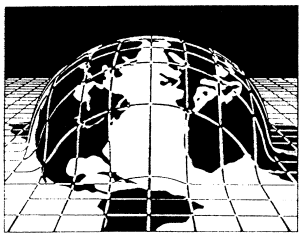
Organization Science
Vol. 22, No. 4, July–August 2011, pp. 1105–1120
ISSN 1047-7039 | EISSN 1526-5455 | 11 | 2204 | 1105

informs

DOI 10.1287/orsc.1100.0557



PERSPECTIVE

Researchers Should Make Thoughtful Assessments Instead of Null-Hypothesis Significance Tests

Andreas Schwab
College of Business, Iowa State University, Ames, Iowa 50011, aschwab@iastate.edu

Eric Abrahamson
Columbia Business School, Columbia University, New York, New York 10027, ea1@columbia.edu

William H. Starbuck
Lundquist College of Business, University of Oregon, Eugene, Oregon 97403, starbuck@uoregon.edu

Fiona Fidler
School of Psychological Science, La Trobe University, Victoria, 3086 Australia, f.fidler@latrobe.edu.au

Null-hypothesis significance tests (NHSTs) have received much criticism, especially during the last two decades. Yet many behavioral and social scientists are unaware that NHSTs have drawn increasing criticism, so this essay summarizes key criticisms. The essay also recommends alternative ways of assessing research findings. Although these recommendations are not complex, they do involve ways of thinking that many behavioral and social scientists find novel. Instead of making NHSTs, researchers should adapt their research assessments to specific contexts and specific research goals, and then explain their rationales for selecting assessment indicators. Researchers should show the substantive importance of findings by reporting effect sizes and should acknowledge uncertainty by stating confidence intervals. By comparing data with naïve hypotheses rather than with null hypotheses, researchers can challenge themselves to develop better theories. Parsimonious models are easier to understand, and they generalize more reliably. Robust statistical methods tolerate deviations from assumptions about samples.

Key words: research design and methods; statistics; analyses
History: Published online in *Articles in Advance* August 20, 2010.

In the mid-1980s, a professor set out to study the language in annual letters to stockholders. Like several other researchers, the professor expected these letters to disclose how managers think about their firms' strategies and performance (Bowman 1984, Fiol 1989, Salancik and Meindl 1984). He formulated hypotheses about differences between letters from unsuccessful and successful companies, and then he compared letters from companies at risk of going bankrupt with letters from successful companies that had closely resembled the failing ones a few years earlier. To his surprise, he found no statistically significant differences between letters from failing and successful companies. He presented his study to a departmental seminar, where participants said they did not find the evidence of no difference convincing. They proposed some new hypotheses and metrics. He incorporated their ideas, but he still found no statistically significant differences.

Repeated lack of support for his theory-based hypotheses led the professor to reframe his paper as a study of corporate communications: companies have reasons to conceal financial problems, and they hire public relations professionals to do so. He sent his manuscript to a prestigious journal. To the professor's excitement, the editor offered an opportunity to revise and resubmit. However, the editor and reviewers did not find the evidence of no difference convincing, and they proposed new hypotheses and metrics. The professor followed their instructions carefully, but the journal's editor and reviewers did not respond enthusiastically to his revised manuscript. Again,

they proposed new hypotheses and metrics, although this time the demanded revisions seemed to be minor. The professor was very hopeful. He revised his manuscript carefully and resubmitted it. However, the editor rejected his manuscript, saying the research methods had been inadequate. The professor was devastated.

Frustrated but determined, the professor submitted his manuscript to another prestigious journal. This time, he supplemented his argument with ideas from political science implying that companies might avoid evaluation. However, the editorial scenario that ensued was very similar to the prior one. Twice, the journal's editor offered opportunity to revise, and each editorial review proposed additional hypotheses and metrics. Twice, the professor revised, following the editor's and reviewers' instructions. Finally, the editor rejected the second revision, saying the research methods had been inadequate.

The professor put the manuscript into a drawer in his desk, locked the drawer, and labeled it "disaster paper." It remains there today.

Determined to surmount statistical hurdles, the professor next analyzed over 2,000 letters to shareholders, and the large sample yielded many significant results. The paper won an award for being the best one published in a very prestigious journal that year. However, the professor thought that his study had found only very small effects.

To the professor's despair, his third study again yielded statistically insignificant results. This time, the professor hired a time-series statistician. After several analyses with different statistical methods and models, they found a pattern of results and published it in a prestigious journal.

The professor drew lessons from these experiences. First, reviewers are more likely to approve of research methods when studies reject null hypotheses. Second, reviewers insist that studies must find differences, even when no difference has important substantive implications. Third, quantitative research was liable to produce findings that he did not trust. He also sensed that such quantitative research might make him highly cynical. He knew scholars who seemed to view their own statistical studies cynically, and he did not like that prospect.

The professor's experiences with these three articles induced him to shy away from quantitative tests of hypotheses. Instead, the professor focused on developing conceptual papers. Several of these won "best paper" awards and appeared in prestigious journals. One award winner, which has received more than 1,400 citations, used simple graphs as evidence.

Underlying the professor's story are major problems with null-hypothesis significance tests (NHSTs). This essay outlines deficiencies and harmful effects of NHSTs and recommends ways to make quantitative research more satisfying and fruitful. Arguments against NHSTs are not novel, but many researchers are unaware of these arguments, and they do not see the harm that NHSTs create. Recognition of the deficiencies of NHSTs is critical for advancement of quantitative research in behavioral and social research. Therefore, the next section of this essay outlines problematic properties of NHSTs, and the ensuing section considers why efforts to move beyond NHSTs have been unsuccessful.

This essay then proposes several ways to improve assessment of research findings while overcoming deficiencies of NHSTs. These recommendations for methodological improvement are not complex, but they involve ways of thinking that may be new to many behavioral and social scientists. This essay's most important recommendation is that researchers should stop relying on NHSTs and think carefully about what assessments are most meaningful in their specific contexts.

What's Wrong with NHSTs Anyway?

NHSTs have been controversial since Fisher (1925) proposed them. For instance, famed statisticians Neyman and Pearson argued in the late 1920s that it makes no sense to test a null hypothesis without testing alternative hypotheses (Hubbard and Bayarri 2003). However, probably because he integrated NHSTs into his very popular textbook, Fisher (1925) was able to persuade many to adopt NHSTs. Complaints about NHSTs have multiplied over time (Cohen 1994, Greenwald 1975, Schmidt and Hunter 1997, Schwab and Starbuck 2009, Seth et al. 2009, Thompson 1999b). However, statistics textbooks have continued to teach their use, and many behavioral and social researchers remain unaware that NHSTs have been subject to strong criticism (Kline 2004, Fidler 2005).

NHSTs cause both conceptual and practical problems. The following sections highlight conceptual problems of NHSTs related to dichotomous conceptions of truth, sample size sensitivities, and implausible null hypotheses.

Conceptual Problem 1: NHSTs Portray Research Findings as Clear-Cut

Paradoxically, an assessment procedure designed for uncertainty about the implications of data does not formally allow for uncertainty about the correctness of hypotheses or ranges of knowledge. Supposedly, data are either "statistically significant" or not so.

Available data define a distribution of probable values for each population parameter of interest. NHSTs replace this distribution with sharply delineated ranges of possible versus impossible values: a confidence interval. NHSTs then portray truth as dichotomous and definite when they either reject or fail to reject null hypotheses. As Tukey (1991, pp. 100–101) stated, "The worst, i.e., most dangerous, feature of 'accepting the null hypothesis' is the giving up of explicit uncertainty . . . .

Mathematics can sometimes be put in such black-and-white terms, but our knowledge or belief about the external world never can."

Of course, many researchers mitigate these dichotomies by using different levels of significance: 0.05, 0.01, and 0.001. However, at any specified level, significance remains dichotomous, and the presence of multiple levels creates dilemmas. Is a null hypothesis rejected at 0.01 more incorrect than one rejected at 0.05?

Any arbitrary threshold for rejecting null hypotheses can amplify very small differences in data into very large differences in implications. In an extreme case, researchers might fail to reject a null hypothesis if data have a probability of 0.0505 and reject this null hypothesis if data have a probability of 0.0495. Such sharp distinctions ignore the possibility that an assumed probabilistic process is an inexact portrayal of events that generated the data or the possibility that data give an inexact portrayal of studied phenomena. As Rosnow and Rosenthal (1989, p. 1277) conjectured,

> That is, we want to underscore that, surely, God loves the .06 nearly as much as the .05. Can there be any doubt that God views the strength of evidence for or against the null as a fairly continuous function of the magnitude of p?

Conceptual Problem 2: Most NHSTs Let Apparent Validity of Findings Depend on Researchers' Efforts to Obtain Enough Data

In the late 1930s, Berkson (1938) noticed that he could obtain a statistically significant chi-squared test by increasing sample size. Since then, researchers have found this sensitivity to sample size in all forms of NHSTs. As Mayo (2006, pp. 808–809) expressed the situation, "[w]ith large enough sample size, an α significant rejection of H_0 can be very probable, even if the underlying discrepancy from μ_0 is substantively trivial. In fact, for any discrepancy from the null, however small, one can find a sample size such as there is a high probability (as high as one likes) that the test will yield a statistically significant result (for any p-value one wishes)."

Extreme sample size sensitivity occurs with so-called "point-null hypotheses," which are tested very, very frequently by behavioral and social researchers. A point-null hypothesis defines an infinitesimal point on a continuum. Typical point-null hypotheses postulate that a correlation, frequency, regression coefficient, mean difference, or variance difference equals zero. All "two-tailed tests" of continuous variables incorporate point-null hypotheses because they require a statistic to exactly equal another statistic or a specific number.

A researcher who gathers a large enough sample can reject any point-null hypothesis. This property of NHSTs follows directly from the fact that a point-null hypothesis defines an infinitesimal point on a continuum. For an NHST to reject a point-null hypothesis, the infinitesimal point corresponding to the null hypothesis must fall outside the confidence interval around the sample estimate. As sample size increases, the confidence interval shrinks and becomes less and less likely to include the point corresponding to the null hypothesis.

Imagine a study of two variables that have no relation whatsoever. Capturing these variables involves measurement errors. Such errors might come from conversion of theoretical constructs into measurement instruments, from rounding of measurements, or from errors by people who provide data. Measurement errors mean that the sample estimate of the correlation between the variables is very unlikely to be exactly zero, although it may differ from zero by only a tiny amount. Thus, if current data do not already reject the point-null hypotheses, additional observations will reduce the confidence interval . . . until NHSTs reject the null hypothesis that the correlation is zero.

A central philosophical issue is whether researchers' efforts and motivation should be sufficient to render research findings worthy of being classified as true or not true. A researcher with enough data is certain to find statistically significant results—even if these findings result from noise in data or from a systematic effect too small to have practical or theoretical relevance. Many researchers tailor their data gathering to obtain statistical significance. Webster and Starbuck (1988) found that the mean correlation in studies with fewer than 70 observations is about twice the mean correlation in studies with over 180 observations.

Indeed, where measurement errors are moderately large, statistical significance can come from medium-large samples, and computer-based data management and data analysis facilitate large samples. Thus, modern technology is helping researchers to convert random measurement errors into significant findings. After reviewing articles published in one prestigious journal, Seth et al. (2009) surmised that a substantial fraction of articles has samples large enough to make substantively trivial differences statistically significant.

Conceptual Problem 3: Most NHSTs Disprove Hypotheses That Could Not Possibly Be Correct

Most NHSTs rely on null hypotheses that could not possibly be correct, but when a null hypothesis offers an implausible description of reality, rejecting it provides no information (Lykken 1968). For example, problems such as those at Worldcom and Enron stimulated research to link firms' performance with governance practices or signs of opportunism. However, such studies tested the implausible null hypotheses that governance practices have no effect whatever on firms' performance. As Tukey (1991, p. 100) pointed out, "All we know about the world teaches us that the effects of A and B are always different—in some decimal place—for any

A and B. Thus asking 'Are the effects different' is foolish."

The important research question is not whether any effects occur, but whether these effects are large enough to matter. Generally, the challenge in behavioral and social research is not to find any factors that have even tiny effects on dependent variables, but to identify factors that have substantial effects and to observe the directions of these effects.

Although some researchers believe that NHSTs falsify incorrect hypotheses as Popper (1959) advocated, use of impossible null hypotheses means that NHSTs violate Popper's requirements. Popper's most important criterion was that, to be considered scientific, theories need to perform well in risky tests. Typical significance tests are not risky because null hypotheses are tested rather than researchers' alternative hypotheses.

Some defenders of NHSTs have argued that they would not cause problems if only people would apply them correctly (e.g., Aguinis et al. 2010). However, the conceptual deficiencies of NHSTs are inherent and even when applied correctly, NHSTs do not make reliable differentiations between important and unimportant effects. In addition to conceptual problems, several practical problems arise when researchers try to apply NHSTs. The following sections outline practical problems related to interpretation of NHST results, differentiation between trivial and important findings, violation of statistical assumptions, and effects of NHSTs on researchers' motivation and ethics.

Practical Problem 1: NHSTs Are Difficult to Understand and Often Misinterpreted

NHSTs are difficult to understand because they involve double negatives and null hypotheses that are obviously false. Many people have more difficulty with double negatives than with positive assertions. Disproving the impossible—a meaningless null hypothesis—is such unusual logic that it makes many people uncomfortable, and it should.

A user of an NHST specifies a null hypothesis and then argues that observed data would be very unlikely if this null hypothesis were true. Often, however, elementary logic or direct experience says the null hypothesis cannot be even approximately true: if so, a finding of statistical significance states that observed data would be very unlikely if the impossible would occur.

It is small wonder that many researchers, as well as the public, invent ways to inject sense into this apparent nonsense. One common version of such sensemaking interprets the significance level (e.g., 0.05) as the probability that the null hypothesis is true given the data, Pr(*Null* | *Data*). According to Bayes' theorem, this probability is

$$\Pr(\mathit{Null} \mid \mathit{Data}) = \Pr(\mathit{Data} \mid \mathit{Null}) * [\Pr(\mathit{Null})/\Pr(\mathit{Data})].$$

NHSTs compute the second term in this equation, Pr(*Data* | *Null*), the probability that the data would occur if the null hypothesis were true. However, there is no way to compute Pr(*Null* | *Data*) from knowledge of Pr(*Data* | *Null*) because both Pr(*Null*) and Pr(*Data*) are unknown. Pr(Data) is always unknown. Pr(*Null*) is unknown unless the null hypothesis is impossible, in which case both $\Pr(\mathit{Null}) = 0$ and $\Pr(\mathit{Null} \mid \mathit{Data}) = 0$. However, if a null hypothesis is impossible, one does not need data or a statistical test to reject it.

Empirical research has documented that many people do not understand NHSTs or the term "statistical significance." Studies of misinterpretation have been conducted by Armstrong (2007), Fidler et al. (2005), Hubbard and Armstrong (2006), Haller and Krauss (2002), Oakes (1986), and Vacha-Haase et al. (2000). Researchers frequently publish incorrect interpretations of significance tests, and researchers who review manuscripts often misinterpret them. Researchers may use NHSTs incorrectly because incorrect usage is what they have often seen and believe to be proper. The result is widespread confusion about NHSTs, by the public and by people who have studied statistics, including even some professional statisticians.

Practical Problem 2: NHSTs Highlight Trivial Findings

Another version of sensemaking about NHSTs has researchers or the public mistaking statistical significance for the theoretical importance of a finding or its practical usefulness. Many studies report statistically significant effects that are too small to be of theoretical or practical interest. Seth et al. (2009, p. 5) surveyed papers published in a prestigious journal during 2007. They concluded that "most strategy scholars emphasize only statistical significance as the criterion of importance in examining empirical results, and ignore substantive or economic significance. Only 12% of the empirical studies used other criteria of importance in addition to considering statistical significance using t- or F-statistics."

NHSTs provide only crude discrimination between important findings and unimportant ones. Empirical findings resemble a large haystack that contains both straws and needles, and NHSTs are the sieve that most researchers use to identify needles. To separate needles from straws effectively, researchers need sieves that reject almost all straws while identifying most needles (Hubbard and Armstrong 1992).

Webster and Starbuck (1988) looked at the haystack of relationships studied by organizational researchers and applied psychologists, at least the published part of the haystack. They examined 14,897 correlations obtained by researchers who published in *Administrative Science Quarterly*, the *Academy of Management Journal*, and the *Journal of Applied Psychology.* These were all correlations among all variables studied, not only variables in

researchers' hypotheses. In all three journals, the correlations had very similar distributions, with both the mean and the median correlations close to +0.09. That 69% of the correlations were positive implies that researchers invert scales retrospectively or anticipate signs of relationships prospectively, both of which would invalidate a null hypothesis of zero correlation. To find statistical significance within such distributions of correlations is easy, especially because researchers obtain larger samples when they have smaller correlations. Imagine that a researcher starts with a target variable and searches randomly in these distributions of correlations for a second variable that correlates significantly with the target, using sample sizes that resemble those reported in actual studies. Random search has a 67% chance of finding a statistically significant correlation on the first try and a 96% chance of finding a statistically significant correlation within three tries.

Many editors and reviewers for academic journals are actually making discrimination worse by refusing to publish manuscripts that fail to reject null hypotheses (Greenwald 1975) and refusing to publish successful or failed replications (Starbuck 1994). By not publishing failed replications or failed extensions into slightly different contexts, journals deprive the research community of opportunities to observe such failures (Rousseau et al. 2008, Starbuck 2006). This behavior distorts meta-analyses of multiple studies—a key methodology for aggregating knowledge. Editorial practices also encourage proliferation of theoretical explanations that have dubious empirical support. NHSTs tend to show that an advocated hypothesis is one of many hypotheses consistent with data, a demonstration that is likely to create a premature belief that the advocated hypothesis is the best hypothesis.

The scarcity of replication studies in the social sciences allows NHSTs to confer deceptive importance on random errors, idiosyncratic factors, and very small effects (Hubbard and Armstrong 1992). In medical research, however, a few appraisals indicate that many published studies reported findings that later studies could not replicate (Ioannidis 2003, 2005a, b; Wacholder et al. 2004). Colhoun et al. (2003) estimated that as many as 95% of reported associations between diseases and genetic properties are false positives. Ioannidis (2005a) reported that later research has disconfirmed 37% of the most cited and discussed medical treatments. After several studies of medical treatments that had been falsely overrated at first, Ioannidis (2005b, p. e124) asserted, "There is increasing concern that in modern [medical] research, false findings may be the majority or even the vast majority of published research claims. However, this should not be surprising. It can be proven that most claimed research findings are false."

Practical Problem 3: NHSTs Obscure Important Findings

In addition to the tendency of NHSTs to assign "significance" to trivial findings, NHSTs also classify substantively important findings as "not significant."

Again, medical research has made valuable replication studies. For example, when doctors began to prescribe hormones to counteract menopausal symptoms, initial assessment studies found only weak evidence of harmful effects, which was not statistically significant, and conjectured benefits expanded to include cardiovascular disease, age-related dementias, osteoporosis, and colon cancer. As a result, doctors prescribed hormone therapies for many women. After several years, however, sufficient evidence accumulated to reveal that estrogen and progestin therapies, especially after long-term use, foster breast cancer, strokes, and heart disease (Greiser et al. 2005, Shah et al. 2005). Obviously, women who suffered such consequences may not have regarded them as insignificant.

When outcomes have severe positive or negative consequences, thresholds for considering them worthy of attention should be low. When outcomes have trivial positive or negative consequences, thresholds for considering them worthy of attention should be high. NHSTs with fixed significance thresholds ignore important trade-offs between costs and benefits of research outcomes. Especially troublesome are analytic procedures, such as stepwise regression, that rely on such fixed significance thresholds to choose variables to include in models (Thompson 1995). Such choices equate statistical significance with substantive importance.

Practical Problem 4: NHSTs Make Assumptions That Much Research Does Not Satisfy

Nonreflective use of NHSTs has promoted applications with nonrandom samples or with samples that comprise large fractions of populations. NHSTs with nonrandom samples have no meaningful interpretation because means and variances computed from sample data bear no knowable relationship to the means and variances of the population. Only with genuinely random samples does statistical theory afford researchers a basis for drawing probability inferences about population parameters.

One prevalent misuse of NHSTs occurs when a researcher gains access to data from a complete subpopulation. For instance, Mezias and Starbuck (2003) obtained data from all senior executives in four divisions of a very large company. With such data, the researchers could learn nothing by making NHSTs. They could compute the means and variances of the data from each division exactly. For such statistics, confidence intervals have a width of zero. On the other hand, the researchers had no basis in statistical theory for claims about other executives within or outside the company or for claims about the world population of executives.

Three of the four authors of this essay have had to deal with journal reviewers who demanded NHSTs even though their data comprised either complete populations or large portions of populations. When data include complete populations, sample means are population means and sampling error is zero. Consequently, NHSTs become completely irrelevant. Similar, but less extreme, effects occur when a random sample comprises a large fraction of a population of known size. In such instances, researchers ought to apply correction factors to account for the fact that sample statistics become increasingly good estimators of population statistics when a sample size approaches population size. However, one almost never sees such corrections in published research.

These issues have grown in importance as researchers have gained access to large databases for financial statistics, proxy statements, patents, and other organizational records. For instance, studies of governance, innovation, and top management teams have examined samples such as all Fortune 250 firms, all S&P 500 firms, all publicly traded U.S. manufacturing companies, or all U.S. patents issued during a specific time period. These are not random samples. If population data are complete, such studies examine populations, and NHSTs provide no information about possible findings during other periods or in other populations. If data are incomplete, missing data are more likely to have common properties than to be random.

Practical Problem 5: NHSTs Corrode Researchers' Motivation and Ethics

The most harmful effect of NHSTs may be erosion of researchers' devotion to their vocations. Repeated and very public misuse of NHSTs creates cynicism and confusion. Unjustified applications of NHSTs bring rewards, and justified deviations from these practices attract extra scrutiny followed by rejection. Frequently, research seminars drift into debates about statistical nuances while participants ignore the substantive importance of findings. Success in research can become a mere game played to achieve promotion or visibility, not a high calling in pursuit of useful knowledge and societal benefit.

Of course, any methodology could create harmful effects if many people misuse it consistently, and some individual researchers will always embrace game playing. However, NHSTs have especially troublesome properties, both conceptual and practical. Collectively, these properties make fertile ground for disillusionment and cynical opportunism.

The problems associated with NHSTs and their harmful effects create mysteries. Why have researchers persisted in using such troublesome methods? Why have researchers failed to adopt better ways to assess research?

Why Do So Many Researchers Cling to NHSTs?

Not everyone uses or endorses NHSTs. During recent years, NHSTs have drawn active opposition in biology, education, forecasting, medicine, and psychology (e.g., Armstrong 2007, Cohen 1994, Cortina and Folger 1998, Schmidt 1996, Starbuck 2006). Yet NHSTs have continued to dominate statistical practice in the life, behavioral, social, and economic sciences. Unfortunately, many researchers believe NHSTs are adequate. Methodology courses do not teach alternatives to NHSTs. Institutionalized practices tolerate or endorse NHSTs. Even researchers who are aware of NHSTs limitations tend to underestimate the detrimental impacts of NHSTs.

For most researchers, the apparent adequacy of NHSTs has roots in misconceptions. For example, the so-called "inverse probability fallacy" leads researchers to believe that p denotes Pr(Null | Data), the probability that the null hypothesis is true given the data. This mistake fosters a second incorrect inference, that $1 - p$ equals the probability that researchers' alternative hypothesis is true. Once researchers believe they know the probabilities of their null hypotheses and their alternative hypotheses being true, what other information could they possible want? Indeed, researchers often take still another unjustifiable leap: they surmise that $1 - p$ is the probability that their substantive theories are correct. This extrapolation assumes that the only alternative to the null hypothesis is the alternative that researchers themselves articulated.

Schmidt and Hunter (1997, p. 37) identified around 80 commonly raised objections to discontinuation of NHSTs and argued that none of the objections has validity. Other proponents of change have pointed to psychological or social reasons. In a personal communication, Meehl (2002) blamed "plain psychic inertia." He said, "If one has been thinking in a certain way since he was a senior in college, ... there is a certain intellectual violence involved in telling a person ... that they've been deceiving themselves." Thompson (1999a, p. 135) argued that substituting statistical significance for theoretical or practical importance allows researchers to "finesse the responsibility for and necessity of declaring and exposing to criticism the personal or societal values that inherently must be the basis for any decree that research results are valuable." Likewise, John (1992) proposed that researchers use statistical significance to portray their work as "objective" and "scientific" because the tests substitute for decisions about whether phenomena are real or effects important. Regarding lack of reporting of confidence intervals, John also said that because so much behavioral and social research produces ambiguous findings, stating wide confidence intervals exposes researchers to embarrassment and undermines their claims to knowledge.

Arguments such as those above place responsibility for methodological choices on individual researchers, and they understate the influence of widespread social norms. The very prevalence of NHSTs has become a major reason for their continued use. Researchers who use NHSTs receive support from their colleagues, journal editors and reviewers, and public media. Researchers who eschew NHSTs have to justify their deviant choices and risk having manuscripts rejected.

Methodologists in education, medicine, and psychology have asked their professional associations to eliminate NHSTs from their journals (Fidler 2005, Fidler et al. 2004). In the mid-1990s, several psychologists well known for their methodological contributions urged the American Psychological Association (APA) to ban NHSTs from its journals, and the APA appointed a task force to develop new recommendations about statistical inference. However, after a brief initial meeting, the task force promptly announced that it "does not support any action that could be interpreted as banning the use of null hypothesis significance testing or p values" (Task Force on Statistical Significance 1996, p. 2). A later, second report by the task force went further in its recommendations, but still short of banning NHSTs (Wilkinson 1999). Finally, the latest version of the American Psychological Association (2010, p. 34) publication manual states:

> For the reader to appreciate the magnitude or importance of a study's findings, it is almost always necessary to include some measure of effect size in the Results section. Whenever possible, provide a confidence interval for each effect size reported to indicate the precision of estimation of the effect size.

Insights from Medicine's Reform

Medical research offers a precedent of rather successful statistical reform. Although some medical researchers still use NHSTs, medicine has moved away from sole reliance on NHSTs. Nearly all medical studies now state confidence intervals, and researchers attempt to estimate the substantive importance of their findings (Fidler et al. 2004).

One force furthering change was strong interventions by journal editors. The most visible and controversial of these editors was Kenneth J. Rothman. As editor of the *American Journal of Public Health*, Rothman's revise-and-resubmit letters to authors stated, "All references to statistical hypothesis testing and statistical significance should be removed from the paper. I ask that you delete p values as well as comments about statistical significance. If you do not agree with my standards (concerning the inappropriateness of significance tests), you should feel free to argue the point, or simply ignore what you may consider to be my misguided view, by publishing elsewhere" (Shrout 1997, p. 1; see also Fleiss 1986, p. 559). Later, Rothman (1998, p. 334) became editor of another journal, where he announced, "When writing for *Epidemiology*, you can enhance your prospects if you omit tests of statistical significance . . . . In *Epidemiology*, we do not publish them at all. Not only do we eschew publishing claims of the presence or absence of statistical significance, we discourage the use of this type of thinking in the data analysis, such as in the use of stepwise regression." During 2000, *Epidemiology* published not a single p-value, and 94% of empirical articles reported confidence intervals (Fidler et al. 2004).

Surprisingly, Rothman's policies established behavioral patterns that persisted after he left those journals, and they influenced the policies of other journals. Opposition to NHSTs continued for many years, and it came from many medical researchers, journal editors, and societies. Rather than offering mere suggestions, editors of medical journals spoke of "requirements" and "expectations." For example, Langman (1986, p. 716) at the *British Medical Journal* (*BMJ*) said, "from 1 July authors of papers submitted to the *BMJ* will be expected to calculate confidence intervals whenever the data warrant this approach."

Editorial policies may have to be quite strict to elicit behavioral change. In contrast to editors of medical journals, editors of psychology journals have generally encouraged behavioral change instead of requiring it. For instance, when Kendall (1997, p. 3) tried to enact changes at the *Journal of Consulting and Clinical Psychology* (*JCCP*), he advised authors as follows: "Evaluations of the outcomes of psychological treatments are favorably enhanced when the published report includes not only statistical significance and the required effect size but also a consideration of clinical significance." His encouragements had much weaker effects than Rothman's requirements. Just 40% of *JCCP*'s authors reported on clinical significance (Fidler et al. 2004). Thompson (1999b, p. 162) argued that mere encouragement amounts to a "self-cancelling message." He said, "To present an 'encouragement' in the context of strict absolute standards regarding the esoterics of author note placement, pagination, and margins is to the send the message, 'these myriad requirements count, this encouragement doesn't.' "

Requirements, bans, or mandates about statistical reporting have often drawn negative reactions. Even some advocates of statistical reform in psychology have viewed requirements as impinging on researchers' intellectual freedom. Although embedded norms that support NHSTs also limit academic freedom, researchers and the public have become accustomed to their effects.

After a comprehensive study of efforts to change statistical practices in ecology, medicine, and psychology, Fidler et al. (2004, p. 615) concluded, "The nature of the editorial policies and the degree of collaboration amongst editors are important factors in explaining the

varying levels of reforms in these disciplines. But without efforts to also re-write textbooks, improve software and research understanding of alternative methods, it seems unlikely that editorial initiatives will achieve substantial statistical reform." Capraro and Capraro (2002) found that statistical textbooks still strongly emphasize statistical significance testing over effect size estimation. Indeed, a third of the textbooks did not cover effect-size estimation at all.

Yet another factor lurks just offstage during discussions of why medical research has changed and behavioral and social research has not. Medical research is more expensive, receives much more funding, makes more of a difference to more people, and draws much more attention. Thus, medical researchers have greater incentive to measure and document effects of their work and to avoid promulgating treatments that turn out later to have been ineffective or harmful.

The next section recommends methodological changes to improve on NHSTs. These recommendations do not represent a comprehensive agenda for methodological change, but they provide guidance for individual researchers who are interested in advancing their research methodology and a starting point for more comprehensive methodological discussions and institutional change.

How Can Researchers Do Better?

Any nonreflective way of assessing research is destined to prove ineffective for the entire range of behavioral and social sciences because it cannot accommodate diverse contingencies and exhibit a spectrum of nuances. Any approach to research assessment that allows for contingencies and nuances has to meet challenges from different consumers of research.

For example, because studies had suggested that many health-care professionals give patients incorrect or confusing advice about nutrition, Cadman and Findlay (1998) investigated effects of training on nurses' knowledge about nutrition. They assessed nurses' knowledge, provided training to the nurses, and then reassessed the nurses' knowledge. On average, nurses scored 21% higher on the reassessment, and nurses' confidence in their knowledge rose from 27% to 88%. These changes led the researchers to propose that the nurses' employer should provide such training.

NHSTs for such a problem would test the point-null hypothesis that training has no effect at all, and most researchers would interpret statistical significance as adequate evidence that training is useful. Of course, NHSTs do not justify such a conclusion. The relevant question is not whether training had any effect, but whether the effect was strong enough and consistent enough to justify using organizational resources for such training.

In addition, a focus on statistical significance tends to suppress reporting of nuances. In the study of nurses, two dieticians trained 59 nurses working in 30 medical practices, so researchers could have described variations across individuals, sites, and trainers. For instance, across practices measured, change in knowledge ranged from -23% to $+73\%$. Either some nurses actually exhibited less knowledge after training or measurements of knowledge lacked reliability.

The study of nurses did not assess consequences for patients; researchers merely assumed that patients would benefit. They also assumed that nurses who are more confident of their knowledge possess better knowledge; they did not examine the correlation between confidence and the correctness of knowledge. Because nurses' confidence rose much more dramatically than their knowledge scores, and because only a few nurses scored above 80% on the test of knowledge even after training, training may have created unjustified confidence with potential negative consequences for misinformed patients.

Good research requires using different methodologies and assessment criteria in different contexts and probing deeply for diverse implications. A single methodology is likely to be inappropriate for many, if not most, studies. Thus, the sections to follow describe versatile practices that promise to improve on use of NHSTs. These recommendations concern assessment of findings, and they follow rather directly from problems discussed above.

Recommendation 1: Tailor Assessment to Research Context

An apparent advantage of statistical significance is that researchers can describe it in much the same way no matter what contexts or phenomena they study. Participants in research seminars believe they can understand presented findings without much knowledge of studies' variables or contexts. Unfortunately, such research descriptions are superficial and apparent comprehension is illusory. People are talking and using familiar words without appreciating how these meanings shift from context to context.

To prevent superficial assessment, researchers need to account for relationships between contexts they study and actions that their findings might stimulate (Breaugh 2003). The following questions suggest starting points for giving research findings more meaning.

What Metrics Make Sense for Dependent Variables? Researchers should describe the effects on dependent variables in the same units that they use to measure the dependent variables—tons, numbers of people, bales, or barrels.

In the special case of random samples with arbitrary scales, researchers can standardize all variables and describe effects probabilistically. For example, a researcher might say that, ceteris paribus, a one-standard-deviation change in student motivation produces a change

in knowledge confidence that is between −0.16 and +0.23 standard deviations. Some methodologists have been seeking dimensionless measures of effect size such as likelihood ratios or correlation coefficients. However, researchers should remain cautious about unobserved or unremarked differences between studied settings that can create deceptive illusions of comparability across studies.

For Whom Do Effects Have Relevance? The researchers who studied nutrition knowledge wanted to improve patients' health, but they obtained data about nurses, not patients. None of their data related directly to patients' health. However, if the researchers had tried to measure changes in patients' health, the connection to training would have been remote, and the number and importance of confounding influences would have been high. Aguinis et al. (2010) have recommended that researchers ought to distinguish between effect size and practical significance and assess them separately.

Should Researchers Relate Benefits of Effects to the Costs of Those Effects? Training of nurses in nutrition is not costless; at a minimum, nurses could be learning other information or skills. Findings stated in cost/benefit terms have more direct relevance for decisions, so when their studies do not capture cost and benefits directly, researchers should consider estimating costs and benefits based on anecdotal or simulation data. To compare benefits with costs, researchers need to state changes in dependent and independent variables in comparable units. However, benefits and costs are often multidimensional, and equivalence can be difficult to establish. For example, training of nurses creates both monetary costs and opportunity costs in terms of nurses' time, and neither of these costs translates readily into the value of nutrition knowledge for patients' health.

Would Multiple Assessments Be Informative? In most studies, different indicators reveal complementary aspects of findings. Researchers' challenge is to enhance readers' understanding by balancing simplicity against depth. Simplicity enhances clarity, whereas complexity fosters future research and further development of indicators.

Recommendation 2: Report Uncertainty Associated with Effect Size

"Effect size" denotes an attempt to estimate the change in a dependent variable that results from change in an independent variable. For example, after training, a researcher might say that, nurses' knowledge scores increased by an average of 21%, but 95% confidence limits for individual nurses ranged from a loss of −41% to a gain of +95%.

Researchers need to think creatively about appropriate ways to estimate effects in their studies. Although researchers have proposed several indicators for effect size (Cortina and Nouri 1999, Ellis 2010, Grissom and Kim 2005), many proposed indicators focus on differences between two discrete treatments, whereas much behavioral and social research does not compare discrete treatments. Researchers should also beware that proponents of various indicators have tended to propose using them with NHSTs.

Researchers should report the uncertainty attending their findings. When data are random samples, one way to communicate this uncertainty is reporting of confidence intervals. Thus, many methodologists and an increasing number of journals recommend reporting confidence intervals for effect sizes (American Educational Research Association 2006, American Psychological Association 2010). Various methodologists distribute software that can perform such calculations (e.g., Algina and Keselman 2003, Cumming and Finch 2001, Smithson 2001, Steiger and Fouladi 1992, Thompson 2002). A disadvantage of confidence intervals is that researchers can interpret them as justifying binary judgments about what is true or false, and thus to make NHSTs covertly.

Instead of confidence intervals, researchers can report likelihood ratios, posterior probability distributions, or entire distributions of inferences (Jeffreys and Berger 1992). For example, Soofi et al. (2009) analyzed executives' expectations about economic change, the impact on their firms, and their firms' possible responses. The researchers used graphs to show estimated values of uncertainty across all 93 executives and to show how researchers' assumptions alter inferences about distributions of uncertainty across executives. Bayesian regression analysis led to graphs showing probability distributions of estimated regression coefficients. Even without making the large conceptual jump from NHSTs to Bayesian inference, researchers can use simple graphs to communicate both size of effects and their variability, as the next section describes.

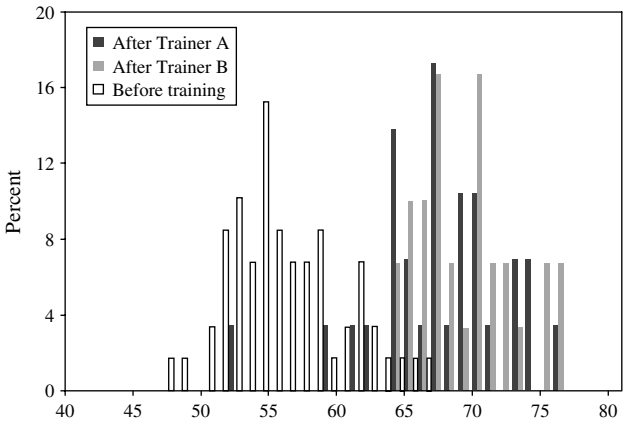
Recommendation 3: Explain and Illustrate Assessment Indicators

Prevalence and general acceptance of NHSTs have fostered an impression that researchers do not have to explain how they assess research findings. Even researchers who use NHSTs should explain why they believe their data satisfy NHSTs' assumptions and what their p-values say about their findings.

To compare treatments or contexts, or to explicate interaction effects, it is useful to graph means, confidence limits, or distributions of possible inferences. Cleveland (1985), Darlington (1973), Doksum (1977), and Wilk and Gnanades (1968) have proposed graphical methods for distributions of effect size.

Figure 1 shows a conjectured extrapolation of the study of nurses' knowledge about nutrition. The hollow columns represent test scores of 59 nurses before training, and the two kinds of solid columns show test scores

Figure 1 Distributions of Knowledge Scores



after training by either of two trainers. The figure postulates that Trainer B is 25% more effective on average than Trainer A is. Such a figure conveys much more information than would mere numbers, such as means or confidence intervals, and it gives audiences a more intuitive appreciation for findings.

Recommendation 4: Compare New Data with Naïve Models Rather Than Null Hypotheses

In place of null hypotheses, researchers can compare their proposed theories with naïve hypotheses that require no understanding of studied phenomena. In contrast to null hypotheses that claim no effect, naïve hypotheses assume a simple effect that occurs, for example, because of stability of social processes, effects of third variables, or random processes. Connor and Simberloff (1986, p. 160) defined a naïve hypothesis (which they called a null model) as "an attempt to generate the distribution of values for the variable of interest in the absence of a putative causal process." Thus, naïve hypotheses are not supposed to provide satisfying explanations, but to offer stronger competition than null hypotheses do. Stronger competition challenges researchers to develop theories that explain more.

Researchers have tested their fields' knowledge against several types of naïve hypotheses. One type proposes that data arise from very simple random processes. In organizational ecology, for example, conceptual arguments suggested higher survival rates for larger and older organizations. Early studies applied traditional NHSTs and rejected the null hypotheses that organizational size and age had no effect on survival rates. However, Levinthal (1991) argued that observed differences between survival rates across organizational size and age are qualitatively consistent with the naïve hypothesis that organizational survival is a random walk. He (1991, p. 416) concluded that a random walk provides a baseline for assessing causal effects of organizational size and age, which exposes more subtle features of size and age dependence.

Powell (2003) combined random processes with other naïve comparisons. Much research has investigated persistence of exceptional performance by business firms, and traditional NHSTs rejected the null hypothesis that all firms perform equally well. Powell (2003) compared data about success patterns among Fortune 500 firms with several naïve hypotheses about the distribution of performance. He produced naïve hypotheses *analytically* (based on simple Pareto-like growth models), *empirically* (based on comparisons to other nonbusiness competitive domains, such as sports, politics, or beauty pageants), and by *simulation* (based on stochastic processes). When he used these naïve hypotheses, he surmised "that nothing unusual is happening in the performance of most industries" (Powell 2003, p. 83). "If firm-specific competitive advantages exist, they are, in all likelihood, local and extreme phenomena, and highly resistant to useful generalization" (Powell 2003, p. 83).

Another type of naïve hypotheses conjectures that crude hypotheses provide at least as much useful information as subtle hypotheses. For example, researchers tested elaborate forecasting models against two naïve hypotheses: (1) tomorrow will be the same as today and (2) the trend since yesterday will continue until tomorrow (Elliott 1973, Makridakis et al. 1982, Pant and Starbuck 1990). Thus, longitudinal research designs should consider not only random-based change patterns but also state-based and trend-based naïve hypotheses.

Another useful standard for comparison can be the crude hypothesis that every independent variable exerts the same influence on the dependent variable. Using both computer simulation and algebraic analyses, psychometricians have discovered (1) that, on average, naïve "same effect" hypotheses make better predictions about new samples than multiple regression does unless the regressions are based on large samples (e.g., $n = 160$ to 400 or larger), and (2) that even when calculated from very large samples, regression coefficients make predictions that are only slightly better on average than those made by the "same effect" hypotheses (Claudy 1972, Dorans and Drasgow 1978, Einhorn and Hogarth 1975, Schmidt 1971). The predictive effectiveness of such naïve hypotheses implies that researchers who gather small samples could make predictions that are more accurate if they did not even gather data.

Recommendation 5: To Support Generalization and Replicability, Frame Hypotheses Within Very Simple Models

Researchers often introduce numerous independent variables into their analyses. They assume that models with more variables are more accurate because they account for more possible influences on data including contingencies and peculiarities of specific situations. However, this argument has serious weaknesses. To estimate coefficients with reliable accuracy, regression requires independent variables that correlate only weakly or not at

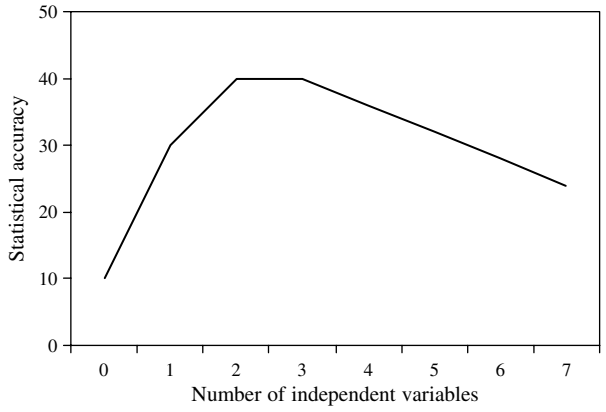
all. When two independent variables correlate, errors in estimates of one regression coefficient can offset errors in estimates of the other coefficient. Thus, the two coefficients may jointly yield a good fit to data even though the individual coefficient estimates are quite inaccurate. Each variable added to represent another influence or idiosyncrasy correlates (if only slightly) with other independent variables, so regression calculations grow more likely to make unreliable estimates as numbers of independent variables increase. Although these effects distort NHSTs, they also distort estimates of effect size and confidence intervals, so researchers have reason to simplify their analytic models no matter what assessments they intend to make.

Of course, such effects vary across situations, and the quality of research findings depends on the quality and properties of data as well as the models used for analysis. Sometimes, statistical procedures can help to address multicollinearity concerns (Thompson 2006).

However, there are reasons to expect parsimonious models to be both more accurate and more understandable. When numbers of independent variables increase, regression calculations climb and descend Ockham's hill, an effect named for William Ockham, a 14th-century advocate of parsimonious theorizing. Figure 2 outlines the general idea. A model that includes too few independent variables fits sample data too loosely: it fails to capture important and explainable variation, and it makes inaccurate inferences about the population. However, additional variables have diminishing returns. When a model starts to include too many independent variables, it fits data too tightly: regression coefficients are more likely to describe random noise or idiosyncratic properties that do not generalize to the population even if the added variables have statistically significant coefficients.

Gauch (2002, 2006) studied Ockham's hills of biological studies via simulations, and he found that the models that give the most accurate generalizations are quite parsimonious. To reduce effects of correlations among independent variables, Gauch first used factor analysis to group independent variables into correlated clusters. Then he compared predictions made by regression equations with population properties assumed when generating the original data. His studies indicate that only two or three such clusters of variables are optimal for making accurate statements about populations.

Figure 2 Ockham's Hill



Large numbers of independent variables also reduce the ability of researchers and their audiences to make logical or intuitive sense of findings (Goldberg 1970, Meehl 1954). Even researchers who advocate multivariate analyses revert to bivariate and trivariate interpretations when they communicate their findings. When Box and Draper (1969) used experiments to improve factory efficiency, they deduced that practical experiments should alter only two or three variables at a time because people had trouble interpreting outcomes of experiments involving four or more variables. Similarly, Faust (1984) observed that scientists have difficulties understanding four-way interactions. He remarked that the greatest theoretical contributions in the physical sciences have exhibited parsimony and simplicity rather than complexity, and he speculated that parsimonious theories have been very influential not because the physical universe is simple but because people can understand simple theories.

Model parsimony is another area in which social norms appear to be degrading research quality. Application of NHSTs has led researchers to test more and more complex models, and ease of collecting and analyzing larger samples has stimulated inclusion of additional variables. Journal reviewers frequently suggest that researchers add more control variables. An unintended outcome has been models that overfit data and findings that are less likely to generalize and replicate.

Behavioral and social scientists have not given parsimony the respect it deserves. Insofar as people and organizations can choose their characteristics (e.g., educations, geographic locations, governance modes, top management teams), random sampling tends to produce correlated variables, which reduce the reliability of statistical analyses. In addition, insofar as people and organizations learn, including learning from reading research studies, replication becomes very difficult if not impossible. Technologies change endlessly, as do economic conditions and political structures and policies. Consequently, sample data come from populations that soon will no longer exist. To formulate useful generalizations, researchers need to focus on the most fundamental, pervasive, and inertial causal relations. To guide human action, researchers need to develop parsimonious and simple models that humans understand.

Recommendation 6: Use Robust Statistics to Make Estimates, Especially Robust Regression

Many statistical methods, including NHSTs, assume that actual population distributions match hypothetical

distributions. These methods give unreliable results when applied to samples from populations that do not satisfy their assumptions or when samples are too small to provide good representations of their populations.

For example, all statistical methods that rely heavily on the squaring of error terms have problems because this squaring raises the influence of low-probability extreme events (outliers). In particular, ordinary least-squares regression (OLS) may yield inaccurate coefficient estimates when sample sizes are smaller than 400 even if sampled populations satisfy the OLS's assumptions (Einhorn and Hogarth 1975, Starbuck 2006).

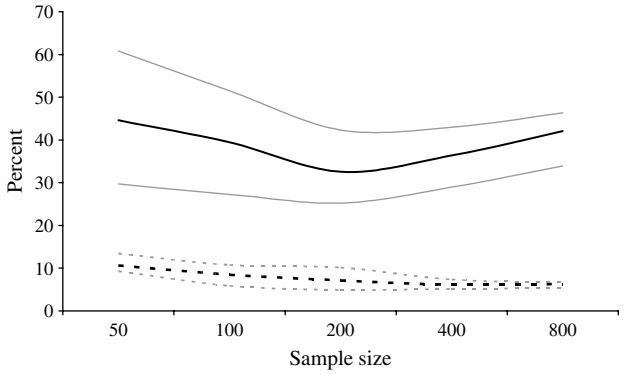
If independent variables have skewed distributions, error terms for regression analyses converge toward a Normal distribution more slowly as sample size increases, so regressions are likely to require larger samples to produce accurate estimates, and the likelihood of outliers is higher. Especially troublesome are distributions with long tails because they increase the probability of outliers. Consequently, the plausibility of assuming Normality of the error term in regressions depends upon sample size and the populations from which data arise. When samples deviate from the normality assumptions of OLS, estimates of regression coefficients and their statistical significance become more inaccurate.

Researchers can investigate the sensitivity of their inferences to properties of their data. For example, they can make trimmed least-squares estimates with different amounts of trimming, or they can test the robustness of their findings by selectively excluding some observations from their analyses. However, when there are several independent variables, it becomes difficult to distinguish outliers from other data.

Therefore, over the last three decades, statisticians have been developing estimation methods that exhibit robustness in the sense that they produce more accurate estimates than traditional methods such as OLS and *t*-tests (Keselman et al. 2007, Rousseeuw and Leroy 1987, Wilcox 1997). The most flexible of these methods adapt automatically to data in that they behave like traditional methods when data satisfy the assumptions of those methods but behave differently when data violate these assumptions. For example, robust MM regression limits the effects of extreme outliers, but when sample data do not include extreme outliers, robust MM regression produces the same coefficient estimates as OLS. (Robust MM regression was developed as a modification of maximum likelihood estimation, or M estimation. The second M in MM estimation symbolizes use of a two-stage process: first choosing a scale parameter to minimize a function that estimates potential loss due to outliers, and then making a maximum likelihood estimate.)

Extremely dangerous for conventional statistical analyses are outliers that result from large errors in sample data. OLS has the serious liability that a single large error in data can greatly distort inferences. In general, measurement errors in independent variables are more likely to cause serious distortions than errors in dependent variables. Audits of frequently used financial databases have found that (1) companies had errors in their accounting, (2) companies reported their accounting data incorrectly, and (3) clerks doing data entry made typing errors. San Miguel (1977), for example, reported a 30% error rate for research and development expenditures on Compustat. Rosenberg and Houglet (1974) found that about 2.4% of the stock prices reported by Compustat and by the Center for Research in Security Prices at the University of Chicago contained errors. Although many of these errors are too small to cause serious concerns, about 0.5% of the stock price errors were large enough to invalidate inferences. Rosenberg and Houglet (1974, p. 1303) concluded, "... there are a few large errors in both data bases, and these few errors are sufficient to change sharply the apparent nature of the data." Again, robust regression procedures have the advantage of deemphasizing extreme outliers caused by errors.

Figure 3 Percentage Errors in Estimated Regression Coefficients with 0.5% of Data Having Extreme Errors



Source. Adapted from Starbuck (2006).

Figure 3 compares the errors in regression coefficients when estimated by OLS (solid lines) and by robust MM (dashed lines) when data contain serious errors. These calculations used sample data in which 0.5% of all variables incorporate data entry errors that shift the decimal point one digit to the right; that is, a random 0.5% of the recorded data are 10 times the correct data. Heavy lines show average errors, and the light lines show quartiles for the errors; 25% of the errors in coefficient estimates fall above the upper quartile lines, 25% fall below the lower quartile lines, and 50% fall between the two quartile lines. The simulations support statisticians' claims that robust MM regression does much better than OLS regression at ignoring unreliable data.[1]

The Opportunity to Change

We began this essay by describing a professor's struggle with institutional pressures that enforced ritualistic

use of NHSTs and a strong bias to equate "statistically significant" with "important for theory or practice." A comprehensive discussion of institutional factors and processes that have locked large parts of the social sciences into the unreflective application of NHSTs would require another essay. Although NHSTs have been remarkably enduring in the face of escalating criticism, we believe change to be inevitable, even if painfully slow.

Progressive defenders of NHSTs continue to argue that wholesale change is not necessary, that it is possible to combine NHSTs usefully with measures of effect size (Aguinis et al. 2010). Conceptual and practical problems outlined in this paper show why such a combination is undesirable. No one has proposed changes to NHSTs that purport to correct the main problems, defenders have been growing fewer, and even defenders acknowledge that NHSTs have deficiencies. Arguments supporting NHSTs appeal to values that seem less than admirable, such as adherence to tradition, resistance to change, convenience of standardization, and disregard for uncertainty.

Critics of NHSTs have been increasing in numbers, and their complaints have been growing more visible, so more and more researchers are becoming aware of the deficiencies of NHSTs. Arguments against NHSTs appeal to values that seem more admirable—ability to cope with complexity, sincerity, willingness to learn, and desire to report findings that matter. Whereas methodologists and researchers have asked their professional societies to ban NHSTs, no one has asked their professional societies to put more emphasis on NHSTs. Two dozen journals in psychology and education now require authors to report effect sizes instead of or in addition to significance tests, and several books and articles have appeared that explain how to compute effect sizes (Algina and Keselman 2003, Breaugh 2003, Cortina and Nouri 1999, Ellis 2010, Grissom and Kim 2005, Smithson 2001).

In spite of these changes, institutional pressures are still strongly supporting the ritualistic use of NHSTs as the default instrument to assess research findings. Research that avoids NHSTs and instead reports effect sizes or confidence intervals, or draws on alternative statistical approaches (e.g., Bayesian statistics), continues to face higher levels of scrutiny and substantial skepticism in review processes. We believe that institutional change, such as changes in review processes, needs grassroots support from individual researchers.

You do not have to wait patiently for others to bring better methodology into your world. When null hypotheses could not possibly be true, you can remark that those NHSTs show only that the impossible did not happen. When research examines a population or a nonrandom sample, you can indicate that NHSTs are inappropriate. When findings are not overwhelmingly conclusive, you can suggest that uncertainty surrounds what is true and what is false. When findings are not statistically significant but they might hold substantive importance, you can highlight their potential importance. When researchers do not report effect sizes, you can ask how big the effects are. Perhaps most importantly, when your colleagues offer such observations, you can support them. Such grassroots support can push necessary institutional changes and help to keep studies with important findings out of file drawers labeled "disaster paper."

> What of the analyst, who may even be a statistician, who says, "This is all about words—I may use the bad words, but I do always think the proper thoughts, and always act in the proper way!" We must reject such a claim as quite inadequate. Unless we can learn to keep what we say, what we think, and what we do all matching one another, and matching a reasonable picture of the world, we will never find our way safely through the thickets of multiple comparisons—and we will not serve ourselves, our friends, and our clients adequately.
>
> (Tukey 1991, p. 100)

Acknowledgments

This essay benefited from comments and suggestions from Linda Argote, Ann Connell, Sergio Koreisha, Walter Nord, Tom Powell, Victor Seidel, and Bruce Thompson.

Endnote

[1]Figure 3 is based on 100 samples—20 samples for each of five sample sizes. Curvature of the OLS's accuracy depends on error rates; relates nonlinearly OLS's accuracy to sample size because smaller samples have lower probabilities of including rare errors. Starbuck (2006, pp. 163–164) gives more details about these simulations.

References

Aguinis, H., S. Werner, J. L. Abbott, C. Angert, J. H. S. Park, D. Kohlhausen. 2010. Customer-centric science: Reporting significant research results with rigor, relevance, and practical impact in mind. *Organ. Res. Methods* **13**(3) 515–539.

Algina, J., H. J. Keselman. 2003. Approximate confidence intervals for effect sizes. *Ed. Psych. Measurement* **63**(4) 537–553.

American Educational Research Association. 2006. Standards for reporting on empirical social science research in AERA publications. *Ed. Res.* **35**(6) 33–40.

American Psychological Association. 2010. *Publication Manual of the American Psychological Association*. American Psychological Association, Washington, DC.

Armstrong, J. S. 2007. Significance tests harm progress in forecasting. *Internat. J. Forecasting* **23**(2) 321–327.

Berkson, J. 1938. Some difficulties of interpretation encountered in the application of the chi-square test. *J. Amer. Statist. Assoc.* **33**(203) 526–536.

Bowman, E. H. 1984. Content analysis of annual reports for corporate strategy and risk. *Interfaces* **14**(1) 61–71.

Box, G. E. P., N. R. Draper. 1969. *Evolutionary Operation*. Wiley, New York.

Breaugh, J. A. 2003. Effect size estimation: Factors to consider and mistakes to avoid. *J. Management* **29**(1) 79–97.

Cadman, L., A. Findlay. 1998. Assessing practice nurses' change in nutrition knowledge following training from a primary care dietitian. *J. Royal Soc. Promotion Health* **118**(4) 206–209.

Capraro, R. M., M. Capraro. 2002. Treatments of effect sizes and statistical significance tests in textbooks. *Ed. Psych. Measurement* **62**(5) 771–782.

Claudy, J. G. 1972. Comparison of five variable weighting procedures. *Ed. Psych. Measurement* **32**(2) 311–322.

Cleveland, W. S. 1985. *The Elements of Graphing Data*. Wadsworth, Monterey, CA.

Cohen, J. 1994. The earth is round ($p < 0.05$). *Amer. Psych.* **49**(12) 997–1003.

Colhoun, H. M., P. M. McKeigue, G. Davey Smith. 2003. Problems of reporting genetic associations with complex outcomes. *Lancet* **36**(9360) 865–872.

Connor, E. F., D. Simberloff. 1986. Competition, scientific method, and null models in ecology. *Amer. Sci.* **74**(2) 155–162.

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

Cortina, J. M., H. Nouri. 1999. *Effect Size for ANOVA Designs*. Sage, Beverly Hills, CA.

Cumming, G., S. Finch. 2001. A primer on the understanding, use and calculation of confidence intervals that are based on central and noncentral distributions. *Ed. Psych. Measurement* **61**(4) 532–574.

Darlington, M. L. 1973. Comparing two groups by simple graphs. *Psych. Bull.* **79**(2) 110–116.

Doksum, K. A. 1977. Some graphical methods in statistics: A review and some extensions. *Statistica Neerlandica* **31**(2) 53–68.

Dorans, N., F. Drasgow. 1978. Alternative weighting schemes for linear prediction. *Organ. Behav. Human Perform.* **21**(3) 316–345.

Einhorn, H. J., R. M. Hogarth. 1975. Unit weighting schemes for decision making. *Organ. Behav. Human. Perform.* **13**(2) 171–192.

Elliott, J. W. 1973. A direct comparison of short-run GNP forecasting models. *J. Bus.* **46**(1) 33–60.

Ellis, P. 2010. *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis and the Interpretation of Research Results*. Cambridge University Press, Cambridge, UK.

Faust, D. 1984. *The Limits of Scientific Reasoning*. University of Minnesota Press, Minneapolis.

Fidler, F. 2005. From statistical significance to effect estimation: Statistical reform in psychology, medicine and ecology. Doctoral dissertation, University of Melbourne, Melbourne, VIC, Australia.

Fidler, F., G. Cumming, M. Burgman, N. Thomason. 2004. Statistical reform in medicine, psychology and ecology. *J. Socio-Econom.* **33** 615–630.

Fidler, F., G. Cumming, N. Thomason, D. Pannuzzo, J. Smith, P. Fyffe, H. Edmonds, C. Harrington, R. Schmitt. 2005. Evaluating the effectiveness of editorial policy to improve statistical practice: The case of the journal of consulting and clinical psychology. *J. Consulting Clinical Psych.* **73**(1) 136–143.

Fiol, C. M. 1989. A semiotic analysis of corporate language: Organizational boundaries and joint venturing. *Admin. Sci. Quart.* **34**(2) 277–303.

Fisher, R. A. 1925. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, UK.

Fleiss, J. L. 1986. Significance tests do have a role in epidemiological research: Reactions to A. A. Walker. *Amer. J. Public Health* **76**(5) 559–560.

Gauch, H. G. 2002. *Scientific Method in Practice*. Cambridge University Press, Cambridge, UK.

Gauch, H. G. 2006. Winning the accuracy game. *Amer. Sci.* **94**(2) 135–143.

Goldberg, L. R. 1970. Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inference. *Psych. Bull.* **73**(6) 422–432.

Greenwald, A. G. 1975. Consequences of prejudice against the null hypothesis. *Psych. Bull.* **82**(1) 1–20.

Greiser, C. M., E. M. Greiser, M. Dören. 2005. Menopausal hormone therapy and risk of breast cancer: A meta-analysis of epidemiological studies and randomized controlled trials. *Human Reproduction Update* **11**(6) 561–573.

Grissom, R. J., J. J. Kim. 2005. *Effect Sizes for Research: A Broad Practical Approach*. Lawrence Erlbaum Associates, Mahwah, NJ.

Haller, H., S. Krauss. 2002. Misinterpretations of significance: A problem students share with their teachers? *Methods Psych. Res.* **7**(1) 1–20.

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

Hubbard, R., J. S. Armstrong. 2006. Why we don't really know what statistical significance means: Implications for educators. *J. Marketing Ed.* **28**(2) 114–120.

Hubbard, R., M. J. Bayarri. 2003. Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *Amer. Statistician* **57**(3) 171–178.

Ioannidis, J. P. A. 2003. Genetic associations: False or true? *Trends Molecular Med.* **9**(4) 135–138.

Ioannidis, J. P. A. 2005a. Contradicted and initially stronger effects in highly cited clinical research. *J. Amer. Medical Assoc.* **294**(2) 218–228.

Ioannidis, J. P. A. 2005b. Why most published research findings are false. *PLoS Med.* **2**(8) e124.

Jeffreys, W. H., J. O. Berger. 1992. Ockham's razor and Bayesian analysis. *Amer. Sci.* **80**(1) 64–72.

John, I. D. 1992. Statistics as rhetoric in psychology. *Australian Psych.* **27**(3) 144–149.

Kendall, P. C. 1997. Editorial. *J. Consulting Clinical Psych.* **65**(1) 3–5.

Keselman, H. J., R. R. Wilcox, L. M. Lix, J. Algina, K. Fradette. 2007. Adaptive robust estimation and testing. *British J. Math. Statist. Psych.* **60**(2) 267–293.

Kline, R. B. 2004. What's wrong with statistical tests—And where we go from here. R. Kline, ed. *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research*, Chapter 3. APA Books, Washington, DC, 61–91.

Langman, M. J. S. 1986. Towards estimation and confidence intervals. *British Med. J.* **292**(6522) 716.

Levinthal, D. A. 1991. Random walks and organizational mortality. *Admin. Sci. Quart.* **36**(3) 397–420.

Lykken, D. T. 1968. Statistical significance in psychological research. *Psych. Bull.* **70**(3, Part 1) 151–159.

Makridakis, S., A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen, R. L. Winkler. 1982. The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *J. Forecasting* **1** 111–153.

Mayo, D. 2006. Philosophy of statistics. S. Sarkar, J. Pfeifer, eds. *The Philosophy of Science: An Encyclopedia*. Routledge, London, 802–815.

Meehl, P. E. 1954. *Clinical versus Statistical Prediction: A Theoretical Analysis and Review of the Evidence*. University of Minnesota Press, Minneapolis.

Meehl, P. E. 2002. Interview by Fiona Fidler. August 27. University of Minnesota, Minneapolis.

Mezias, J. M., W. H. Starbuck. 2003. Studying the accuracy of managers' perceptions: A research odyssey. *British J. Management* **14**(1) 3–17.

Oakes, M. 1986. *Statistical Inference. A Commentary for the Social and Behavioural Sciences*. Wiley, Chichester, UK.

Pant, P. N., W. H. Starbuck. 1990. Innocents in the forest: Forecasting and research methods. *J. Management* **16**(2) 433–460.

Popper, K. R. 1959. *The Logic of Scientific Discovery*. Basic Books, New York.

Powell, T. C. 2003. Varieties of competitive parity. *Strategic Management J.* **24**(1) 61–86.

Rosenberg, B., M. Houglet. 1974. Error rates in CRSP and Compustat data bases and their implications. *J. Finance* **29**(4) 1303–1310.

Rosnow, R., R. Rosenthal. 1989. Statistical procedures and the justification of knowledge in psychological science. *Amer. Psych.* **44**(10) 1276–1284.

Rousseau, D. M., J. Manning, D. Denyer. 2008. Evidence in management and organizational science: Assembling the field's full weight of scientific knowledge through syntheses. *Acad. Management Ann.* **2**(1) 475–515.

Rousseeuw, P. J., A. M. Leroy. 1987. *Robust Regression and Outlier Detection*. Wiley, New York.

Rothman, K. J. 1998. Writing for epidemiology. *Epidemiology* **9**(3) 333–337.

Salancik, G. R., J. R. Meindl. 1984. Corporate attributions as strategic illusions of management control. *Admin. Sci. Quart.* **29**(12) 238–254.

San Miguel, J. G. 1977. The reliability of R&D data in Compustat and 10-K Reports. *Accounting Rev.* **52**(3) 638–641.

Schmidt, F. L. 1971. The relative efficiency of regression and simple unit predictor weights in applied differential psychology. *Ed. Psych. Measurement* **31**(3) 699–714.

Schmidt, F. L. 1996. Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psych. Methods* **1**(2) 115–129.

Schmidt, F. L., J. E. Hunter. 1997. Eight common but false objections to the discontinuation of significance testing in analysis of research data. L. Harlow, S. Mulaik, J. Steiger, eds. *What If There Were No Significance Tests?* Lawrence Erlbaum Associates, Mahwah, NJ, 37–63.

Schwab, A., W. H. Starbuck. 2009. Null-hypothesis significance tests in behavioral and management research: We can do better. D. Bergh, D. Ketchen, eds. *Research Methodology in Strategy and Management*, Vol. 5. Elsevier, New York, 29–54.

Seth, A., K. D. Carlson, D. E. Hatfield, H. W. Lan. 2009. So what? Beyond statistical significance to substantive significance in strategy research. D. Bergh, D. Ketchen, eds. *Research Methodology in Strategy and Management*, Vol. 5. Elsevier, New York, 3–28.

Shah, N. R., J. Borenstein, R. W. Dubois. 2005. Postmenopausal hormone therapy and breast cancer: A systematic review and meta-analysis. *Menopause* **12**(6) 668–678.

Shrout, P. E. 1997. Should significance tests be banned? *Psych. Sci.* **8**(1) 1–2.

Smithson, M. 2001. Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Ed. Psych. Measurement* **61**(4) 605–632.

Soofi, E. S., P. C. Nystrom, M. Yasai-Ardekani. 2009. Executives' perceived environmental uncertainty shortly after 9/11. *Comput. Statist. Data Anal.* **53**(9) 3502–3515.

Starbuck, W. H. 1994. On behalf of naïveté. J. A. C. Baum, J. V. Singh, eds. *Evolutionary Dynamics of Organizations*. Oxford University Press, New York, 205–220.

Starbuck, W. H. 2006. *The Production of Knowledge: The Challenge of Social Science Research*. Oxford University Press, Oxford, UK.

Steiger, J. H., R. T. Fouladi. 1992. R2—A computer-program for interval estimation, power calculations, sample-size estimation, and hypothesis testing in multiple regression. *Behav. Res. Methods, Instruments, Comput.* **24**(4) 581–582.

Task Force on Statistical Significance. 1996. Initial report. Board of Scientific Affairs, American Psychological Association, Washington, DC.

Thompson, B. 1995. Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Ed. Psych. Measurement* **55**(4) 525–534.

Thompson, B. 1999a. Why "encouraging" effect size reporting is not working: The etiology of researcher resistance to changing practices. *J. Psych.* **133**(2) 133–140.

Thompson, B. 1999b. Journal editorial policies regarding statistical significance tests: Heat is to fire as p is to importance. *Ed. Psych. Rev.* **11**(2) 157–169.

Thompson, B. 2002. What future quantitative social science research could look like: Confidence intervals for effect sizes. *Ed. Res.* **31**(3) 25–32.

Thompson, B. 2006. *Foundations of Behavioral Statistics: An Insight-Based Approach*. Guilford, New York.

Tukey, J. W. 1991. The philosophy of multiple comparisons. *Statist. Sci.* **6**(1) 100–116.

Vacha-Haase, T., J. E. Nilsson, D. R. Reetz, T. S. Lance, B. Thompson. 2000. Reporting practices and APA editorial policies regarding statistical significance and effect size. *Theory Psych.* **10**(3) 413–425.

Wacholder, S., S. Chanock, M. Garcia-Closas, L. El ghormli, N. Rothman. 2004. Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. *J. Natl. Cancer Inst.* **96**(6) 434–442.

Webster, E. J., W. H. Starbuck. 1988. Theory building in industrial and organizational psychology. C. L. Cooper, I. Robertson, eds. *International Review of Industrial and Organizational Psychology 1988*. Wiley, London, 93–138.

Wilcox, R. R. 1997. *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press, San Diego.

Wilk, M. B., R. Gnanades. 1968. Probability plotting methods for analysis of data. *Biometrika* **55**(1) 1–17.

Wilkinson, L. 1999. APA Task Force on Statistical Inference. Statistical methods in psychology journals: Guidelines and explanations. *Amer. Psych.* **54**(8) 594–604.

Andreas Schwab is an assistant professor in the College of Business at Iowa State University. His research focuses on learning opportunities, learning challenges, and their contingencies during the formation and execution of innovative project ventures. He serves on the editorial board of the *Strategic Entrepreneurship Journal.* He received his Ph.D. in strategic management and organization theory from the University of Wisconsin–Madison.

Eric Abrahamson is the Hughie E. Mills Professor of Business at the Graduate School of Business at Columbia University and Faculty Leader of the Sanford C. Bernstein & Co. Center for Leadership and Ethics. His award-winning research focuses on management fashions in business techniques, bandwagons, interorganizational culture, and disorganized systems. His first book, *Change Without Pain,* was awarded one of the top business books of 2004; *A Perfect Mess* has been translated into 23 languages.

William H. Starbuck is a professor-in-residence at the University of Oregon and professor emeritus at New York University. He has published numerous articles, authored two books, and edited 17 books. His latest book, *The Production of Knowledge*, reflects on lessons from his own academic journey and on the challenges associated with management and social science research.

Fiona Fidler is an Australian Research Council Postdoctoral Fellow in the School of Psychological Science at La Trobe University, Melbourne, Australia, where she works on various projects related to statistical cognition. Her Ph.D. in the history and philosophy of science at the University of Melbourne examined criticisms of null hypothesis significance testing and calls for statistical reform in psychology, medicine, and ecology.