengthen such defensiveness, which is ultimately brought about by an erosion of scientific confidence (Chalmers, 1999; Kuhn, 1996).

Hence, the central conclusion of this article is that the proponents of institutional reform must steer clear of the Scylla of small, hardly perceptible methodological changes within an unchanged prereform epistemology and the Charybdis of emotive, factual exaggeration, which only invites defensiveness (see also Orlitzky, 2011b, for more details). A number of cultural-cognitive, normative, and regulative reforms were proposed instead. More specifically, as shown in Figure 1, the suggested reforms across the three institutional domains can be graphically displayed along the *y*-axis of perceived importance and the *x*-axis of required time horizon. The vertical coordinate of importance represents a subjective assessment of the relative importance of the particular reform effort. Importance, in turn, is a function of the extent to which each change effort represents a genuine, substantive shift away from hypothetico-deductivism and can be internalized and to what extent its resultant change in data analysis practice is likely to persist in OS. The horizontal coordinate of time horizon is the extent to which each particular reform is expected to take time (''long term'') or has, in fact, already emerged in past or current OS research practice (''short term''). In general, the more radical and revolutionary the proposed reform, the longer the time horizon required for its implementation. In the context of Figure 1, it should be emphasized that the reformers cannot exclusively rely on the two upper quadrants representing high-importance change. Rather, all institutional reforms tabulated in Figure 1 ought to be pursued in conjunction by those who consider the deinstitutionalization of NHST a vital precondition for meaningful quantitative research. Conversely, to preserve NHST (if readers disagree with the authors cited on the first few pages of this article on the severity of the NHST problem), opponents to methodological reform will have to prevent the developments listed in all four quadrants of Figure 1.

The relatively low importance of a NHST ban, as shown in Figure 1, should not be interpreted as suggestive of the low importance of all regulative forces in general. Although, admittedly, self-interested compliance with rules and sanctions is not as important within institutional theory as it would be from an economic or political science perspective (W. R. Scott, 2008, p. 52), researchers should still be offered incentives to abandon, or at a minimum reduce reliance on, NHST. What,

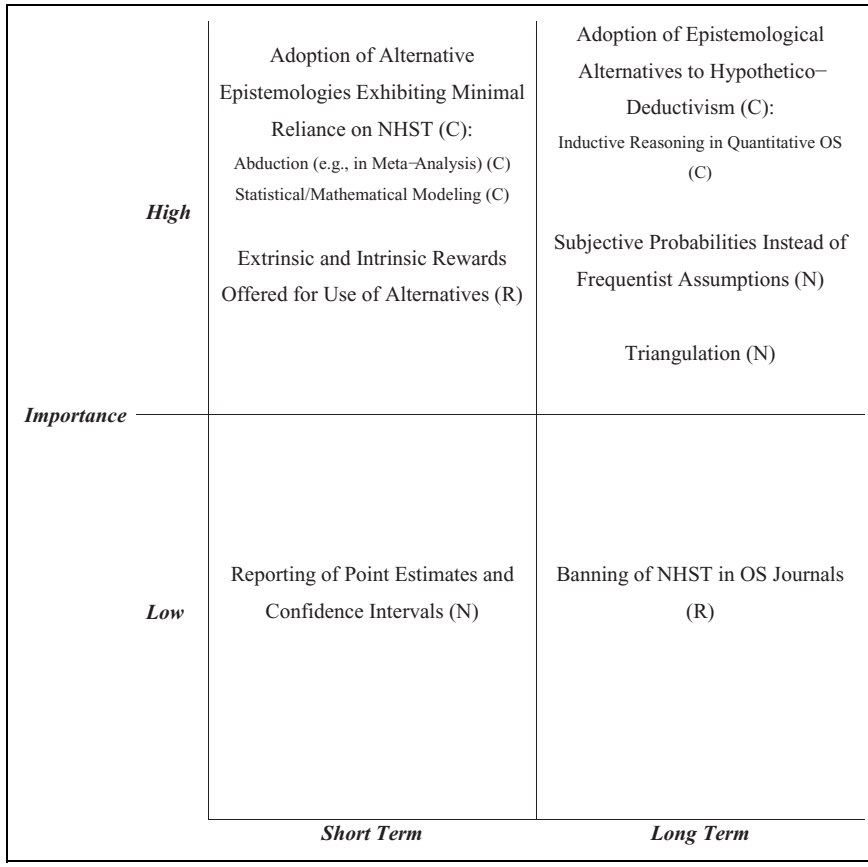

Figure 1. Reform proposals sorted by importance and expected time horizon
Note: Letters in parentheses indicate main institutional forces invoked in specific reform proposals: C = cultural-cognitive forces; N = normative forces; R = regulative forces. NHST = null hypothesis significance testing; OS = organization studies.

from the institutional perspective of this article, is most important is that, as already noted above, researchers are incentivized to make fundamental cultural-cognitive and normative changes rather than only comply with surface-level, and possibly temporary, coercive solutions, such as a NHST ban (Fidler, Thomason, et al., 2004; Ziliak & McCloskey, 2008). A good starting point in that respect would be the continuation of the Academy of Management preconference workshops on NHST and their alternatives, organized by Schwab and his colleagues under the auspices of the Research Methods and other Divisions. In the final analysis, however, such relatively small-scale workshops are likely to be ineffective if there is no fundamental global change in how epistemology and statistics are taught in undergraduate and graduate programs, as suggested in the previous sections on cultural-cognitive and normative reforms.

Generally, the discussion in this article emphasized that the reforms required for the deinstitutionalization of NHST are wide-ranging—and certainly not limited to methodological improvements or individual-level adjustments in researcher decisions. Broader regulative reforms, such as changes in publication manual and journal style guidelines, are undoubtedly necessary to delegitimize NHST. However, because so many wider epistemological and normative issues are implicated, rational methodological critiques by themselves are unlikely to make much difference in changing institutional practice in the social and organizational sciences. They would work only if NHST were not

a deeply embedded institution within the social project of OS. For NHST reforms to happen *and* stick, an OS-wide debate about the pros and cons of NHST must begin so that the methods experts opposing NHST will not remain "voices crying in the wilderness" (Schmidt, 1996, p. 116). Any deinstitutionalization of NHST must be based on the realization that the probability is low that the legitimacy of NHST will erode only because the technique can be shown to be scientifically dysfunctional. Rather, to delegitimize such a firmly institutionalized practice as NHST, political and social pressures for collective change will most likely have to accompany all those rational-functional critiques (Oliver, 1992). For example, political pressure could emerge from demographic changes among quantitative OS researchers, that is, increasing proportions of Bayesian statisticians and of researchers who rely on exploratory data analysis (Tukey, 1977), modeling (Rodgers, 2010), robust statistics (Wilcox, 1998), or resampling techniques, which are also known as computer-intensive methods (Diaconis & Efron, 1983). Generally, research sponsors, methods experts, and journal editors have an important role to play in accelerating these demographic shifts by facilitating the implementation, and/or convincingly highlighting the advantages, of non-NHST procedures.

However, to what extent these reforms can really constitute a *concerted* effort remains very much an open question. On the one hand, there are important compatibilities among the potential reform initiatives. For example, Bayesian reasoning is, in general, consistent with Rodgers's (2010) emphasis on the epistemology of modeling (Krauss, Martignon, & Hoffrage, 1999). Or, as another example, triangulation and abduction are consistent with the incremental estimation of point estimates and confidence intervals in replications. On the other hand, Figure 1 also indicates that pressure will have to be applied from very different research communities characterized by fundamental epistemological incommensurabilities. For example, inductive reasoning tends to be philosophically aligned with an objectivist ontology underpinning the reliability of human sensory perception (Locke, 2007), whereas many[8] Bayesians believe in the social construction of knowledge (Pruzek, 1997; Rindskopf, 1997). That is, inductivists assume that the OS research community can turn back time to an era when human observation and reason were not philosophically problematized yet. Although a joining of forces of the different research communities that are dispersed across the four quadrants of Figure 1 cannot be presumed, it is an open question whether such disjointed attacks on NHST bode well or ill for the abandonment of NHST. It is certainly plausible to suggest that NHST may indeed be delegitimized more quickly if different, highly dispersed research communities assail it from different angles.

Beneath the surface of this article lurks a Darwinian conjecture. Would not demographic shifts (i.e., by significance testers gradually going extinct) and other evolutionary changes ensure that only the most appropriate, or "fittest," methods survive in the long run? In cultural and scientific evolution, however, things may be a bit more complicated than in natural evolution (see also Gould, 2002, p. 969), as the arguments presented in this article implied. Scientific progress (à la Spencerian selection pressure via "survival of the fittest")—however such progress may ultimately be defined (see, e.g., D. T. Campbell, 1990; Rescher, 1990; Wuketits, 1990, pp. 173-175)—cannot necessarily be assumed when evolutionary dynamics are marked by relatively stable periods of normal science interrupted by periods of methodological upheaval (Feyerabend, 1975; Kuhn, 1996). In fact, the insight that there is "no innate tendency to progressive development" may represent Darwin's "greatest conceptual advance over previous evolutionary theories" (Gould, 2002, p. 468).[9] In cultural and scientific evolution, one major reason why teleological progress toward an appropriate, rational goal (e.g., of superior methodology) does not necessarily occur is the impact of sociological, political, or psychological forces rather than rational calculation during periods of discord, upheaval, transition, or political consensus building (Gersick, 1991; Gould, 2002, pp. 999-1022). To the point, the organizational and social sciences have a record of many false starts and myths[10] and overall a distinct lack of progress (see, e.g., Donaldson, 1995; Ghoshal, 2005; Nola, 2003; Pfeffer, 1993). Specifically, the methodological history of

NHST seems to suggest the same conclusion (Kline, 2004; Seth et al., 2009; Ziliak & McCloskey, 2008). More generally, the plausible argument has been advanced that much could be gained if only the organizational and other social sciences went back to first principles in epistemology (J. Cohen, 1990; Donaldson, 1996; Hambrick, 2007; Hoppe, 2007; Locke, 2007; Rosenthal, 1995; Schmidt, 1996; Tukey, 1977).

In many social practices (including social science practices such as NHST), there is a knowing–doing gap (e.g., see Pfeffer & Sutton, 2000); that is, individuals and collectives follow institutional practice, or a social ritual, rather than best practice.[11] In an exploratory rather than definitive attempt at closing this gap, the present article first highlighted the gravity and extent of the problem of statistical significance testing. Then, it described the most important cultural-cognitive, normative, and regulative preconditions for the deinstitutionalization of NHST. In this effort, a neoinstitutional perspective was combined with ideas from the philosophy of science to develop a set of recommendations for quantitative research reforms that transcend purely methodological change efforts. The central question now is whether the community of OS researchers will really be able to drop the "heavy tool" (Weick, 1996, p. 301) of NHST, if epistemology and the institutional practice of the social and organizational sciences are indeed the root causes of the problem.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Not all null hypotheses are so-called nil hypotheses, which refers to null hypotheses that postulate no effect or an effect size of zero (J. Cohen, 1994, p. 1000). It should be noted that many of the problems associated with null hypothesis significance testing (NHST) are not due only to nil hypotheses but to statistical significance testing more generally. For the purposes of this article, the abbreviation *NHST* refers to both *null hypothesis significance testing* and *null hypothesis significance tests*.
2. Because methodological critiques of NHST typically emerge from social science disciplines that approach problems from an individualist epistemology (i.e., psychology rather than sociology), critics' focus on the responsibility or irresponsibility of individual quantitative researchers is not surprising. In contrast, this article assumes that accounts about the sociology of science—like descriptions of all social phenomena—are incomplete if based exclusively on individualist explanations (see also Kincaid, 1996, Chapter 5; Knorr-Cetina, 1999; Kuhn, 1996).
3. First, methodological individualism is the idea that "all true theories of social science are *reducible* to theories of individual human action" (Nozick, 1977, p. 353). Second, in this article, *methodological individualism* refers to the related and more specific idea of researchers making atomistic, individual decisions about the methods used in their research. Because of their asocial and decontextualized nature, both theses lack credibility.
4. This percentage was estimated by using Van Maanen's (1998) introduction to qualitative research in *Administrative Science Quarterly*. The 1986–1996 column shows 178 quantitative articles, and the total number of empirical articles in that 10-year period was 36 (qualitative) + 178. A study of the *Journal of Applied Psychology* (Hubbard, Parsa, & Luthy, 1997) and my own review of the *Academy of Management Journal* (discussed later) suggest that these estimates of the total number of quantitative articles (83%) and the number of NHST in those articles (95%) are most likely conservative estimates.

5. Hempel (1965) is commonly seen as the architect and codifier of hypothetico-deductivism. However, the way NHST is currently used by most social and organizational scientists is actually closer to Popper's (1969, 1972) doctrine of falsificationism, emphasizing the importance of *disconfirming* evidence, rather than Hempel's law-like generalizations and possibility of verification (for supportive interpretations of hypothetico-deductivism in the context of Hempel's and Popper's logics of science, see, e.g., B. P. Cohen, 1989; Chow, 1998b, 2000; Chalmers, 1999; Gorski, 2004; Kincaid, 1996).
6. The influence of this meta-analysis was first postulated by Vogel (2005, p. xvi) and, with over 120 Google Scholar citations per year, most recently included, as the fourth most important study, in Hoffman's tabulation of 75 seminal articles in the field of business and the natural environment (http://oneaomonline.blogspot.com/2011/07/thirty-five-years-of-research-on_13.html). The study also won the 2004 Moskowitz award for outstanding quantitative research relevant to the social investment field. More details can be found in Orlitzky (2008, p. 114).
7. As noted in a later section, regulative pressures by editors and reviewers effectively cause many researchers to adopt this framing of meta-analysis as a theory-testing rather than fact-finding tool.
8. Not all Bayesians follow a subjectivist epistemology (Press, 2003; Reichardt & Gollob, 1997).
9. Yet other statements by Darwin show his reluctance to abandon "his culture's central concern with progress, if only to respect a central metaphor that appealed so irresistibly to most of his contemporaries" (Gould, 2002, p. 468).
10. The idea of progress in the organizational and social sciences is itself a myth, a "great psychic balm" (Gould, 2002, p. 588), possibly born from the desire of each generation of social scientists to preserve and enhance their self-esteem and scientific identity.
11. A less charitable conclusion, supported by considerable empirical evidence (e.g., Gigerenzer, 2004; Oakes, 1986; Schmidt, 1996; Seth et al., 2009; Ziliak & McCloskey, 2008), is that many OS researchers are still unaware of the problems related to NHST and, therefore, currently assume that NHST represents best practice in data analysis.

References

Abelson, R. P. (1995). *Statistics as principled argument.* Hillsdale, NJ: Lawrence Erlbaum.

Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 117-144). Mahwah, NJ: Erlbaum.

Abrahamson, E. (1996). Management fashion. *Academy of Management Review*, *21*(1), 254-285.

Abrahamson, E. (2007, August). *The case against null hypothesis significance testing: Flaws, alternatives, and action plans*. Paper presented at the Academy of Management Annual Meeting, Philadelphia, PA.

Algera, J. A., Jansen, P. G. W., Roe, R. A., & Vijn, P. (1984). Validity generalization: Some critical remarks on the Schmidt-Hunter procedure. *Journal of Occupational Psychology*, *57*, 197-210.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management*, *64*, 912-923.

Armstrong, J. S. (2007). Significance tests harm progress in forecasting. *International Journal of Forecasting*, *23*(2), 321-327.

Arthur, W. B. (1994). Inductive reasoning and bounded rationality. *American Economic Review*, *84*(2), 406-411.

Austin, J. T., Boyle, K. A., & Lualhati, J. C. (1998). Statistical conclusion validity for organizational science researchers: A review. *Organizational Research Methods*, *1*(2), 164-208.

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

Barber, B. (1971). Resistance by scientists to scientific discovery. *Science*, *134*(3479), 596-602.

Barnes, B., Bloor, D., & Henry, J. (1996). *Scientific knowledge: A sociological analysis.* Chicago: University of Chicago Press.

Bedeian, A. G., Sturman, M. C., & Streiner, D. L. (2009). Decimal dust, significant digits, and the search for stars. *Organizational Research Methods*, *12*(4), 687-694.

Begg, C. B. (1994). Publication bias. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 399-409). New York: Russell Sage.

Beyer, J. M., Chanove, R. G., & Fox, W. B. (1995). The review process and the fates of manuscripts submitted to *AMJ*. *Academy of Management Journal*, *38*(5), 1219-1260.

Bloor, D. (1991). *Knowledge and social imagery* (2nd ed.). Chicago: University of Chicago Press.

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *Philosophical Review*, *97*, 303-352.

Bovens, L., & Hartmann, S. (2004). *Bayesian epistemology.* New York: Oxford University Press.

Brandstaetter, E. (2001). Confidence intervals as an alternative to significance testing. *Methods of Psychological Research*, *4*(2), 33-46.

Brown, R. H., & Malone, E. L. (2004). Reason, politics, and the politics of truth: How science is both autonomous and dependent. *Sociological Theory*, *22*(1), 106-122.

Burawoy, M. (1990). Marxism as science. *American Sociological Review*, *55*, 775-793.

Burrell, G. (1996). Normal science, paradigms, metaphors, discourses and genealogies of analysis. In S. R. Clegg, C. Hardy, & W. R. Nord (Eds.), *Handbook of organization studies* (pp. 642-658). London: Sage.

Campbell, D. T. (1990). Epistemological roles for selection theory. In N. Rescher (Ed.), *Evolution, cognition, and realism: Studies in evolutionary epistemology* (pp. 1-19). Lanham, MD: University Press of America.

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

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

Chalmers, A. F. (1999). *What is this thing called science?* (3rd ed.). Indianapolis, IN: Hackett.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*(7), 287-291.

Chow, S. L. (1998a). The null-hypothesis significance-test procedure is still warranted. *Behavioral and Brain Sciences*, *21*(2), 228-239.

Chow, S. L. (1998b). Précis of statistical significance: Rationale, validity, and utility. *Behavioral and Brain Sciences*, *21*(2), 169-239.

Chow, S. L. (2000). The Popperian framework, statistical significance, and rejection of chance. *Behavioral and Brain Sciences*, *23*(2), 294-298.

Cohen, B. P. (1989). *Developing sociological knowledge: Theory and method* (2nd ed.). Chicago: Nelson-Hall.

Cohen, J. (1965). Some statistical issues in psychological research. In D. D. Wolman (Ed.), *Handbook of clinical psychology.* New York: McGraw-Hill.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304-1312.

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

Colbert, B. A. (2004). The complex resource-based view: Implications for theory and practice in strategic human resource management. *Academy of Management Review*, *29*(3), 341-358.

Cooper, G. F. (1990). The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, *42*, 393-405.

Cooper, H. M., & Hedges, L. V. (1994). *The handbook of research synthesis.* New York: Russell Sage Foundation.

Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods*, *2*, 161-172.

Cortina, J. M., & Folger, R. G. (1998). When is it acceptable to accept a null hypothesis: No way, Jose? *Organizational Research Methods*, *1*(3), 334-350.

Coursol, A., & Wagner, E. E. (1986). Effect of positive findings on submission and acceptance rates: A note on meta-analysis bias. *Professional Psychology*, *17*, 136-137.

Creswell, J. W. (1998). *Qualitative inquiry and research design: Choosing among five traditions.* Thousand Oaks, CA: Sage.

Cumming, G. (2005). Understanding the average probability of replication: Comment on Killeen (2005). *Psychological Science*, *16*, 1002-1004.

Daft, R. L., & Weick, K. E. (1984). Toward a model of organizations as interpretation systems. *Academy of Management Review*, *9*(2), 284-295.

Dalton, D. R., & Dalton, C. M. (2008). Meta-analyses: Some very good steps toward a bit longer journey. *Organizational Research Methods*, *11*(1), 127-147.

Dawes, R. M. (1991). Probabilistic versus causal thinking. In D. Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology: Matters of public interest: Essays in honor of Paul Everett Meehl* (Vol. 1, pp. 235-264). Minneapolis: University of Minnesota Press.

Deem, R., & Brehony, K. J. (2000). Doctoral students' access to research cultures—Are some more unequal than others? *Studies in Higher Education*, *25*(2), 149-165.

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American*, *248*(5), 116-130.

di Bacco, M., d'Amore, G., & Scalfari, F. (2004). *Applied Bayesian statistical studies in biology and medicine.* New York: Springer.

Dixon, P. (2003). The *p*-value fallacy and how to avoid it. *Canadian Journal of Experimental Psychology*, *57*(3), 189-202.

Donaldson, L. (1995). *American anti-management theories of organization: A critique of paradigm proliferation.* Cambridge, UK: Cambridge University Press.

Donaldson, L. (1996). *For positivist organization theory: Proving the hard core.* Thousand Oaks, CA: Sage.

Doros, G., & Geier, A. B. (2005). Probability of replication revisited: Comment on "An Alternative to Null-Hypothesis Significance Tests." *Psychological Science*, *16*, 1005-1006.

Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of Management Review*, *14*, 532-550.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory and Psychology*, *5*(1), 75-98.

Feyerabend, P. K. (1975). *Against method: Outline of an anarchistic theory of knowledge.* London: New Left Books.

Fidler, F. (2002). The fifth edition of the APA publication manual: Why its statistics recommendations are so controversial. *Educational and Psychological Measurement*, *62*(5), 749-770.

Fidler, F., Cumming, G., Burgman, M., & Thomason, N. (2004). Statistical reform in medicine, psychology and ecology. *Journal of Socio-Economics*, *33*, 615-630.

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but they can't make them think: Statistical reform lessons from medicine. *Psychological Science*, *15*, 119-126.

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical significance in the *Journal of Applied Psychology*: Little evidence of reform. *Educational and Psychological Measurement*, *61*(2), 181-210.

Finch, S., Thomason, N., & Cumming, G. (2002). Past and future American Psychological Association guidelines for statistical practice. *Theory and Psychology*, *12*(6), 825-853.

Friedman, M. (1953). *Essays in positive economics.* Chicago: University of Chicago Press.

Fuchs, S. (2001). What makes "sciences" scientific? In J. H. Turner (Ed.), *Handbook of sociological theory* (pp. 21-35). New York: Kluwer Academic/Plenum Publishers.

Gans, J. S., & Shepherd, G. B. (1994). How are the mighty fallen: Rejected classic articles by leading economists. *Journal of Economic Perspectives*, *8*, 165-179.

Gatsonis, C., Kass, R. E., Carling, B., Carriquiry, A., Gelman, A., & Verdinelli, I. (2001). *Case studies on Bayesian statistics.* (Vol. 5). New York: Springer-Verlag.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis.* London: Chapman & Hall.

Gersick, C. J. G. (1991). Revolutionary change theories: A multilevel exploration of the punctuated equilibrium paradigm. *Academy of Management Review*, *16*(1), 10-36.

Ghoshal, S. (2005). Bad management theories are destroying good management practices. *Academy of Management Learning and Education*, *4*(1), 75-91.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Vol. 1. Methodological issues* (pp. 311-339). Hillsdale, NJ: Lawrence Erlbaum.

Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences*, *21*, 199-200.

Gigerenzer, G. (2000). *Adaptive thinking: Rationality in the real world.* New York: Oxford University Press.

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, *33*, 587-606.

Gigerenzer, G., & Hoffrage, U. (1999). Overcoming difficulties in Bayesian reasoning: A reply to Lewis and Keren (1999) and Mellers and McGraw (1999). *Psychological Review*, *106*, 425-430.

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 391-408). Thousand Oaks, CA: Sage.

Godfrey, P. C. (2005). The relationship between corporate philanthropy and shareholder wealth: A risk management perspective. *Academy of Management Review*, *30*(4), 777-798.

Goldstone, J. A. (2004). Response: Reasoning about history, sociologically. *Sociological Methodology*, *34*, 35-61.

Gorski, P. S. (2004). The poverty of deductivism: A constructive realist model of sociological explanation. *Sociological Methodology*, *34*, 1-33.

Gould, S. J. (2002). *The structure of evolutionary theory.* Cambridge, MA: Belknap/Harvard.

Griffin, J. J., & Mahon, J. F. (1997). The corporate social performance and corporate financial performance debate: Twenty-five years of incomparable research. *Business and Society*, *36*, 5-31.

Gujarati, D. N. (1988). *Basic econometrics* (2nd ed.). New York: McGraw-Hill.

Guttman, L. (1985). The illogic of statistical inference for cumulative science. *Applied Stochastic Models and Data Analysis*, *1*, 3-10.

Hacking, I. (1965). *The logic of statistical inference.* Cambridge, UK: Cambridge University Press.

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

Haig, B. D. (2000). Statistical significance testing, hypothetico-deductive method, and theory evaluation. *Behavioral and Brain Sciences*, *23*(2), 292-293.

Haig, B. D. (2005a). An abductive theory of scientific method. *Psychological Methods*, *10*(4), 371-388.

Haig, B. D. (2005b). Exploratory factor analysis, theory generation, and scientific method. *Multivariate Behavioral Research*, *40*(3), 303-329.

Hakala, J. (2009). Socialization of junior researchers in new academic research environments: Two case studies from Finland. *Studies in Higher Education*, *34*(5), 501-516.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research* [Online serial], *7*(1), 1-20.

Hambrick, D. C. (2007). The field of management's devotion to theory: Too much of a good thing? *Academy of Management Journal*, *50*(6), 1346-1352.

Hannan, M. T., Polos, L., & Carroll, G. R. (2007). *Logics of organization theory: Audiences, codes, and ecologies.* Princeton, NJ: Princeton University Press.

Hanson, N. R. (1965). *Patterns of discovery.* Cambridge, UK: Cambridge University Press.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997). *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum.

Harris, R. J. (1997). Significance tests have their place. *Psychological Science*, *8*(1), 8-11.

Hawking, S. (1996). *The illustrated A Brief History of Time (Updated and expanded ed.)*. New York: Bantam.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? The empirical cumulativeness of research. *American Psychologist*, *42*(2), 443-455.

Hempel, C. G. (1965). *Aspects of scientific explanation and other essays in the philosophy of science.* New York: Free Press.

Henisz, W. J., & Zelner, B. A. (2004). *Resistance to illegitimate multilateral influence on reform: The political backlash against private infrastructure investments.* Philadelphia, PA: Unpublished manuscript.

Hill, T. L. (2003). Null hypothesis significance testing, the file drawer problem, and intercessory prayer: A Bayesian analysis of the effect (or effectiveness) of each. *Dissertation Abstracts International, 63,* 3475B.

Hofmann, S. G. (2002). Fisher's fallacy and NHST's flawed logic. *American Psychologist*, *57*(1), 69-70.

Hoover, K. D., & Siegler, M. V. (2008). Sound and fury: McCloskey and significance testing in economics. *Journal of Economic Methodology*, *15*(1), 1-37.

Hoppe, H.-H. (2007). *Economic science and the Austrian method.* Auburn, AL: Ludwig von Mises Institute.

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.

Howson, C., & Urbach, P. (2005). *Scientific reasoning: The Bayesian approach* (3rd ed.). Chicago: Open Court.

Hubbard, R., & Armstrong, J. S. (1992). Are null results becoming an endangered species in marketing? *Marketing Letters*, *3*(2), 127-136.

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

Huff, A. S. (1999). *Writing for scholarly publication.* Thousand Oaks, CA: Sage.

Hunt, M. (1997). *How science takes stock: The story of meta-analysis.* New York: Russell Sage Foundation.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, *8*(1), 3-7.

Hunter, J. E. (1998). Testing significance testing: A flawed defense. *Behavioral and Brain Sciences*, *21*(2), 204.

Hunter, J. E., & Gerbing, D. W. (1982). Unidimensional measurement, second order factor analysis, and causal models. *Research in Organizational Behavior*, *4*, 267-320.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

Jepperson, R. L. (1991). Institutions, institutional effects, and institutionalism. In W. W. Powell & P. J. DiMaggio (Eds.), *The new institutionalism in organizational analysis* (pp. 143-163). University of Chicago Press.

Jick, T. D. (1979). Mixing qualitative and quantitative methods: Triangulation in action. *Administrative Science Quarterly*, *24*, 602-611.

Johnson, D. H. (1999). The insignificance of statistical significance testing. *Journal of Wildlife Management*, *63*(3), 763-772.

Jones, L. V., & Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods*, *5*(4), 411-414.

Judge, T. A., Cable, D. M., Colbert, A. E., & Rynes, S. L. (2007). What causes a management article to be cited—Article, author, or journal? *Academy of Management Journal*, *50*(3), 491-506.

Kaufman, A. S. (1998). Introduction to the special issue on statistical significance testing. *Research in the Schools*, *5*, 1.

Killeen, P. R. (2005a). An alternative to null-hypothesis significance tests. *Psychological Science*, *16*(5), 345-353.

Killeen, P. R. (2005b). Replicability, confidence, and priors. *Psychological Science*, *16*, 1009-1012.

Killeen, P. R. (2006). The problem with Bayes. *Psychological Science*, *17*, 643-644.

Kimball, A. W. (1957). Errors of the third kind in statistical consulting. *Journal of the American Statistical Association*, *52*(278), 133-142.

Kincaid, H. (1996). *Philosophical foundations of the social sciences.* New York: Cambridge University Press.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research.* Washington, DC: American Psychological Association.

Knorr-Cetina, K. D. (1999). *Epistemic cultures: How the sciences make knowledge.* Cambridge, MA: Harvard University Press.

Krauss, S., Martignon, L., & Hoffrage, U. (1999). Simplifying Bayesian inference: The general case. In L. Magnani, N. Nersessian, & P. Thagard (Eds.), *Model-based reasoning in scientific discovery* (pp. 165-180). New York: Plenum.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist*, *56*(1), 16-26.

Kuhn, T. S. (1996). *The structure of scientific revolutions* (3rd ed.). Chicago: The University of Chicago Press.

Labovitz, S. (1972). Statistical usage in sociology: Sacred cows and rituals. *Sociological Methods and Research*, *1*(1), 13-37.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge.* Cambridge, UK: Cambridge University Press.

Latour, B., & Woolgar, S. (1986). *Laboratory life: The construction of scientific facts* (2nd ed.). Princeton, NJ: Princeton University Press.

Lewis, C., & Keren, G. (1999). On the difficulties underlying Bayesian reasoning: A comment on Gigerenzer and Hoffrage. *Psychological Review*, *106*, 411-416.

Locke, E. A. (2007). The case for inductive theory building. *Journal of Management*, *33*(6), 867-890.

Longino, H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry.* Princeton, NJ: Princeton University Press.

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

Lynch, M. (1993). *Scientific practice and ordinary action: Ethnomethodology and social studies of science.* Cambridge, UK: Cambridge University Press.

Macdonald, S., & Kam, J. (2007). Ring a ring o' roses: Quality journals and gamesmanship in management studies. *Journal of Management Studies*, *44*(4), 640-655.

MacKenzie, D. A. (1981). *Statistics in Britain, 1865-1930: The social construction of scientific knowledge.* Edinburgh, Scotland: Edinburgh University Press.

Mahon, J. F. (2002). Corporate reputation: A research agenda using strategy and stakeholder literature. *Business and Society*, *41*(4), 415-445.

Mahoney, M. J. (1976). *Scientist as subject: The psychological imperative.* Cambridge, MA: Ballinger.

Malgady, R. G. (2000). Myths about the null hypothesis and paths to reform. In R. H. Dana (Ed.), *Handbook of cross-cultural and multicultural personality assessment* (pp. 49-62). Mahwah, NJ: Lawrence Erlbaum.

Masson, M. E., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology*, *57*(3), 203-220.

Matthews, R. A. J. (2000). Facts versus factions: The use and abuse of subjectivity in scientific research. In J. Morris (Ed.), *Rethinking risk and the precautionary principle* (pp. 247-282). Butterworth-Heinemann.

McCloskey, D. N. (2002). *The secret sins of economics.* Chicago: Prickly Paradigm Press.

McCloskey, D. N., & Ziliak, S. T. (1996). The standard error of regression. *Journal of Economic Literature*, *34*(1), 97-114.

McGrath, R. E. (1998). Significance testing: Is there something better? *American Psychologist*, *53*(7), 796-797.

McNemar, Q. (1960). At random: Sense and nonsense. *American Psychologist*, *15*, 295-300.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*(Monograph Suppl. 1-V66), 195-244.

Meehl, P. E. (1991). Why summaries of research on psychological theories are often uninterpretable. In R. E. Snow & D. E. Wiley (Eds.), *Improving inquiry in social science: A volume in honor of Lee J. Cronbach* (pp. 13-59). Hillsdale, NJ: Lawrence Erlbaum.

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictors. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 393-426). Mahwah, NJ: Lawrence Erlbaum.

Mendoza, P. (2007). Academic capitalism and doctoral student socialization: A case study. *Journal of Higher Education*, *78*(1), 71-96.

Merton, R. K. (1957). Priorities in scientific discovery: A chapter in the sociology of science. *American Sociological Review*, *22*(6), 635-659.

Meyer, J. W., & Rowan, B. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology*, *83*, 340-363.

Miner, J. B. (2003). The rated importance, scientific validity, and practical usefulness of organizational behavior theories: A quantitative review. *Academy of Management Learning and Education*, *2*(3), 250-268.

Mitroff, I. I., & Silvers, A. (2009). *Dirty rotten strategies: How we trick ourselves into solving the wrong problems precisely.* Stanford, CA: Stanford Business Books.

Morrison, D., & Henkel, R. E. (1970). *The significance test controversy: A reader.* Chicago: Aldine.

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. 65-115). Mahwah, NJ: Lawrence Erlbaum.

Nelson, J. S., Megill, A., & McCloskey, D. N. (1987). Rhetoric of inquiry. In J. S. Nelson, A. Megill, & D. N. McCloskey (Eds.), *The rhetoric of the human sciences: Language and argument in scholarship and public affairs* (pp. 3-18). Madison: University of Wisconsin Press.

Neyman, J. (1957). "Inductive behavior" as a basic concept of philosophy of science. *Revue de l'Institut International de Statistique/Review of the International Statistical Institute*, *25*(1/3), 7-22.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241-301.

Nola, R. (2003). *Rescuing reason: A critique of anti-rationalist views of science and knowledge.* Boston: Kluwer Academic.

Nozick, R. (1977). On Austrian methodology. *Synthese*, *36*(3), 353-392.

Oakes, M. L. (1986). *Statistical inference: A commentary for the social and behavioral sciences.* New York: John Wiley.

Oliver, C. (1992). The antecedents of deinstitutionalization. *Organization Studies*, *13*(4), 563-588.

Orlitzky, M. (2001). Does organizational size confound the relationship between corporate social performance and firm financial performance? *Journal of Business Ethics*, *33*(2), 167-180.

Orlitzky, M. (2006). Links between corporate social responsibility and corporate financial performance: Theoretical and empirical determinants. In J. Allouche (Ed.), *Corporate social responsibility: Vol. 2. Performances and stakeholders* (pp. 41-64). London: Palgrave Macmillan.

Orlitzky, M. (2008). Corporate social performance and financial performance: A research synthesis. In A. Crane, A. McWilliams, D. Matten, J. Moon, & D. Siegel (Eds.), *The Oxford handbook of CSR* (pp. 113-134). Oxford, UK: Oxford University Press.

Orlitzky, M. (2011a). Institutional logics in the study of organizations: The social construction of the relationship between corporate social and financial performance. *Business Ethics Quarterly*, *21*(3), 409-444.

Orlitzky, M. (2011b). Institutionalized dualism: Statistical significance testing as myth and ceremony. *Journal of Management Control, 22,* 47-77.

Orlitzky, M., & Benjamin, J. D. (2001). Corporate social performance and firm risk: A meta-analytic review. *Business and Society*, *40*(4), 369-396.

Orlitzky, M., Schmidt, F. L., & Rynes, S. L. (2003). Corporate social and financial performance: A meta-analysis. *Organization Studies*, *24*(3), 403-441.

Orlitzky, M., Siegel, D. S., & Waldman, D. A. (2011). Strategic corporate social responsibility and environmental sustainability. *Business and Society*, *50*(1), 6-27.

Oswald, A. J. (2007). An examination of the reliability of prestigious journals: Evidence and implications for decision-makers. *Economica*, *74*, 21-31.

Overall, J. E. (1980). Power of chi-square tests for 2 × 2 contingency tables with small expected frequencies. *Psychological Bulletin*, *87*(1), 132-135.

Park, S. H., & Gordon, M. E. (1996). Publication records and tenure decisions in the field of strategic management. *Strategic Management Journal*, *17*, 109-128.

Pearce, J. L. (2004). Presidential address: What do we know and how do we really know it? *Academy of Management Review*, *29*(2), 175-179.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach.* Hillsdale, NJ: Lawrence Erlbaum.

Peters, D. P., & Ceci, S. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences*, *5*, 187-195.

Pfeffer, J. (1993). Barriers to the advance of organizational science: Paradigm development as a dependent variable. *Academy of Management Review*, *18*(4), 599-620.

Pfeffer, J., & Sutton, R. I. (2000). *The knowing–doing gap: How smart companies turn knowledge into action.* Boston: Harvard Business School Press.

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

Popper, K. R. (1962). *The open society and its enemies. Vol. II: The high tide of prophesy: Hegel, Marx, and the aftermath.* London: Routledge and Kegan Paul.

Popper, K. R. (1969). *Conjectures and refutations.* London: Routledge.

Popper, K. R. (1972). *The logic of scientific discovery.* London: Hutchinson.

Porter, T. M. (1992). Quantification and the accounting ideal in science. *Social Studies of Science*, *22*, 633-652.

Press, S. J. (2003). *Subjective and objective Bayesian statistics: Principles, models, and applications.* New York: John Wiley.

Press, S. J. (2005). *Applied multivariate statistics: Using Bayesian and frequentist methods of inference* (2nd ed.). Mineola, NY: Dover.

Pruzek, R. M. (1997). An introduction to Bayesian inference and its applications. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 287-318). Mahwah, NJ: Lawrence Erlbaum.

Reichardt, C. S., & Gollob, H. F. (1997). When confidence intervals should be used instead of statistical tests, and vice versa. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 259-284). Mahwah, NJ: Lawrence Erlbaum.

Rescher, N. (1990). *A useful inheritance: Evolutionary aspects of the theory of knowledge*. Savage, MD: Rowman & Littlefield.

Rindskopf, D. M. (1997). Testing "small," not null, hypotheses: Classical and Bayesian approaches. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 319-332). Hillsdale, NJ: Lawrence Erlbaum.

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist, 65*(1), 1-12.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin, 118,* 183-192.

Rothman, K. J. (1986). *Modern epidemiology*. New York: Little, Brown.

Royall, R. (1997). *Statistical evidence—A likelihood paradigm*. London: Chapman and Hall.

Rozeboom, W. W. (1961). Ontological induction and the logical typology of scientific variables. *Philosophy of Science, 28,* 337-377.

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 335-391). Hillsdale, NJ: Lawrence Erlbaum.

Sackett, P. R., & Larson, J. R., Jr. (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. 419-489). Palo Alto, CA: Consulting Psychologists 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.

Schaffer, S. (1986). Scientific discoveries and the end of natural philosophy. *Social Studies of Science, 16*(3), 387-420.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47,* 1173-1181.

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

Schmidt, F. L. (2008). Meta-analysis: A constantly evolving research integration tool. *Organizational Research Methods, 11*(1), 96-113.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37-64). Hillsdale, NJ: Lawrence Erlbaum.

Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology, 71,* 432-439.

Schofer, E. (2004). Cross-national differences in the expansion of science, 1970-1990. *Social Forces, 83*(1), 215-248.

Schwab, A., Abrahamson, E., Starbuck, W. H., & Fidler, F. (2011). Researchers should make thoughtful assessments instead of null-hypothesis significance tests. *Organization Science, 22*(4), 1105-1120.

Schwab, A., & Starbuck, W. H. (2009). Null-hypothesis significance tests in behavioral and management research: We can do better. *Research Methodology in Strategy and Management, 5,* 29-54.

Scott, J. G. (2009). Nonparametric Bayesian multiple testing for longitudinal performance stratification. *Annals of Applied Statistics, 3*(4), 1655-1674.

Scott, W. R. (1994). Institutions and organizations: Toward a theoretical synthesis. In W. R. Scott & J. W. Meyer (Eds.), *Institutional environments and organizations: Structural complexity and individualism* (pp. 55-80). Thousand Oaks, CA: Sage.

Scott, W. R. (2008). *Institutions and organizations: Ideas and interests* (3rd ed.). Thousand Oaks, CA: Sage.

Seth, A., Carlson, K. D., Hatfield, D. E., & Lan, H.-W. (2009). So what? Beyond statistical significance to substantive significance in strategy research. *Research Methodology in Strategy and Management, 5,* 3-28.

Shrout, P. E. (1997). Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychological Science, 8*(1), 1-2.

Sober, E. (2005). Is drift a serious alternative to natural selection as an explanation of complex adaptive traits? *Royal Institute of Philosophy Supplements, 56,* 125-154.

Starbuck, W. H. (2005). How much better are the most-prestigious journals? The statistics of academic publications. *Organization Science, 16*(2), 180-200.

Starbuck, W. H. (2007, August). *Croquet with the queen of hearts*. Paper presented at a professional development workshop at the Academy of Management Annual Meeting, Philadelphia, PA.

Steel, P. D. G., & Kammeyer-Mueller, J. D. (2008). Bayesian variance estimation for meta-analysis: Quantifying our uncertainty. *Organizational Research Methods, 11*(1), 54-78.

Steele, M., Hurst, C., & Chaseling, J. (2008). The power of chi-square type of goodness-of-fit test statistics. *Far East Journal of Theoretical Statistics, 26*(1), 109-119.

Stephenson, W. W. (1961). Scientific creed—1961. *Psychological Record, 11,* 1-25.

Stinchcombe, A. L. (1968). *Constructing social theories*. Chicago: University of Chicago Press.

Suchman, M. C. (1995). Managing legitimacy: Strategic and institutional approaches. *Academy of Management Review, 20,* 571-610.

Suddaby, R., & Greenwood, R. (2005). Rhetorical strategies of legitimacy. *Administrative Science Quarterly, 50,* 35-67.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. New York: Doubleday.

Tang, Y.-C., & Liou, F.-M. (2009). Does firm performance reveal its own causes? The role of Bayesian inference. *Strategic Management Journal, 31*(1), 39-57.

Thompson, B. (2002). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling and Development, 80,* 64-71.

Townley, B. (2002). The role of competing rationalities in institutional change. *Academy of Management Journal, 45*(1), 163-179.

Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Bayes's theorem. *Psychological Review, 110*(3), 526-535.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminancy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods, 6*(4), 371-386.

Tsang, E. W. K., & Kwan, K.-M. (1999). Replication and theory development in organizational science: A critical realist perspective. *Academy of Management Review, 24*(4), 759-780.

Tsoukas, H., & Knudsen, C. (2003). Introduction: The need for meta-theoretical reflection in organization theory. In H. Tsoukas & C. Knudsen (Eds.), *The Oxford handbook of organization theory: Meta-theoretical perspectives* (pp. 1-36). Oxford, UK: Oxford University Press.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Ullmann, A. (1985). Data in search of a theory: A critical examination of the relationship among social performance, social disclosure, and economic performance. *Academy of Management Review, 10,* 540-577.

Van Maanen, J. (1995). Style as theory. *Organization Science, 6*(1), 133-143.

Van Maanen, J. (1998). Different strokes: Qualitative research in the *Administrative Science Quarterly* from 1956-1996. In J. E. Van Maanen (Ed.), *Qualitative studies of organizations* (pp. ix-xxxii). Thousand Oaks, CA: Sage.

Vogel, D. (2005). *The market for virtue: The potential and limits of corporate social responsibility*. Washington, DC: Brookings Institution Press.

Waddock, S. A., & Graves, S. B. (1997). Quality of management and quality of stakeholder relations: Are they synonymous? *Business and Society, 36*(3), 250-279.

Wagenmakers, E.-J., & Gruenwald, P. (2006). A Bayesian perspective on hypothesis testing: A comment on Killeen (2005). *Psychological Science, 17,* 641-642.

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods, 4*(2), 212-213.

Webb, E. J., Campbell, D., Schwartz, R., Sechrest, L., & Grove, J. (1981). *Nonreactive measures in the social sciences*. Boston: Houghton Mifflin.

Weick, K. E. (1996). Drop your tools: An allegory for organizational studies. *Administrative Science Quarterly, 41,* 301-313.

Wiener, P. P. (1958). Book review of T. S. Kuhn, *The Copernican Revolution*. *Philosophy of Science, 25*(4), 297-299.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53*(3), 300-314.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594-604.

Woodward, J. (1989). Data and phenomena. *Synthese, 79,* 393-472.

Wuketits, F. M. (1990). *Evolutionary epistemology and its implications for humankind*. Albany: State University of New York Press.

Zellner, A. (1996). *An introduction to Bayesian inference in econometrics*. New York: John Wiley.

Zellner, A. (2009). Honorary lecture on S. James Press and Bayesian analysis. *Review of Economic Analysis, 1,* 98-118.

Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Ann Arbor: University of Michigan Press.

Zucker, L. G. (1991). The role of institutionalization in cultural persistence. In W. W. Powell & P. J. DiMaggio (Eds.), *The new institutionalism in organizational analysis*. Chicago: University of Chicago Press.

Bio

Marc Orlitzky (PhD, U. of Iowa) is Associate Professor of Management at the Pennsylvania State University Altoona. Previously, he served on the faculties of the University of New South Wales (UNSW) and University of Auckland (Senior Lecturer Above the Bar) and was a Visiting Research Fellow at the ICCSR at the University of Nottingham and a Visiting Professor at the Technical University of Dresden. His research has been published in *Organization Studies, Business Ethics Quarterly, Journal of Management & Organization, Organizational Research Methods, Business & Society, Personnel Psychology, Academy of Management Review, Academy of Management Learning & Education, Journal of Business Ethics, International Journal of Human Resource Management, Small Group Research,* and many other publication outlets. He has also coauthored (with D. Swanson) the book *Toward Integrative Corporate Citizenship: Research Advances in Corporate Social Performance* and coedited (with J. Moon and G. Whelan) *Corporate Governance and Business Ethics.* His widely cited papers won several research awards, including the 2004 Moskowitz award for outstanding quantitative study in the field of social investing (for Orlitzky, Schmidt, & Rynes, 2003) and the 2001 Best Article Prize by the International Association for Business & Society in association with *California Management Review* (for Orlitzky & Benjamin, 2001).