Comment

Contents

Schmidt & Hunter on Krueger ..........65
Brand on Krueger .......................66
Guenther on Krueger .................67
Markus on Krueger ...................68
Hofmann on Krueger ................69
Krueger replies ........................70

DOI: 10.1037//0003-066X.57.1.65

Are There Benefits From NHST?

Frank Schmidt
University of Iowa

John Hunter
Michigan State University

Krueger’s (January 2001) thesis can be summarized as follows: Induction is central to science (true). From a philosophical perspective, induction cannot be defended logically but can be defended pragmatically—because it leads to progress in science (true). Null hypothesis significance testing (NHST) is a good tool—perhaps the essential tool—for induction and inference from research data (false). NHST is like induction: It cannot be defended logically (true) but can be defended pragmatically (false). The pragmatic defense Krueger offered is the contention that NHST has “proven useful” (p. 24) and that it “rewards the pragmatic scientist” (p. 23; both false).

Induction as used by Krueger (2001) is the same as hypothesis testing; it is undisputed that hypothesis testing is indispensable for science. Krueger’s position, then, reduces to the proposition that NHST is the best procedure—and perhaps the essential procedure—for testing hypotheses. This false argument has long been offered as a defense of significance testing (Schmidt & Hunter, 1997, pp. 42–44). In its strong form, the argument is that without significance testing, psychologists could not have a science because they would no longer be able to test hypotheses.

The physical sciences, such as physics and chemistry, do not use NHST or statistical significance tests of any kind, yet these sciences test hypotheses and have done so for centuries. In fact, in contrast to many psychologists, most researchers in the physical sciences regard reliance on significance testing as unscientific (Schmidt, 1996). If the argument is that hypothesis testing requires the use of significance tests, then the logical implications are that physicists and chemists are not really testing hypotheses and that their research is not really scientific. How plausible is this? If the argument is that significance testing is the best method of testing hypotheses in science, then the logical implication is that the hypothesis-testing methods used in physics and chemistry are suboptimal and inferior to those based on NHST and typically used in psychology. Does anyone really believe this? Hence, although induction (hypothesis testing) is central to science, neither NHST nor any other form of statistical significance testing is required for hypothesis testing.

Significance testing almost invariably retards the search for knowledge by producing false conclusions about research literature. The evidence is strong that the null hypothesis is almost always false in psychological research. For example, Lipsey and Wilson (1993) examined 302 meta-analyses of psychological interventions of all kinds in many areas of psychology. In only 2 of these meta-analyses were the effect sizes (ESs) zero or near zero (less than 1%). An examination of all published meta-analyses would produce a similar figure for psychology as a whole. If the null hypothesis is typically false, then Type I error is not important because it is impossible to make a Type I error when the null is false. What is important is Type II error: failing to detect the effect or relation that is there. One minus the Type II error rate is the statistical power of the study: the probability of detecting the effect or relation. The evidence is clear that the average level of statistical power in psychological research is between .40 and .60 (e.g., see Cohen, 1962, 1994; Sedlmeier & Gigerenzer, 1989). The operational decision rule used by researchers is “if it is significant, it is real; if it is not significant, it is zero” (Schmidt, 1996).

Hence, the error rate in the typical psychological research literature is approximately 50%—that is, half of all studies reach false conclusions about the null hypothesis, a situation of maximal apparent conflict in the literature. As discussed by Schmidt (1996), this leads to one of two false conclusions about the meaning of the literature. The first is that the literature is so conflicting that nothing can be concluded. The second is that there are interactions or moderator variables that cause the effect to exist in some studies and to be nonexistent in others and that research should be directed at finding these moderator variables. Meta-analysis typically indicates that both of these conclusions are false—by revealing that the effect exists in all studies (Schmidt, 1996).

Significance tests are a disastrous method for testing hypotheses, but a better method does exist: use of point estimates (ESs) and confidence intervals (CIs). First, unlike significance tests, CIs hold the real error rate to .05 (or whatever confidence level is set); there is no possibility of a higher error rate as with significance tests. In particular, the true error rate will never be 50% when the researcher thinks it is 5% (because the alpha level is set at .05). Second, almost all of the CIs from different studies overlap each other, correctly suggesting that the studies are not contradictory. Third, the CI clearly reveals the level of uncertainty in the study results; unlike the significance test, the CI provides an index of the effects of sampling error on the results. Finally, the ES provides the information needed for subsequent meta-analyses, whereas the significance test does not.

Krueger (2001) stated, “In daily research activities, NHST has proven useful. Researchers make decisions concerning the validity of hypotheses, and although their decisions sometimes disagree, they are not random or arbitrary” (p. 24). First, Krueger presented no evidence to support his assertion of NHST’s usefulness. As shown above, significance testing creates confusion and false conclusions about research literature. How is this useful? Second, researchers’ conclusions disagree more often than sometimes—in the typical research literature, they disagree 50% of
the time. Third, making decisions about hypotheses on the basis of studies with a power of .50—which is what typically happens—is at least irrational, if not arbitrary. Fourth, when Krueger stated that researchers make decisions about the validity of hypotheses, he meant that they do so for their individual studies—something that cannot be justified given the low power of the typical study. In fact, Krueger criticized CIs because “they steer researchers away from drawing categorical conclusions about hypotheses (which is what they are trained to aim for . . .)” (p. 24). It should be clear that it is a deficiency of significance tests that they encourage such conclusions. No single study is ever sufficient to support a conclusion about the validity of a hypothesis; this is especially true of typical studies with their statistical power of about .50. Such conclusions should be based on multiple studies as processed through meta-analysis methods (Hunter & Schmidt, 1990).

Krueger (2001) also stated, “Second, NHST rewards the pragmatic scientist. Much has been learned from this method in the past and presumably more can be learned in the future” (p. 23). Krueger presented no evidence or examples to support his assertion that NHST promotes discovery of knowledge; he just asserted it. Where is the reward to the pragmatic statistician in a procedure that obfuscates, elides, or renders knowledge, or ture? A more accurate conclusion would be that some things have been learned in some areas despite NHST.

Finally, Krueger (2001) contended that significance tests help researchers deal with sampling error in their data. Empirical research clearly requires some mechanism for induction. NHST, especially when bolstered by Bayesian assumptions, fares quite well relative to its alternatives. Effect sizes alone do not weed out findings with large sampling errors (Frick, 1996). Significance level help do this; they indicate how hard a researcher has worked to reduce error to at least make a judgment about the direction of the effect. (p. 24)

However, effect sizes are not used alone; they are used with CIs, and CIs present much more information about the effects of sampling error in data than do significance tests. Even Krueger admitted that “confidence intervals are attractive because they contain more information than p values do” (p. 24). In this statement, Krueger is endorsing Frick’s conclusion that significance tests work quite well if one’s goal is merely to determine the direction of an effect and one is not interested in its magnitude. Sedlmeier (1999, pp. 206–215) showed that this conclusion is false. Even in this circumstance, point estimates and CIs work better than significance tests.

REFERENCES


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. 38–64).

Mahwah, NJ: Erlbaum.


DOI: 10.1037//0003-066X.57.1.66

Why Chance Is a Good Theory

Jay L. Brand
Haworth, Inc.

I adhere to the view that assumptions (*values* seems to be the prevailing applied term, although *axioms, givens, presuppositions, or a priori postulates* would all work as well) are rarely as astutely examined by scientists—and in particular by psychologists—as are the details of their prevailing methods. The trees, it would seem, stand up for endless scrutiny while the forest escapes without much notice.

Although Krueger (January 2001) provided another helpful reminder of the logical difficulties of null hypothesis significance testing (NHST), as well as an intriguing exploration of its pragmatic advantages, he also carelessly included a dubious claim that weakened at least some of his contentions. Although I agree with many of the arguments advanced by rightly celebrated critics of NHST, I do not agree that its widespread use represents a ubiquitous ignorance of its logical pitfalls, as these critics invariably seem to imply. I am encouraged by Krueger’s admitting that there is a larger context in which most scientific psychologists use NHST.

I have spoken many times about NHST with colleagues in, for example, mathematics departments, and these colleagues frequently reiterate many of Krueger’s (2001) concerns. They insist, for example, that establishing a priori rejection regions for $H_0$ and/or $H_1$ is completely arbitrary from a purely logical point of view, as indeed is what one calls $H_0$ and $H_1$. However, when they are embedded within the hypothetico-deductive approach to testing causal theories, the arbitrary logical symmetry between $H_0$ and $H_1$ breaks down (cf. Dawes, 1991). In that instance, $H_1$ ideally represents a predictive failure for the theory being tested, whereas $H_0$ (and any more extreme, i.e., improbable, event) represents a predictive success for the theory. Evaluating NHST purely within its mathematical, logical, or causal framework, misses its special power as a tool for testing causal theories. And contrary to Krueger’s (2001) claims, within this larger causal framework, the null hypothesis—or, conceptually, chance—remains the best theory available absent compelling evidence to the contrary. Together with the principle of parsimony (another assumption beyond the purely logical confines of NHST), chance and its statistical instantiation, the null hypothesis, remain the best guesses about any phenomenon until a better theoretical alternative—one made up of empirically testable conceptual linkages—can be developed. I would thus doubt that winning at bingo or other games of chance rests on the ability to communicate with dead relatives; in this case, $H_0$—the hypothesis that chance alone has produced good performance—should not be lightly discarded. (I suppose this would be a *risky experiment* in Krueger’s, 2001, terminology, at least if one’s interest involves evidence for life after death.) Many researchers simply recognize and accept that science uses logic but not vice versa. Science represents a protracted attempt to contribute to a public edifice of knowledge founded on probabilistic evidence that the piecemeal construction achieves some important similarities with reality. No one sincerely believes that his or her single experiment will answer any useful question once and for all.