Joachim Krueger (1999) Significance Testing Does not Solve the Problem of Induction. Psycoloquy: 10 (015) Social Bias (16)

Volume: 10 (next, prev) **Issue**: 015 (next, prev) **Article:** 16 (next prev first) **Alternate versions:** ASCII Summary

Topic: Social bias View Topic

Article: 16) Krueger 10(015) Significance Testing Does not Solve the View Article

PSYCOLOQUY (ISSN 1055-0143) is sponsored by the American Psychological Association (APA).

SIGNIFICANCE TESTING DOES NOT SOLVE THE PROBLEM OF INDUCTION
Reply to Chow on Krueger on Social-Bias

Joachim Krueger
Department of Psychology
Brown University, Box 1853
Providence, RI 02912
http://www.brown.edu/Departments/Psychology/faculty/krueger.hml

Joachim_Krueger@Brown.edu

Abstract

Chow (1999) presents an "if only" defense of Null Hypothesis Significance Testing (NHST). If investigators only recognized the distinctions between (1) theory corroboration experiments and utilitarian experiments, and between (2) substantive hypotheses and statistical hypotheses, then NHST could take its rightful place in empirical psychology. By contrast, I suggest that these distinctions divert attention away from the fundamental problems of NHST, namely, that (1) point-specific hypotheses (null or other) cannot be verified, and that (2) increases in statistical power favor any non-null hypotheses and hence the substantive claims associated with them.

Keywords

Bayes' rule, bias, hypothesis testing, individual differences probability, rationality, significance testing, social cognition, statistical inference

1. I laid much of the blame for the current state of bias research in social cognition on ritualistic and bi-directional significance testing (Krueger 1998). With sufficient persistence on the part of the experimenters, sufficient power on the part of the study sample, and sufficient precision on the part of

the measurement instruments, respondents can be shown to over- or underestimate the degree of their own typicality within a group, their own qualities relative to the qualities of others, and the relative impact of dispositional and situational causes on behavior. The point of no bias (rationality) is identified with the null (or nil) hypothesis. Bias lies on either side of that point, and it is therefore always detectable. Rationality is not detectable but only 'retainable' so long as bias has not been detected.

2. In defense of NHST, Chow (1999, see also 1998) draws a distinction between theory corroboration experiments and utilitarian experiments. In theory corroboration experiments, effect sizes do not matter; what matters is whether the null hypothesis, H0, is rejected. Chow describes a hypothetical memory experiment designed to test the idea that subsequent learning interferes with previous learning. According to Chow, the size of the learning decrement leading to the rejection of H0 is irrelevant. In contrast, he acknowledges the relevance of effect sizes in utilitarian experiments. More fertilizer, or fertilizer of a different kind, for example, may increase crop yield, and one would wonder by how much. I am skeptical about the relevance of the distinction between theory corroboration experiments and utilitarian experiments because this distinction does not affect the way NHST is done. As Chow himself observes, H1 (i.e., the effect size) "plays no role in the statistical decision." Nevertheless, and although it may not make any difference, I would say that the typical study of social-cognitive bias is meant to corroborate a theory.

3. Next, Chow draws a distinction between substantive and statistical hypotheses. Substantive hypotheses are posed on the conceptual or theoretical level, whereas statistical hypotheses are posed on the numerical data level. Chow suggests that the corroboration of substantive hypotheses "must be different from the statistical distinction itself." What is the nature of this difference and what is its relevance for NHST? Chow proposes that deductive syllogisms can bridge the gap between the theoretical and the statistical levels. In brief, when H0 is rejected, H1 is accepted by a disjunctive syllogism. Then, when H1 is accepted, the substantive theory is corroborated (i.e., accepted). This proposal amounts to an attempt to justify of NHST by logical inference.

4. The "embedding" of NHST within a set of syllogisms creates the appearance of logical rigor, but it does not overcome the probabilistic nature of the inferences drawn from data. The output of NHST is a probability, namely, the probability of the observed data (or data more extreme) given that H0 is true (i.e., p(D|H0)). No matter how small Fisher's p is, no certainty about the falsity of H0 can come from it. The convention to 'reject' H0 if $p < .05$ cannot serve as a premise for any logically valid conclusion. While it may be true that in a given experiment the probability of the data under H0 is low, it does not follow that H0 is false. Hence it does not follow, by disjunctive syllogism, that H1 is true. In the same way, it does not follow that H0 is true (or that H1 is false) if the probability of the data under H0 is not quite so low (i.e., if $p > .05$).

5. Demonstrations of social bias are rarely claims that bias has been proven; instead, they are claims that bias has been detected at conventional levels of improbability under H0. These claims benefit from the asymmetry built into NHST. Increases in statistical power (or precision) make rejections of H0 (and thus rationality) more likely. Chow is unconcerned about this asymmetry and suggests that H0 can have a fair chance of being retained (and even proven) when theory corroborative experiments are designed properly. By proper design he means that the implementation of control procedures (e.g., random assignment to conditions) may limit response variability to chance variation. In other words, he suggests that, first, H0 can be true, and that, second, this truth can be verified.

6. In response to the first claim, I note that the falsity of H0 is a matter of mathematical calculus. Like any other point-specific hypothesis, H0 is always false, and this state of affairs does not depend on the nature of the observed data (Bakan, 1966; Lykken, 1968; Meehl, 1978). When a variable is continuous rather than discrete, any individual point on its distribution has a value on a density function, but it has

no probability. Only areas under the curve have probabilities. Reflecting this, NHST yields the probability of the tail areas of the distribution (i.e., the probability of extreme data under H0). NHST cannot yield estimates for the probability of H0 being true. When it is not even possible to estimate the probability of H0 being true, it is certainly impossible to prove H0 to be true. It is remarkable that the primary outcome of NHST is a conditional probability (namely, p(D|H0)) whose condition has no marginal probability value itself. Yet, claims about the possible truth of H0 continue to surface. These claims are rhetorical, however, rather than logical.

7. Lewandowsky and Mayberry (1998), responding to Chow's (1998) target article in BEHAVIORAL AND BRAIN SCIENCES < http://www.cogsci.soton.ac.uk/bbs/Archive/ >, also expressed skepticism concerning the idea that H0 is always false. To them, this would amount to an a priori acceptance of the idea that "that knocking on wood will prevent the occurrence of dreaded events, that black cats crossing the road are better predictors of future mishaps than white cats, or [. . .] any other superstition that can be put to an experimental test with sufficiently large effect sizes." This argument is rhetorical as it appeals to the conviction of the enlightened reader that superstitions are categorically false. It does not, however, make the desired methodological claim. As stated, each of the three superstitions has a 50% chance of being supported by large-sample studies. The correlation between knocking on wood and the occurrence of dreaded events is presumably near zero, but it cannot be exactly zero. Indeed, knocking on wood might have a small positive association with the occurrence of dreaded events. This might be the case, for example, when people have valid fears of calamities they cannot control (by knocking or whatever). The superstition would only be true (though meaningless) a priori if it stated that dreaded occurrences are either more likely or less likely after knocking on wood than after doing nothing.

8. Elaborating Chow's second claim, namely, that H0 can be verified, Lewandowsky and Mayberry suggested "an experiment with 25,000 subjects may fail to reject H0." Such large-scale experiments may produce what one might call moral certainty about the truth of H0. They remain logically inconclusive, however, because they cannot solve Hume's problem of induction. No matter how many white swans you observe (while not observing any non-white swans), you cannot reach the categorical conclusion that all swans are white. Large-scale experiments that fail to reject H0 beg the question of whether the obtained differences -- however small they might be -- would become significant if the sample size were doubled or tripled. Consider Karl Pearson's famous attempt to test the fairness of a coin (see Lockhart 1998, p. 165). After flipping it 24,000 times, he found that it came up heads 12,012 times. Was the coin fair? NHST suggests that is was because p(D|H0) = .88. The same small bias would entail the rejection of H0, however, if the coin were tossed 4 million times. This small bias may well be trivial in most contexts, and its statistical detection would require more time and effort than most coin flippers are willing to invest. Still, any sample size, no matter how large it is, can always be multiplied. But, the argument goes, if the sample size were multiplied, the difference between the expected value under H0 and the obtained value would probably approach zero. This would only be so if H0 were exactly true, which is a condition that is never really satisfied. The truth of H0 is a fantasy. H0 "can only be true in the bowels of a computer processor running a Monte Carlo study (and even then a stray electron may make it false). If it is false, even to a tiny degree, it must be the case that a large enough sample will produce a significant result and lead to its rejection. So if the null hypothesis is always false, what's the big deal about rejecting it?" (Cohen, 1990, p. 1308).

9. Theory corroboration experiments, as described by Chow, involve inferences about populations of unspecified (i.e., infinite) size. The situation would be different if the observed events were discrete and populations were finite. Here, H0 could be verified by counting up all events. Suppose we know that an urn contains 50 red balls and 50 blue balls because we have counted them. Given this knowledge about the truth of H0, we can predict the probabilities of all possible samples taken from this urn. Notice that this scenario eliminates the purpose of NHST, however, which is the drawing of inductive inferences from samples to unknown population parameters.

10. Aside from the mathematical objections against the truth of H0 in infinite populations, there are practical concerns. Casino operators know that gambling machines, which are designed to ensure the truth of H0, cannot meet this exacting goal. Even small biases in the apparatus can become costly in the long run. To avoid patient and determined gamblers figuring out which numbers come up slightly more often than others at the roulette wheel, casino operators periodically exchange wheels so that biases become diluted in the long run. This strategy does not eliminate biases, but it makes them so remote that their detection is beyond the capability of even the hardiest of gamblers. In other words, casino operators do not blindly rely on the classical theory of probability. According to that theory, all possible events (e.g., the six faces of a die), can be equally likely. But the unempirical character of this view has been criticized. "We can never be really sure that the possibilities are all equally likely -- a die can be loaded, a coin can be off balance, and some cards in a deck can stick together [and experimental controls may be imperfect, and] to forestall this kind of criticism, we often hedge our probability statements by inserting the word 'if'" (Freund, 1993, p. 40). The problem of induction remains unsolved because this hedge cannot be evaluated by the very method (i.e., NHST) that relies on it for inductive inferences.

11. Chow tries to justify induction by superimposing 'embedding' syllogisms on the standard practice of NHST. None of these syllogisms is logically valid (Erwin, 1998), however, and even if they were, they would add no information to that which is yielded by NHST itself. Hence, the attempt to rescue inductive inferences by grafting deductive inferences onto them disguises rather than solves the problem of induction.

REFERENCES

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

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

Chow, S. L. (1998). Multiple Book Review of "Statistical Significance: Rationale, Validity and Utility." Behavioral and Brain Sciences 21: 169-240.
ftp://ftp.princeton.edu/pub/harnad/BBS/WWW/bbs.chow.html

Chow, S. L. (1999). In defence of significance tests. Commentary on Krueger on social-bias. PSYCOLOQUY 10(006). ftp://ftp.princeton.edu/pub/harnad/Psycoloquy/1999.volume.10/ psyc.99.10.0006.social-bias.15.chow http://www.cogsci.soton.ac.uk/cgi/psyc/newpsy?10.006

Erwin, E. (1998). The logic of null hypothesis testing. Behavioral and Brain Sciences, 21: 197-198.

Freund, J. E. (1993). Introduction to probability. New York: Dover Publications.

Krueger, J. (1998). The bet on bias: A foregone conclusion? PSYCOLOQUY 9(46). ftp://ftp.princeton.edu/pub/harnad/Psycoloquy/1998.volume.9/ psyc.98.9.46.social-bias.1.krueger http://www.cogsci.soton.ac.uk/cgi/psyc/newpsy?9.46

Lewandowsky, S. & Mayberry, M. (1998). The critics rebutted: A Pyrrhic victory. Behavioral and Brain Sciences 21: 210-211.

Lockhart, R. S. (1998). Introduction to statistics and data analysis for th behavioral sciences. New York: Freeman.

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

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology 46: 806-834.

Volume: 10 (next, prev) **Issue**: 015 (next, prev) **Article:** 16 (next prev first) **Alternate versions:** ASCII Summary

Topic: Social bias View Topic

Article: 16) Krueger 10(015) Significance Testing Does not Solve the View Article