BIASES TO THE LEFT, FALLACIES TO THE RIGHT: STUCK IN THE MIDDLE WITH NULL HYPOTHESIS SIGNIFICANCE TESTING
Commentary on Krueger on Social-Bias

Ralph Hertwig & Peter M. Todd
Center for Adaptive Behavior and Cognition
Max Planck Institute for Human Development
Lentzeallee 94
14195 Berlin, Germany
http://www.mpib-berlin.mpg.de/ABC/Staff/hertwig
http://www.mpib-berlin.mpg.de/users/ptodd

hertwig@mpib-berlin.mpg.de ptodd@mpib-berlin.mpg.de

Abstract

Krueger (1998) argues that biases are simple to find because of the ease of disproving overly specific null hypotheses of normative behavior. We support his argument with examples of biases falling on both sides of the normative null hypothesis in behavioral decision making, and we show how a broader definition of normativity can lead to opposite conclusions about human rationality. To overcome the problems of asymmetric hypothesis testing -- encouraging the discovery of biases and discouraging the specification of particular cognitive processes -- we recommend two steps: submit precisely specified theories to symmetrical tests, and model sound reasoning based on problem-specific psychological assumptions.

Keywords

biases, heuristics, narrow norms, Bayesian reasoning, autocorrelation, symmetric hypothesis testing

1. In a recent presidential address to the members of the Society for Judgment and Decision Making, Barbara Mellers (1996) observed that five decades of research on judgment and decision making have created a "lopsided view of human competence" (p. 3) in which the "demonstrations of one more error for the sake of an error...are nothing new" (p. 3). Why are we so preoccupied with reasoning errors, "biases," "fallacies," and "cognitive illusions"?

2. Krueger (1998) has provided us with a very insightful answer to this question: Finding reasoning errors is just plain simple. It is simple because of the asymmetric way research hypotheses are tested in our field. Psychology has institutionalized the habit of not specifying the predictions of any particular research hypothesis of interest (e.g., a specific heuristic), but instead specifying the predictions of a less interesting hypothesis, namely, the null hypothesis. The policy then is to try to reject this null hypothesis and thereby claim credit for the unspecified and untested research hypothesis.

3. In much of the research on people's reasoning competence, the null hypothesis typically represents the numerically precise point-specific prediction of some normative principle from probability theory or SEU theory. The predicament is that any difference between such a theoretically predicted single value and the empirical value observed in the experiment can be made significant, if the sample size n is made large enough. Thus, psychology's way of testing hypotheses stacks the deck in favor of rejecting predictions derived from normative principles (i.e., the null hypothesis). In order words: Asymmetric hypothesis testing has a built-in bias to find biases.

I. THE SCOPE OF THE PREDICAMENT IN BEHAVIORAL DECISION

MAKING

4. Clearly, this dilemma is widespread. Krueger has focused on its manifestation in classic effects from social cognition research. We can add further illustrative examples from research on behavioral decision-making, demonstrating that there, too, any significant deviation from the null hypothesis can be and usually has been classified as a reasoning error. Take Bayesian reasoning research as an example. This research has mainly fallen into two traditions that focus on opposite deviations from the normative response predicted by Bayes' theorem. In the 1960s, the conclusion was that people mistrust new data and give too much weight to the prior probabilities of events in a given situation -- a phenomenon called "conservatism." In the words of Ward Edwards (1968, p. 18): "It takes anywhere from two to five observations to do one observation's worth of work in inducing a subject to change his opinions." In contrast, in the 1970s and 1980s, researchers of the "heuristics-and-biases" program arrived at the opposite conclusion: People systematically neglect prior probabilities and give too much weight to new data -- a phenomenon called "base-rate neglect." In the words of Kahneman and Tversky (1972, p. 450): "... man is apparently not a conservative Bayesian: he is not a Bayesian at all." In both cases, deviations from the single "correct" Bayesian updating of new and old knowledge were seen as indications of erroneous reasoning.

5. Research on the perception of randomness provides another example for how assumed irrationality lurks to either side of the null hypothesis. Truly random sequences are usually assumed to show no autocorrelations (Wagenaar, 1972). People's judgments of randomness have been reported to deviate from this norm in both directions. Wrongly predicting negative autocorrelations -- believing that successive values should change appreciably from previous trends -- has been identified as the "gambler's fallacy" (Tversky & Kahneman, 1974). On the other side, wrongly seeing positive autocorrelations -- believing that past trends will continue -- has been called the "hot hands" effect (Gilovich, Vallone, & Tversky, 1985). Interestingly, both beliefs, despite pointing in opposite directions from the "true" state of affairs, have been explained by the same relatively ill-specified hypothesis, the representativeness heuristic (see Ayton & Fischer, 2000).

6. Krueger's account provides us with an explanation for why previous lines of research like these seem doomed to end up in biases to the left and fallacies to the right of the normative prediction. But his account may also help us understand when and why some researchers come to opposite conclusions about whether people are even rational at all, namely, when one team defines normative reasoning as a single value and another team adopts a more encompassing definition region. For instance, consider research on the multiplication rule for dependent and independent events (for a short history of this research see Hertwig & Chase, 1998). Cohen and colleagues focused on this and other "Achilles heels" (Cohen, Chesnick, & Haran, 1972, p. 46) of probability judgment in teenagers and adults. Cohen and Hansel (1958) had participants estimate conjoint probabilities, for example, the probability of winning two gambles together when there is a 10% chance of winning each one independently. Participants' responses were compared for equality with the result calculated from the multiplication rule for independent events ($p(A\&B) = p(A) p(B)$; in this case, .01). Finding that most 12- to 15-year-olds (Cohen et al., 1972; Cohen & Hansel, 1958) overestimated the conjoint probability, Cohen et al. (1972) concluded that a "grasp of the multiplicative character of a compound probability is far from being in any sense a 'primitive' property of mental processes in relation to the external world" (p. 44).

7. In contrast, Peterson and Beach (1967) argued that the laws of probability theory and statistics could be used to build psychological models that "integrate and account for human performance in a wide range of inferential tasks" (p. 29). They marshaled what they saw as people's consistency with the multiplication rule as evidence for this argument. In two experiments, Peterson, Ulehla, Miller, Bourne, and Stilson (1965) asked participants to estimate conditional and unconditional probabilities such as the number of people out of 100 who are witty and the number of those witty people who are brave. According to probability theory, the product of these two values, $p(witty) p(brave \mid witty)$, is equal to the product of another pair of values they asked for, $p(brave) p(witty \mid brave)$. Unlike Cohen and colleagues, however, Peterson et al. did not define numerical equality to be the decisive criterion; rather they analyzed the correlation between pairs of products like these. Finding significant correlations (.67 and .90 in Experiments 1 and 2, respectively), Peterson et al. (1965) concluded that participants' probability judgments showed "a high degree of internal consistency" (p. 528) in accordance with the multiplication rule.

8. In short: Adopting any particular test for adherence to normative principles is likely to affect our results and conclusions -- in particular, as Krueger's examples and our own demonstrate, narrow definitions of normativity will lead to widespread findings of irrationality. In addition, analyzing the way other researchers test their hypotheses may help us to understand why various lines of research sometimes arrive at completely opposing conclusions about people's reasoning abilities.

II. IMPLICATIONS FOR RESEARCH STRATEGIES

9. While we mostly agree with Krueger's suggestions for how to redesign research on human reasoning competence, we would like to highlight two implications that deserve particular attention.

IIA. BEYOND THE MECHANISTIC USE OF NORMS: TOWARD A THOUGHTFUL ANALYSIS OF THE SITUATION

10. The practice of null hypothesis testing described earlier is not the only mechanistic ritual in psychological research. Norms of sound reasoning are also often employed in a mechanistic way. Recall the prevailing research strategy in much of research on human reasoning. Typically, people are presented with word problems designed such that reasoning according to a normative principle (e.g., from probability theory or logic) leads to the "correct" response. Reasoning according to other principles (e.g., using the

representativeness heuristic) in these problems results in a qualitatively different and thus "incorrect" response.

11. For this research strategy, the content of the word problems (e.g., the Linda problem, the cab problem, the engineer-lawyer problem) is irrelevant, because the content-blind normative principles are assumed to apply, irrespective of the particular subject matter of the problems. The function of the content is merely decorative, to deliver the values or pieces of information that are to be mechanically plugged into the normative equation. In contrast to this mechanistic use of norms, any realistic psychological modeling of rational judgment requires making assumptions about how people decide on the numbers (e.g., prior probabilities, likelihoods) that should enter the equations, or on the particular information in a word problem that is relevant for the required judgment.

12. Take Birnbaum's (1983) thoughtful explication of a rational response to the cab problem as an example. This problem involves a cab that is involved in a hit-and-run accident at night. The text provides the information that a witness identified the cab as being blue, along with information about the eyewitness's ability to discriminate blue and green cabs, and the base rate of blue and green cabs in the city. Rather than mechanically plugging these values into Bayes's formula as typically is done in the heuristics-and-biases program, Birnbaum started with the content of the problem and made assumptions about various psychological processes a witness may use. In terms of a signal detection model, for instance, a witness may try to minimize some error function: If witnesses are concerned about being accused of incorrect testimony, then they may adjust their criterion so as to maximize the probability of a correct identification. If instead witnesses are concerned about being accused of other types of errors, then they can adjust their criterion so as to minimize those specific errors. Obviously, different goals will lead to different posterior probabilities (see Gigerenzer, 1998, and Mueser, Cowan & Mueser, 1999, for the detailed arguments).

13. The lesson from Birnbaum's analysis is that "the normative solution to the cab problem requires the assumption of the theory of the witness, whether by the subject or the experimenter" (1983, p. 93). More generally, psychological assumptions are indispensible for constructing sensible norms of good reasoning (for two more examples, see the analysis of the Linda problem by Hertwig & Gigerenzer, 1999, and of strategies people use in Bayesian reasoning tasks and their possible adaptive fit to the environment by Gigerenzer & Hoffrage, 1995, and Goodie & Todd, 2000). Those assumptions determine the choice among possible candidates for a normative model, and different sets of assumption may favor different models. As a consequence, there is often not one single "correct" response that could be equated with the null hypothesis, but, depending on the psychological assumptions, multiple "correct" responses. Given this, how can different models be tested?

IIB. BEYOND THE STRAW MAN "NULL HYPOTHESIS:" TOWARD A TEST OF COMPETING PRECISE THEORIES

14. As we indicated earlier, the prevailing form of hypothesis testing in social and cognitive psychology is asymmetric -- only the predictions of the null hypothesis of "chance" or "rationality" are precisely specified. The predictions of the second hypothesis, typically the research hypothesis (e.g., a heuristic), remain unspecified. Psychology pays a high price for this logic. It discourages or even punishes the formulation of precise theories and predictions, because precise predictions exclude a wider range of empirical observations, and thus are more likely to be false. In contrast, indeterminate theories can much more readily resist and survive attempts to disprove them.

15. In our view, statistical testing ought to be symmetrical -- and various recent theoretical developments in behavioral decision making encourage such testing. We have, for instance, recently proposed the notion of the "adaptive toolbox" (Gigerenzer, Todd, and the ABC Research Group, 1999), that is, the collection of specialized cognitive mechanisms that evolution has built into the human mind for specific domains of inference and reasoning. This notion calls for symmetric hypothesis testing because the adaptive toolbox typically includes multiple heuristics that can be applied in a particular situation, which means that it can provide multiple hypotheses to account for a given behavior. Symmetric testing of hypotheses requires first that the precise predictions of several heuristics be derived for specific experimental situations. Then, each individual participant's judgments are compared with the set of predictions, and if there is a match with one of the predictions (at the level of the hypothesized processes, or the outcomes, or ideally both), their judgments are classified as being consistent with the corresponding heuristic.

16. The result of this testing procedure will not be just a single p-value. Rather, the researcher may find for instance that 60% of the participants used heuristic A, 10% employed heuristics B or C, and the rest appeared to be using idiosyncratic strategies. Consistent with Krueger's suggestion, such a symmetric testing strategy will discourage researchers from averaging across individuals -- a practice which is unjustified when multiple strategies are in use in a population. Instead, it will encourage researchers to analyze inter-individual differences in people's use of heuristics.

17. To conclude: In a debate with Gigerenzer (1996), Kahneman and Tversky (1996) conceded that "it soon became apparent that 'although errors of judgments are but a method by which some cognitive processes are studied, the method has become a significant part of the message'" (p. 582). Gigerenzer, in contrast, has pointed out that he has "always found it difficult to understand why Kahneman and Tversky have persisted in one-word explanations such as representativeness" (1996, p. 594). Krueger's elegant argument helps us to see how the overwhelming focus on reasoning errors and the simultaneous lack of theoretical progress in the heuristics-and-biases program are two sides of the same coin. Asymmetric hypothesis testing reinforces both the discovery of biases and the reluctance to specify underlying cognitive processes. In our view, the crucial steps needed to overcome this state of affairs are to submit precisely specified theories to symmetrical tests, and to model sound reasoning (by considering specific psychological mechanisms underlying people's comprehension of a decision situation) rather than mechanistically imposing normative rules.

REFERENCES

Ayton, P., & Fischer, I. (2000) The gambler's fallacy and the hot-hand fallacy: Two faces of subjective randomness? Unpublished manuscript.

Birnbaum, M. H. (1983) Base rates in Bayesian inference: Signal detection analysis of the cab problem. American Journal of Psychology 96:85-94.

Cohen, J., Chesnick, E. I., & Haran, D. (1972) A confirmation of the inertial-y effect in sequential choice and decision. British Journal of Psychology 63:41-46.

Cohen, J. & Hansel, C. E. M. (1958) The nature of decisions in gambling: Equivalence of single and compound subjective probabilities. Acta Psychologica 13:357-370.

Edwards, W. (1968) Conservatism in human information processing. In B. Kleinmuntz (Ed.), Formal represenation of human judgment. New York: Wiley.

Gilovich, T., Vallone, R. & Tversky, A. (1985) The hot hand in basketball: On the misperception of random sequences. Cognitive psychology 17:295-314.

Gigerenzer, G. (1996) On narrow norms and vague heuristics: A rebuttal to Kahneman and Tversky. Psychological Review 103:592-596.

Gigerenzer, G. (1998) Psychological challenges for normative models. In D. M. Gabbay & Ph. Smets (Eds.), Handbook of defeasible reasoning and uncertainty management systems (Vol 1, pp. 441-467). Dordrecht: Kluwer.

Gigerenzer, G, & Hoffrage, U. (1995) How to improve Bayesian reasoning without instruction: Frequency formats. Psychological Review 102:684-704.

Gigerenzer, G., Todd, P. M., & the ABC Research Group (1999) Simple heuristics that make us smart. New York: Oxford University Press. http://www.cogsci.soton.ac.uk/bbs/Archive/bbs.todd.html

Goodie, A.S., & Todd, P.M. (2000) The ecological rationality of base-rate neglect. Unpublished manuscript.

Hertwig, R. & Chase, V. M. (1998) Many reasons or just one: How response mode affects reasoning in the conjunction problem. Thinking and Reasoning 4:319-352.

Hertwig, R. & Gigerenzer, G (1999) The 'conjunction fallacy' revisited: How intelligent inferences look like reasoning errors. Journal of Behavioral Decision Making 12: 275-305.

Kahneman, D & Tversky, A. (1972) Subjective probability: A judgment of representativeness. Cognitive Psychology 3:430-454.

Kahneman, D., & Tversky, A. (1996) On the reality of cognitive illusions. Psychological Review 103:582-591.

Krueger, J. (1998). The bet on bias: A foregone conclusion? PSYCOLOQUY 9(046). ftp://ftp.princeton.edu/pub/harnad /Psycoloquy.1999.volume.9/ psyc.98.9.46.social-bias.1.krueger http://www.cogsci.soton.ac.uk/cgi/psyc/newpsy?9.046

Mellers, B. (1996) From the president. J/DM Newsletter 15:3.

Mueser, P. R., Cowan, N. & Mueser, K. T. (1999) A generalized signal detection model to predict rational variation in base rate use. Cognition 69: 267-312.

Peterson, C. R., & Beach, L. R. (1967) Man as an intuitive statistician. Psychological Bulletin 68: 29-46.

Peterson, C. R., Ulehla, Z. J., Miller, A. J., Bourne, L. E., & Stilson, D. W. (1965) Internal consistency of subjective probabilities. Journal of Experimental Psychology 70: 526-533.

Tversky, A. & Kahneman, D. (1974) Judgment under uncertainty: Heuristics and biases. Science 185: 1124-1131.

Wagenaar , W. A. (1972) Generation of random sequences by human subjects: A critical survey of the literature. Psychological Bulletin 77: 65-72.

Topic: Social bias   View Topic

Article: 20) Hertwig 11(028) Biases to the Left, Fallacies to the   View Article