round, regression towards the mean produced a mirror pattern that looked like *underconfidence bias:* When participants answered 100% correctly, their mean confidence was lower, such as 80%. They found no real bias. The same unsystematic error is a sufficient condition for two other phenomena listed in Table 1, people's apparent error of overestimating low risks and underestimating high risks (Lichtenstein et al. 1978), as well as the hard–easy effect (see Gigerenzer & Fiedler, 2003; Juslin et al. 2000).

Consider next how stimulus objects are sampled from an environment and a class of phenomena known as "contingency illusions," which were attributed to irrelevant prior beliefs or prejudices against minorities. Versions of the contingency illusion have been claimed in research on self-fulfilling prophecies (Jussim 1991; Kukla 1993), on confirmation biases in hypothesis testing (Snyder 1984), and on alleged memory advantages for negative behaviors in minorities (Hamilton & Gifford 1976; Hamilton & Sherman 1989).

Let me use evaluative judgments of minorities as an illustration. It is an ecological truism that minorities are smaller than majorities, and a recurrent property of social environments is that the rate of positive, norm-conforming behaviors is higher than the rate of negative, norm-violating behaviors (Fiedler 1991; Parducci 1968). When these two ecological assumptions are built into the stimulus distribution presented in a social psychological experiment, participants may be exposed to the following description:

Group *A* (Majority):	18 positive	and	8 negative behaviors
Group *B* (Minority):	9 positive	and	4 negative behaviors

Note that the same ratio of positive to negative behaviors (18:8 = 9:4) holds for both groups, but people nevertheless tend to conclude that there is significantly more positive behavior in the majority than in the minority: a "contingency illusion." Given the unequal sample sizes, however, an unbiased mind using an (unconscious) binomial test would infer that there are significantly more positive than negative behaviors in the majority group ($p = .038$), but not in the minority group ($p = .13$). Thus, unequal sample size is a sufficient condition for a class of phenomena labeled "contingency illusions." Again, one can empirically test whether an additional bias exists because of prior knowledge, such as by replacing real groups by neutral labels, to rule out any influence of prior knowledge (Fiedler et al. 1993; 1999).

Table 1 lists two other phenomena that can be deduced from sampling. One has been called "overconfidence bias," and is defined as mean confidence minus proportion correct (many different phenomena have been labeled overconfidence). Note that "miscalibration" does not imply this phenomenon. It can be logically deduced from unrepresentative sampling of stimulus items and an unbiased mind (Gigerenzer et al. 1991). An analysis of 135 studies showed that "overconfidence bias" practically disappears when stimuli are randomly selected from an environment (Juslin et al. 2000). The second phenomenon is that people erroneously judge that there are more English words with a letter (such as "R") in first position than in third position, which has been attributed to "availability" (Tversky & Kahneman 1973). When one uses a representative sample of letters, rather than the five letters selected by Tversky and Kahneman (which are among the few that are more frequent in the third position), people's apparently systematic bias disappears (Sedlmeier et al. 1998).

The other "cognitive illusions" listed in Table 1 can be deduced in the same way for the task structure, including that of social environments (see also Gigerenzer 2000; 2001; Gigerenzer & Fiedler 2003; Krueger & Mueller 2002). An objection to my general argument is, "But people do commit errors!" No doubt, people commit errors; but I am talking about a blunder committed by a research program. The fact that little attention is paid to establishing what is good and bad reasoning cannot be excused by blaming John Q. Public.

Errors might be a window to cognitive processes, but falsely identified errors do not seem to be so, which is consistent with the fact that after 30 years of collecting errors, no model of cognitive processes has emerged from overconfidence bias, the conjunction fallacy, or any of the other celebrated errors – only vague labels. In contrast, the study of amazing performance seems to be a better window to cognitive processes, such as the less-is-more effect, which led to the discovery of the recognition heuristic (Goldstein & Gigerenzer 2002).

The story is told that there are two personalities among psychologists, optimists and pessimists, who see the glass as half full or half empty, respectively. According to this legend, people like Funder, Krueger, and myself are just kinder and more generous, whereas the pessimists enjoy a darker view of human nature. This story misses what the debate about human irrationality is about. It is not about how much rationality is in the glass, but what good judgment is in the first place. It is about the kinds of questions asked, not just the answers found.

Null hypothesis statistical testing and the balance between positive and negative approaches

Adam S. Goodie

Department of Psychology, University of Georgia, Athens, GA 30602-3013.
goodie@egon.psy.uga.edu
http://www.uga.edu/psychology/faculty/agoodie.html

Abstract: Several of Krueger & Funder's (K&F's) suggestions may promote more balanced social cognition research, but reconsidered null hypothesis statistical testing (NHST) is not one of them. Although NHST has primarily supported negative conclusions, this is simply because most conclusions have been negative. NHST can support positive, negative, and even balanced conclusions. Better NHST practices would benefit psychology, but would not alter the balance between positive and negative approaches.

Balance is hard to achieve, but Krueger & Funder (K&F) are right to try. Several of their concrete suggestions for the better balancing of positive and negative approaches in social cognition are constructive and likely to be beneficial. First, the canonical literature is not likely to change overnight, but a more balanced reading of it would help, and the canon is indeed more balanced than the treatment it has received in recent decades. Second, K&F avoid the temptation to implore research that is more likely to support positive conclusions. Their formulation, which is entirely reasonable, is to encourage research in which normative responding is at least possible. Third, it would be of salutary value if more researchers examined modal as well as mean responding. When most participants respond normatively, it is appropriate to model this behavior. This is not intended as a complete list of their constructive suggestions: there are more. However, one of their recommendations, although helpful in other ways, is likely to have no impact on the balance between negative and positive approaches. This is the advocacy of increased skepticism with regard to null hypothesis statistical testing (NHST).

To be sure, the limitations of NHST, as it is currently practiced, are real, and the widespread lack of understanding of these limitations has led to a great deal of ritualistic, often inappropriate, application of NHST (Gigerenzer 1993). Anything that may diminish the misunderstanding and ritualistic, inappropriate application of NHST in psychology is to be applauded.

However, although a more thoughtful approach to NHST by more researchers would help psychology in many ways, it would probably have little impact on the balance between positive and negative approaches in psychology. This is because NHST, whether used thoughtfully or ritualistically, may be used equally in support of positive conclusions as in support of negative conclusions.

It is true that the tools of NHST have mostly been used in social cognition research to promote negative conclusions. But this

https://doi.org/10.1017/S0140525X0432008X Published online by Cambridge University Press

is true only because of the predominance of negative conclusions in the literature. Computers and pencils have also been used primarily in the support of negative conclusions, for the same reason, but this does not make them inherently negative tools. Similarly, the use (and misuse) of NHST is equally amenable to supporting positive conclusions as negative ones. NHST can and has been used to promote positive conclusions, to dispute negative conclusions, and to promote balanced approaches and conclusions.

One available tactic in using NHST to support positive conclusions is to install an idealized version of an error in the role of null hypothesis, and then to "show" by rejecting the null hypothesis that the error in question is not real. I will illustrate this with the example of base-rate neglect. Sometimes it is used in a truly balanced way, which is exemplified by its use in the literature of confidence calibration.

The vast literature on base-rate neglect is replete with statistical demonstrations of the effect of just the sort that is criticized by K&F, that is, constructing a normative standard as a null hypothesis, and then showing that mean responding is sufficiently different from that standard to justify rejecting the normative standard as a description of performance. This leads to exactly the problem K&F identify in their Figure 1: pointed (and therefore impossible) rationality alongside a ranging bias.

However, the same tool is available to those of a different perspective. For example, Koehler (1996) drew the influential conclusion that "[w]e have been oversold on the base rate fallacy" (p. 1). One important rationale for this conclusion is that base rates are not, in fact, ignored entirely. Some proponents of the negative approach have come to extreme conclusions, such as "base rate information concerning categories in a population is ignored" (Nisbett & Borgida 1975, p. 935). This has led critics, such as Koehler, to note that people almost never ignore base rates completely. NHST, used to test the null hypothesis that manipulations of base rates have no impact, is guaranteed to show that people are sensitive to base rates, just as tests of a null hypothesis of rationality are guaranteed to show that people are not perfectly rational. Researchers who try to argue that people are insufficiently sensitive (although not totally insensitive) to base rates have been compelled to combat the perception that they are really arguing for a complete lack of sensitivity (e.g., Goodie & Fantino 1996; 2000).

In the literature on confidence calibration, the equivalent tactic would be to report a statistically significant correlation between confidence and accuracy as proof that people are sensitive to their own abilities and limitations. I am glad to be unaware of any research deploying this tactic, although the tool remains available for this kind of misuse. More commonly, though, researchers use NHST for the more appropriate and balanced approach of decomposing Brier (1950) probability scores into components, such as difficulty, calibration, discrimination, and noise (Carlson 1993). Here, NHST is not used in service of either positive or negative approaches, but for quantifying the balance between human achievement and human limitations.

NHST, like any tool, carries the potential for misuse as well as constructive use. It strongly appears to have been used (and misused) more for negative ends than for positive ends, and it is tempting for this reason to suspect that NHST is inherently slanted toward negative ends. But it has served negative approaches predominantly only because negative approaches have been predominant, and have had more occasions to invoke it. It is all to the good to encourage the psychological community to use NHST more judiciously, but this is an improvement that would likely be separate from any shift in the balance between positive and negative approaches to psychology.

Another route to broadening the scope of social psychology: Ecologically valid research

Samuel D. Gosling

Department of Psychology, The University of Texas at Austin, Austin, TX 78712-0187. **gosling@psy.utexas.edu**
http://homepage.psy.utexas.edu/homepage/faculty/Gosling/index.htm

Abstract: An imbalance is identified in social psychology between controlled experimental studies (which are common) and real-world, ecologically valid studies (which are rare). The preponderance of experimental studies (which provide mere existence proofs and lack realism) helps fuel social psychology's fault-finding focus. Laboratory experiments and ecological studies should be pursued jointly to examine social life in the real world.

Krueger & Funder (K&F) highlight an imbalance in social psychology, persuasively arguing that researchers devote far more attention to how people get things wrong (e.g., bias) than to how people get things right (e.g., accuracy). Here, I highlight another imbalance in social psychology, mentioned only briefly in K&F's article: that between the highly controlled experimental studies that characterize much of social psychological research, and real-world, ecologically valid nonexperimental studies, which are rather rare. I argue that this methodological imbalance contributes to the broader substantive imbalance lamented by K&F.

Thanks in part to the influence of social cognition, modern social psychology has become closely associated with one particular method – carefully controlled laboratory experiments. Although most, if not all, social psychological research is inspired by real-world events, it has become standard practice to control reality out of the picture. As has been argued elsewhere (e.g., Funder 1999; Kenny 1994; Rozin 2001), this over-reliance on laboratory experiments can limit the real-world relevance of the research. Two specific features of experiments fuel the field's fixation on finding faults.

First, experimental research is concerned with existence proofs. That is, laboratory experiments show only what *can* happen under some circumstances, but, without some ideas about the everyday context of the phenomenon, the experiments say very little about what *will* happen in the real world. Laboratory experiments tell us what is possible, not what is likely. To map out where the limits of possibility lie, experiments tend to focus on the points at which processes (e.g., decision making) break down. Consider the research programs on majority and minority influence. Essentially, these programs tell us merely that there exist circumstances under which the majority can influence the minority (e.g., Asch 1956), and there exist circumstances under which the minority can influence the majority (e.g., Moscovici 1980). However, the pinpoint focus of experimental studies on either the majority or the minority allows researchers to study these intrinsically related processes independently, bypassing the incontrovertible fact that in the real world, every time a minority exists so too does a majority. And, by overlooking the real-world contexts in which these social processes typically occur, the limit-testing experimental paradigm draws attention to the boundaries, where things break down, and neglects the question of how often these boundaries are approached by the circumstances of everyday life. If the conditions promoting majority or minority influence occur very rarely, then the negative influence of the majority or the minority could be practically inconsequential. Experiments show us where the boundaries of possibility lie, but social psychologists now need to map the real-world topography within those generous boundaries.

Second, the artificial experimental stimuli in experiments do not permit tests of accuracy. In their reasonable attempts to exert control, experimental social psychologists typically rely on artificial stimuli (e.g., vignettes), so that the variable of interest can be manipulated while keeping all other variables constant. Although useful from the standpoint of identifying causal processes, the

https://doi.org/10.1017/S0140525X0432008X Published online by Cambridge University Press