Theory & Psychology

Article

Significance tests as sorcery: Science is empirical—significance tests are not

Theory & Psychology
22(1) 67–90
© The Author(s) 2012
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0959354311429854
tap.sagepub.com
SAGE

Charles Lambdin
Intel Corporation

Abstract

Since the 1930s, many of our top methodologists have argued that significance tests are not conducive to science. Bakan (1966) believed that "everyone knows this" and that we slavishly lean on the crutch of significance testing because, if we didn't, much of psychology would simply fall apart. If he was right, then significance testing is tantamount to psychology's "dirty little secret." This paper will revisit and summarize the arguments of those who have been trying to tell us—for more than 70 years—that *p* values are not empirical. If these arguments are sound, then the continuing popularity of significance tests in our peer-reviewed journals is at best embarrassing and at worst intellectually dishonest.

Keywords

controversy, effect size, meta-analysis, null hypothesis, practical significance, replication, science, significance, statistics

In 1972, Polish-British sociologist Stanislav Andreski published *Social Sciences as Sorcery*, a luminous work that is still one of the most vitriolic diagnoses ever assembled of everything wrong with the social sciences. One of Andreski's key messages to the social science world is simple: real science is empirical, pseudo-science is not. Many social scientists make the claims they do, Andreski states, not because they have corroborated, diverse evidence supporting them as accurate descriptors of reality, but rather because they desire their opinions to *become reality*. This, Andreski argues, is shamanistic, not scientific. In this paper, the case is made that the social sciences (and particularly psychology) are rife with another kind of sorcery, a form of *statistical* shamanism—the test of "significance."

Corresponding author:
Charles Lambdin, Intel Corporation-Ronler Acres, 2501 Northwest 229th Avenue, Hillsboro, OR 97124-5506, Mailstop RA1-222, USA.
Email: charles.g.lambdin@intel.com

According to Andreski, many social science publications constitute little more than a translating of platitudes into jargon, where sophisticated statistical lingo lends a specious scientific air to one's pet hypotheses. I will here suggest that consequent to our generations-long obsession with *p* values and the statistical buffoonery which as a result passes for empirical research in our peer-reviewed journals, many psychologists are in fact guilty of what Andreski charges, and typically without even knowing it. Indeed, in the social sciences, the mindless ritual significance test is applied by researchers with little appreciation of its history and virtually no understanding of its actual meaning, and then—despite this alarming dearth of statistical insight—is held up as the hallmark of confirmatory evidence.

Methodologists have attempted to draw our attention to the foibles of significance tests for generations—indeed since well before our obsession with them even developed—and yet the fad persists, much to the severe detriment of the social sciences. This article chronicles many of the criticisms that have been leveled against "significance" testing and then comments on what the author feels is the most regrettable outcome of our observance of this null ritual, which is a vast and confused body of literature consisting largely of idiosyncratic results.

Psychology, *p* values, and science

These remarks are not intended to imply that the social sciences and science proper never overlap. Psychologists, for instance, certainly strive to be empirical, though whether psychology is a science is and has long been hotly debated (for an excellent discussion, see Rosenberg, 1988). Within psychology, the debate often boils down to a bar fight between experimental and clinical psychologists, with the former assuming scientific status while denying it to the latter (e.g., Dawes, 1994).

One of psychology's greatest thinkers, Paul Meehl, was an outspoken clinician, researcher, and prolific author. In 1978, Meehl (in)famously noted that theories in psychology, like General MacArthur's description of old generals, never really die; they just slowly fade away. After looking over 30 years of research, Meehl observed that theories in psychology do not accumulate as in the hard sciences; they are rather more akin to fads or trends that come into and go out of style. Rosenthal (1993) echoes this sentiment, noting that in psychology "we seem to start anew with each succeeding volume of the psychological journals" (p. 519).

Despite such arguments, experimental psychologists typically maintain that their work is science *because it is experimental*. But are the methods they typically employ actually scientific? Scientific research, after all, is scientific because it is empirical, not because it is experimental. Bakan (1974), for instance, argues that experimentation in no way guarantees empiricism, adding that much of the research in psychology is not empirical precisely because of the experimentation employed.Thus, experimental ≠ empirical.

Indeed, there seems to be a lack of appreciation among some researchers (not to mention the media and the public) that the results of any study can be preordained by the selection of stimuli, how variables are operationally defined, the construction of the experimental protocol, the level at which the data are aggregated, or the particular analysis the researcher chooses to utilize. Psychologist Mary P. Koss, for instance, once famously claimed that 27.5% of college women are victims of rape or attempted rape. What Koss didn't tell us is that her operational definition of "rape" was at odds with what the women surveyed actually believed. Going off the latter, only 4% of college women considered themselves rape victims (Murray, Schwartz, & Lichter, 2001). To take another example, whenever variables are controlled in a way that ignores their real-world interrelatedness, the spurious findings that emerge are not generalizable beyond the meretricious method used to artificially control the variables in question. Brunswik (1952, 1956) and Petrinovich (1979) famously argue that this fact alone casts doubt on a great deal of the research in psychology.

When Bakan (1974), however, warned us that the nature of our experimentation may actually be precluding psychological research from being truly empirical, he was referring to one practice in particular. In psychology we typically analyze our data in a way that often creates the impression of a finding that simply is not there (or that misses a finding that is). This ritualistic statistical observance is often called "null hypothesis significance testing" (NHST). There is nothing empirical, Bakan (1966) argues, about formulating a hypothesis and then collecting data until a significant result is obtained. Indeed, using this as the sole, shoddy litmus of evidential discovery, one can easily ferret out "support" for almost any bias.

In 1972, Andreski observed that in the social sciences, "[i]n comparison with half a century ago, the average quality of the publications (apart from those which deal with techniques rather than substance) has declined" (p. 11). I would argue that today the situation has grown even worse. Social scientists, he argues, all too commonly employ a jargonized statistical analysis to smokescreen the seriously flawed reasoning underpinning their conclusions. He calls this practice "quantification as camouflage."[1]This had gotten so bad, Andreski thought, that the "quantophrenics" in psychology should take a break from learning statistical technique to study some elementary logic and analytic philosophy.[2]

Twenty-eight years after Andreski published his book, Gigerenzer (2000) wrote that he spent an entire day and night in a library reading issues of the *Journal of Experimental Psychology* from the 1920s and 1930s. He became depressed because these 70-some-year-old articles made today's efforts pale in comparison in regards to experimental design, methodology, and statistical thinking.

Ranting to the wind?

In a recent article, Armstrong (2007) points out that, contrary to popular belief, "there is no empirical evidence supporting the use of statistical significance tests. Despite repeated calls for evidence, no one has shown that the applications of tests of statistical significance improve decision making or advance scientific knowledge" (p. 335). He is by no means alone in arguing this. Many prominent researchers have now for decades protested NHST, arguing that it often results in the publication of peer-reviewed and journal endorsed pseudo-science. Indeed, this history of criticism now extends back more than 90 years (e.g., Armstrong, 2007; Bakan, 1966, 1974; Berkson, 1938; Boring, 1919; Campbell, 1982; Carver, 1978, 1993; Cohen, 1990, 1994; Edwards, 1965; Falk, 1998;

Fidler, Thomason, Cumming, Finch, & Leeman, 2004; Fisher, 1955, 1956; Gigerenzer, 1987, 1993, 2004; Gigerenzer et al., 1989; Gill, 1999; Granaas, 2002; Greenwald, 1975; Hubbard & Armstrong, 2006; Hubbard & Lindsay, 2008; Hubbard & Ryan, 2000; Jones & Tukey, 2000; Kirk, 1996, 2003; Lindsay, 1995; Lykken, 1968, 1991; Meehl, 1967, 1978, 1990; Nunnally, 1960; Rosnow & Rosenthal, 1989; Rozeboom, 1960; Schmidt, 1992, 1996; Schmidt & Hunter, 1997; Sedlmeier & Gigerenzer, 1989; Skinner, 1972; Thompson, 1996, 1999, 2002, 2006, 2007; Tukey, 1991).

Despite a long line of top researchers making such points (e.g., Kirk, 1996, 2003), despite the fact that many methodologists will admit such things in conversation, the use of NHST is still standard practice in the social sciences. In his outgoing comments as editor of the *Journal of Applied Psychology*, Campbell (1982) wrote:

> It is almost impossible to drag authors away from their *p* values, and the more zeros after the decimal point, the harder people cling to them. It is almost as if all the statistics courses in the world stopped after introducing Type I error. …Perhaps *p* values are like mosquitos. They have an evolutionary niche somewhere and no amount of scratching, swatting, or spraying will dislodge them. Whereas it may be necessary to discount a sampling error explanation for results of a study, investigators must learn to argue for the significance of their results without reference to inferential statistics. (p. 698)

Nineteen years later, Finch, Cumming, and Thomason (2001) noted that little had changed in almost two decades.

> Why has [statistical] reform proceeded further in some other disciplines, including medicine, than in psychology? … What happened in psychology was not inevitable. We leave to historians and sociologists of science the fascinating and important question of why psychology has persisted for so long with poor statistical practice. (pp. 205–206)

Thus, to make the same points now that have been made repeatedly for decades is not, by any means, to beat a dead horse. A horse that is winning the race is not dead. The fact that this horse is still in the lead unfortunately suggests that the long history of authors making such observations have largely been ranting to the wind.

So what exactly is wrong with significance testing? It certainly has its defenders after all (e.g., Dixon, 1998; Hagen, 1997; Mulaik, Raju, & Harshman, 1997), but its defenders fail to address all of the complaints chronicled herein. Furthermore, their supportive arguments in no way get us around that fact that NHST does not lend objectivity to the process of making inferences (Armstrong, 2007; Bakan, 1966; Carver, 1978; Cohen, 1994; Kirk, 1996; Tukey, 1991), a fact which led Thompson (1999) to dub significance testing "pseudo-objectivity."

Even more troubling, significance testing creates in the minds of researchers the impression of automating the difficult process of thinking through inferences, seemingly reducing the complex notion of scientific support to the mindless task of an assembly line inspector, stamping "accept" or "reject" on every good that is rolled along. This

> practice of focusing exclusively on a dichotomous reject–non reject decision strategy of null hypothesis testing can actually impede scientific progress. … [F]ocusing on *p* values and

> rejecting null hypotheses actually distracts us from our real goals: deciding whether data support our scientific hypotheses and are practically significant. The focus of research should be on our scientific hypotheses, what data tell us about the magnitude of effects, the practical significance of effects, and the steady accumulation of knowledge. (Kirk, 2003, p. 100)

Journal editors, however, probably like this use of significance tests because it doubtlessly makes their jobs easier (Loftus, 1991). It makes the jobs of researchers easier as well. I would go so far as to say that many who today call themselves "scientists" would be unable to do so without the crutch of NHST to lean on. Just as most would rather take a weight-loss pill than have to diet and exercise, researchers seem all too readily lulled into a "false sense of science" by the convenience of clicking a few buttons and checking whether $p < .05$. The alternative is work—a lot of hard thinking and critical reasoning.

It is certainly a "significant" problem for the social sciences that significance tests do not actually tell researchers what the overwhelming majority of them think they do (Bakan, 1966). Bakan thought that "everybody knows this" and that to say it out loud is to be like the child who pointed out that the emperor is wearing no clothes. He argues that if we pull out the strand of NHST much of the tapestry of psychology would fall apart. Indeed, NHST, Gerrig and Zimbardo (2002) state, is the "backbone of psychological research" (p. 46). So, instead of abandoning it, which could be very embarrassing, we make the adjustment of simply misinterpreting what it actually tells us, which is … not much.

Significance testing and the infertility of psychology

Meehl (1967) likens NHST to a sterile intellectual rake who ravishes maidens but creates no viable scientific offspring. Cohen (1994) jokingly states he had to resist the temptation to call NHST "statistical hypothesis inference testing," which presumably would have yielded a more appropriate acronym. Countless methodologists tell us we should be focusing on effect sizes and confidence intervals, which—though both built on the foundation of significance testing—are together far more informative and meaningful. Even Fisher (1925) himself stated that ANOVA *p* values should be supplemented with η (see also Kirk, 1996). If so many of our best methodologists tell us we should be focusing on confidence intervals and effect sizes, why don't we listen?

Much of the blame lies with our journals, our statistics texts, and our graduate school statistics courses. The NHST orthodoxy is not only required for publication in most journals but its tenets continue to be proselytized from the pulpits of graduate school classrooms across the fruited plain. This trend has continued unabated since the 1960s. In 1972, B.F. Skinner complained that graduate schools teach "statistics in lieu of scientific method" (p. 319). This curriculum, he argued, is

> incompatible with some major features of laboratory research. As now taught, statistics plays down the direct manipulation of variables and emphasizes the treatment of variables after the fact. If the graduate student's first result is not significant, statistics tells him to increase his sample size. (p. 319)

And,

> What statisticians call experimental design (I have pointed out elsewhere that this means design which yields data to which the methods of statistics are appropriate) usually generates a much more intimate acquaintance with a calculating machine than with a behaving organism. (p. 320)

There are, of course, exceptions to this. A few journals have certainly striven to turn the tide. Bruce Thompson, for instance, has long campaigned for journals to require the reporting of effect sizes. A famous exception is found in the remarks of the great statistician William Kruskal, who, in response to an author using *p* values to assess the importance of differences, once wrote:

> So I'm sorry that this ubiquitous practice received the accolade of use by you and your distinguished coauthors. I am thinking these days about the many senses in which relative importance gets considered. Of these senses, some seem reasonable and others not so. Statistical significance is low on my ordering. Do forgive my bluntness. (as cited in Bradburn, 2007, p. 263)

Such exceptions to the norm are important, but are still too few and far between. Social science journals must require the use of effect sizes and the comparison of confidence intervals. Statistics texts must quit teaching NHST and ignorantly misrepresenting it as a happy compromise between Fisher, Neyman, and Pearson. Researchers must quit exploiting public trust in the social sciences by employing the $p < \alpha$ pseudo-litmus of empirical support to gain publications and continued paychecks.

The reason many researchers are so hesitant to comply is, Cohen (1994) suggests, that in psychology most confidence intervals are embarrassingly large. By not reporting miniscule effect sizes or hiding the girth of our confidence intervals, we can present our simple, nonsensical $p < .05$ results while keeping our uninformed readers wholly in the dark regarding the actual size, nature, and practical significance (or lack thereof) of the effects in question (Kirk, 1996). To give an example, it may seem impressive (but unsurprising) to state there is a significant correlation between being religious and having religious parents, but it is less impressive (and more surprising) if it is pointed out that for this statistic, $N = 2084$ and $r^2 = .01$ (Shermer, 2000)!

This example brings to mind Lykken (1968), who argues that many correlations in psychology have effect sizes so small that it is questionable whether they constitute actual relationships above the "ambient correlation noise" that is always present in the real world. Blinkhorn and Johnson (1990) persuasively argue, for instance, that a shift away from "culling tabular asterisks" in psychology would likely cause the entire field of personality testing to disappear altogether. Looking at a table of results and highlighting which ones are significant is, after all, akin to throwing quarters in the air and noting which ones land heads.

As a slight aside, some might here object to the author's use of correlations as examples, claiming that correlational research is not experimental and therefore not scientific. Because of this (and because of a reviewer for another journal making just this argument), the author would like to point out that this claim is patently false (see Cattell, 1978). Experimental design ≠ statistical analysis.What is usually meant by this argument is that *passive observational studies* are not experimental, not that "correlational" studies are not experimental. To misstate this is to misinform.

A true experiment must have three things: (a) random assignment of cases to levels; (b) manipulation of the levels of at least one independent variable (if you do not manipulate a variable, you cannot randomly assign cases to levels); and (c) control of extraneous variables (Tabachnick & Fidell, 2001). You *can* do these three things and then follow the design with a correlational analysis. And you might not manipulate a variable and then analyze your data using ANOVA. In the former case, a true experiment has been conducted (despite the use of correlation), and in the latter, the design is not experimental (despite the fact that an ANOVA was run). (This confusion becomes even more comical when one realizes that regression and ANOVA are basically the same thing anyway.)

Clearing the weeds so that something healthy might grow

It has been stated that NHST does not tell us what most researchers think it does. So what are the misconceptions? Bakan (1966) and Thompson (1996, 1999) catalogue some of the most common:

1. A *p* value is the probability the results will replicate if the study is conducted again (false).
2. We should have more confidence in *p* values obtained with larger *N*s than smaller *N*s (this is not only false but backwards).
3. A *p* value is a measure of the degree of confidence in the obtained result (false).
4. A *p* value automates the process of making an inductive inference (false, you still have to do that yourself—and most don't bother).
5. Significance testing lends objectivity to the inferential process (it really doesn't).
6. A *p* value is an inference from population parameters to our research hypothesis (false, it is only an inference from sample statistics to population parameters).
7. A *p* value is a measure of the confidence we should have in the veracity of our research hypothesis (false).
8. A *p* value tells you something about the members of your sample (no it doesn't).
9. A *p* value is a measure of the validity of the inductions made based on the results (false).
10. A *p* value is the probability the null is true (or false) given the data (it is not).
11. A *p* value is the probability the alternative hypothesis is true (or false; this is false).
12. A *p* value is the probability that the results obtained occurred due to chance (very popular but nevertheless false).

In their defense of significance testing, Mulaik et al. (1997) argue that such misconceptions about NHST are irrelevant as to whether we should continue its use. This is an awkward position. The point that Mulaik et al. seem to be missing is that methodologists are not trying to refute NHST when they write of how it is commonly misconstrued. The case made in this context is that the very concept of a *p* value is so seldom accurately grasped that the deleterious impact on the quality of research of which the social sciences are comprised is undoubtedly great—so much so that it should be quite uncontroversial to state that most published research is in fact nonsense. The great philosopher of science Imre Lakatos considered most published research in the social sciences to be little more than "intellectual pollution" (reported by Meehl, 1990) and Ioannidis (2005) argues that the spread of the *p* value from psychology to other fields has resulted in a world where, even in medicine, most published findings are probably wrong.

Mulaik et al.'s (1997) point is still a sound observation. A proper response, however, is not to continue with our blind use of NHST, but to do something about it. Before proceeding, let us state just what exactly a *p* value tells us. *A p value is the probability of obtaining the results in hand, assuming that the statistical null hypothesis is true in the population*. That is all and nothing more. As Schopenhauer (1851/2004) reminds us, nothing more is implied by a premise than what is already contained in it, and this, it is time we admit, does not imply much.

Faddish falsities

Let us now attempt to extirpate some of these statistical delusions so that our attention may then be focused on NHST's actual flaws. This is no easy feat. Hubbard and Armstrong (2006) argue that the misconceptions regarding significance testing among researchers are wider and deeper than even the critics appreciate. Indeed, there is undeniably something incredibly amiss when the APA's own Task Force on Statistical Inference, formed to address this very problem, itself embarrassingly misses the fact that Fisher's evidential statistic (*p* value) and Neyman–Pearson's error estimate (α) are not in any meaningful way combined when stating that "$p < .05$" (Hubbard, 2004; Wilkinson & The TSFI, 1999).

The most common and destructive delusions are, in my opinion, that *p* values somehow tell you (a) the odds your data are due to chance, (b) the odds your research hypothesis is correct, (c) the odds your result will replicate, and (d) the odds the null is true. Let us now take a closer look at each of these falsities.

The odds your data are due to chance

Even after reading articles such as this one, most researchers simply file it away as "interesting" and then go right on treating NHST as though it somehow tells them the odds of their results. This is likely the most common misconception regarding *p* values. The statement that if a result is significant, its odds of occurring due to chance are only 1 out of 20, or 5 out of 100, seems omnipresent and yet is wholly false. Carver (1978) calls this the "odds-against-chance fantasy." Of this misconception, Mulaik et al. (1997) comment that "[i]t is hard to imagine how someone would arrive at this conclusion" (p. 74) and then go on to speculate what process of reasoning might lead one there. For my own part, I doubt that many reason their way to this conclusion. Many were simply taught this interpretation and then—it itself sounding so plausible on its face—few likely go on to question it.

Indeed, this misconception is widely taught in graduate school classes and can easily be found asserted by prominent thinkers (e.g., Anastasi, 1976; Hebb, 1966; Paulos, 2003). What matters, however, is the fact that it is so widely held, regardless of why. Carver (1978, 1993) believed this misconception to be widespread in education and Bakan (1966) believed it predominant in psychology. I would argue that this is still true and would also submit that I have seen this misconception prevalent in both academia and industry. Quite embarrassingly, even a reviewer of this article for a leading methodology journal (though not the one you are currently reading) objected to this very discussion, insisting that *p* values do indeed tell one the odds of one's data occurring due to chance.

As Falk (1998) and Falk and Greenbaum (1995) observe, this misconception is derived from a legitimate concern. If your research results are based on a random sample (or participants have been randomly assigned to levels), then the worry can and should arise that the results obtained may in fact be a fluke. NHST attempts to address this concern—the problem is that it fails. And fail it must, as Carver (1978) succinctly points out.

A *p* value does not and cannot tell you the odds your results are due to chance, because it is calculated based on the assumption that this probability is already 100%. In other words, a *p* value tells you the odds of your results given that you assume they are in fact due to chance. This brings back to mind the Andreskian admonition that less statistics classes and more courses in simple analytic philosophy are in order. It certainly seems that one of the most robust empirical conclusions collectively reached by the social sciences is that the overwhelming majority of social scientists do not know what a conditional probability is.

The proper use of a *p* value is to assist in deciding whether the probability is in fact 100% that your results are due to chance. Perhaps then the best wording for a low *p* value is simply to state: "Assuming my results are due to chance, my obtained mean difference is very unlikely. Therefore chance may not be the culprit. Now it is up to me to employ *other methods* to determine what that culprit might be."

The odds your research hypothesis is correct

We have seen that experimental ≠ empirical and that experimental design ≠ statistical analysis. To properly interpret *p* values, one must also keep in mind both that statistical hypotheses ≠ research hypotheses and that $P(D|H_0) \neq P(H_1|D)$. This seems obvious when read as an explicit statement, but not when reading the discussion sections of psychology papers. The equation, $P(D|H_0) \neq P(H_1|D)$, implies that a low probability of a result (or data) given the truth of the null does not indicate the probability of the alternative given the data. Such a misstatement all too readily lends itself to the erroneous belief that rejecting the null indicates that your treatment works. That the treatment does not work is not your null hypothesis and that the treatment works is not the statistical alternative. Your null is likely $\mu_A - \mu_B = 0$ and your alternative is likely $\mu_A - \mu_B \neq 0$. Rejecting H_0 implies only that $\mu_A - \mu_B \neq 0$, not that your treatment works. There are virtually an infinite number of reasons why $\mu_A - \mu_B \neq 0$.

With a null of no difference and an α of .05, what a significant result indicates is that you would find the difference obtained less than 5% of the time if in reality $\mu_A - \mu_B = 0$.

The null refers only to the statistical hypothesis. That the treatment works is the research hypothesis. As Granaas (2002) reminds us, the rejection of any nil hypothesis (a null of no difference) supports *all* research hypotheses predicting an effect, *not just yours*—and there may be an infinite number of explanations for the effect in question. In fact, all significance here tells you is that you are justified in proceeding to test your research hypothesis, not that your research hypothesis is supported. And yet this fantasy leads to many articles being accepted for publication whenever p values are erroneously taken to suggest the research hypothesis is likely correct when the methodology and hypotheses involved are themselves doubtful (Carver, 1976, 1978; Lykken, 1968).

A related error, and a comical one at that, occurs whenever one sees talk of p values measuring "degrees of significance." If a $p < .05$ result is "significant," then a $p = .067$ result is not "marginally significant." Similarly, if α is fixed in advanced at .05 (as it typically is), then it is nonsensical to say that a $p < .001$ result is "highly significant." It is either significant or it is not. Hubbard and Armstrong (2006) conducted a survey of marketing journals and found that 54.9% of articles examined committed this error.

Supporting a research hypothesis against all competing rival hypotheses which explain a given effect is not something significance testing can help with. Such support (a.k.a., scientific support) is gained only after meticulous theorizing, sound methodology, and numerous replications lead to diverse, corroborating evidence demonstrating the effect in a variety of situations (Carver, 1978).

The odds your result will replicate

It is an illusion to think one can learn anything about the replicability of a finding from a p value. Fisher himself was well aware of this fact (e.g., Fisher, 1929; see also Salsburg, 2001; Tukey, 1991), though many of those who have adopted the idea of significance testing seem to have altogether forgotten it. As Thompson (2003) states

> If p calculated informed the researcher about the truth of the null in the population, then this information would directly test the replicability of results. … Unfortunately, this is not what statistical significance tests, and not what the associated p calculated values evaluate. (p. 96)

Consequently, it is entirely false to state or imply that $1 - p$ = the probability that the results are replicable/reliable (Carver, 1978).

As Rosenthal (1993) observes, if there is a real effect in nature with a d of .5 ($r = .24$) and a researcher conducts a study with an N of 64 (and so the power of the study is .5—a typical power level in psychology), and then someone else replicates this study, there is only a one in four chance that both researchers will find that $p < .05$, even though the effect is real. If three more researchers replicate the study, there is only a 50/50 chance that three or more of the studies will find a significant result. This is obviously not conducive to the accumulation of knowledge. This is unfortunate in that replication is greatly needed and of the utmost importance in psychology. As Lykken (1991) points out, when we bother to look, many of psychology's findings actually do not replicate.

What matters in psychology is preponderance of evidence. And if one takes Meehl's neo-Popperian/neo-Lakatosian view (and there is much to be said for this view), then—stemming both from its rejection of strict falsificationism and from the observation that in psychology all theories are technically false (in that they are incomplete)—what also matters is (a) “money in the bank” and (b) “damn strange coincidences” (Meehl, 1990). In other words, since no psychological theory can ever be “the whole truth,” they should seek to earn a sort of “good enough verisimilitude” to warrant our continuing to entertain them. This is done, Meehl (1990) argues, by their accruing “money in the bank” via Wesley Salmon’s notion of “damn strange coincidences” (Salmon, 1984). As Meehl (1990) argues, “*The main way a theory gets money in the bank is by predicting facts that, absent the theory, would be antecedently improbable*” (p. 115). The role of significance tests in this process is “minor and misleading” (p. 115). Indeed, psychology’s track record at making predictions is extremely embarrassing. As many know, predictions made by expert psychologists typically do not outperform those made by laypersons (see, e.g., Faust & Ziskin, 1988).

If replication had been stressed by our journal editors over and above the specious requirement of small *p* values, things might today be different. Sadly, journals typically want “original” ideas, not replications, even though it is likely that the impracticable expectation that every researcher be “original” only encourages intellectual dishonesty and dilutes the overall quality of published research as a whole (Andreski, 1972).[3] Further, a sound replication that does not achieve statistical significance is not likely to be published anyway, even though the results of a study—so long as its methodology is sound—should be irrelevant as to whether it gets published (Mahoney, 1977).

The odds the null is true

It should be remembered that science and significance testing do not ask the same question. A *p* value is the probability of the results in hand assuming that the null is true in the population. Many researchers get this definition backwards: a *p* value does not tell us the probability that the null is true in the population given the results (which would be science). Thus, the results of significance tests are typically lent an entirely inappropriate Bayesian interpretation.Those who defend NHST are typically guilty of this error in elementary logic (Cohen, 1994).

This is likely due to wishful thinking. As scientists, what we should be interested in is the Bayesian probability the null hypothesis is correct given the evidence or data, $P(H_0|D)$, not the odds of obtaining the data we did (or more extreme data) assuming the null is true in the population, $P(D|H_0)$. Significance tests can only tell us the latter. Unfortunately, $P(D|H_0) \neq P(H_0|D)$.

To illustrate, take the probability of an abused child having nightmares, P(N|A) and the probability that a child who has nightmares is abused, P(A|N). Clearly $P(N|A) \neq P(A|N)$: that is, knowing that abused children are likely to have nightmares does not imply that a child who has nightmares is likely abused (Dawes, 2001). Knowing the odds are low of the mean difference obtained given the assumed truth of the statistical null does not imply that the null is likely false given the evidence. Ignoring this is described by Falk and Greenbaum (1995) as the “illusion of probabilistic proof by contradiction” and by Gigerenzer (1993) as the “permanent illusion.”

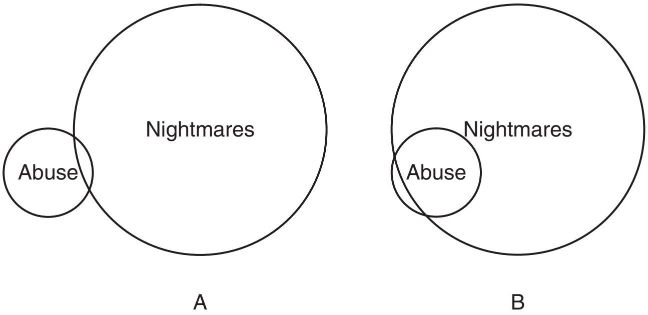

Figure 1. Venn diagram: Abuse and nightmares.

It is true, however, that knowing $P(D|H_0)$ can *influence* $P(H_0|D)$. Though $A \rightarrow B \neq B \rightarrow A$ is a true statement, so is the following: If $A\uparrow B$ denotes the situation $P(B|A) > P(B)$, then $A\uparrow B \rightarrow B\uparrow A$. In terms of significance testing, if $P(D|H_0) < P(D)$, then showing that the probability of our data is low given the assumed truth of the null (i.e., a low *p* value) *does reduce* the conditional probability of the null given the data, $P(H_0|D)$. The problem is that it does not indicate $P(H_0|D)$ as a precise probability (Falk, 2008). In other words, though it is true that if $P(D|H_0) < P(D)$, then $D\downarrow H_0 \rightarrow H_0\downarrow D$, this in no way indicates $P(D|H_0) = P(H_0|D)$ (Falk, 2008). As noted above, NHST tries to address a legitimate concern. The problem is that it fails to address it. Demonstrating that $P(D|H_0)$ is low may indeed reduce $P(H_0|D)$, but it does not demonstrate that $P(H_0|D)$ *is also low*, which is what (as scientists) we would be interested in seeing (Carver, 1978; Cohen, 1994; Kirk, 1996).

Returning to the example above, if the probability that an abused child will have nightmares is greater than the unconditional probability of children having nightmares, and if the probability that an abused child will have nightmares is high, this does increase the probability that a child who has nightmares has also been abused, but it in no way indicates that it is in any way likely. It is still doubtlessly very unlikely that a child who has nightmares is also an abuse victim. This can easily be seen by inspecting Venn diagrams (see Figure 1).

In A, abuse and nightmares barely overlap, which means fewer children who have nightmares are also abused. In B, the overlap is more pronounced, which means, compared to A, more children who have nightmares are necessarily also abuse victims, but the overwhelming majority of children who have nightmares are still abuse-free. In short, just because $P(D|H_0)$ is low does not mean that $P(H_0|D)$ is also low.

Here is an illustration from mathematician John Allen Paulos (2003): We want to know if a suspect is the DC sniper. He owns a white van, rifles, and sniper manuals. We think he's innocent. Does the probability that a man who owns these items is innocent = the probability that an innocent man would own these items? Let us assume there are about 4 million innocent people in the DC area, one guilty person (this is before we knew there were two) and that 10 people own all the items in question. Thus, the first probability, the probability that a man who has these items is innocent, is 9/10, or 90%. The latter probability, the probability that an innocent man has all these items, is 9/4,000,000, or .0002%.

Having examined some common misconceptions, let us now move on to NHST's actual problems.

Etiology and evaluation of the NHST virus

Owing chiefly to the inaccuracies in many of our college textbooks, the origins of NHST have practically become cloaked in myth (Gigerenzer, 2004). The NHST ritual is a hybrid of Fisher's *p* value and null hypothesis test and Neyman and Pearson's alternative hypothesis and Type I error rate. It bears mentioning that the man who started the practice of significance testing and then introduced such methods to the social sciences was not Fisher, but Francis Ysidro Edgeworth, a distant cousin of Francis Galton (Edgeworth, 1886; Stigler, 1986). Further, it should be made plain when discussing NHST that its pioneers (e.g., Fisher, Neyman, and Pearson) are not to blame for our current ill-considered misapplication of their ideas. As Gigerenzer (2004) states, "Each of these eminent statisticians would have rejected the null ritual as bad statistics" (p. 589).

The two models differ not only in statistical method but in philosophic views. Neyman and Pearson rejected null hypothesis testing and emphasized error detection and the determination of β, "which is not a part of the null ritual" (Gigerenzer, 2004, p. 589). Indeed, the appropriate use of the Neyman–Pearson model is far narrower than current statistical (mal)practice would suggest (a typical legitimate application would be quality control: Gigerenzer, 2004; Hubbard & Armstrong, 2006). Though Fisher's early work (e.g., Fisher, 1925, 1935) advocated the use of cutoffs for significance, and though it has been argued that the force of his early rhetoric likely prompted our adoption of significance testing (Yates, 1951), it should be noted that Fisher (1955, 1956) himself later objected both to the use of significance tests for accept–reject decisions and to the idea of fixed alpha levels. He eventually considered it naïve to assume that scientists actually conduct the same test repeatedly. Instead, he argued, one should report the exact *p* value without making a dichotomous accept–reject decision (Gigerenzer, 2004).

Sadly, the social sciences took no notice and the influence of Fisher's early work is still ubiquitous (Hubbard, 2004). The typical social science researcher wishing to conduct a statistical analysis decides on α, calculates *p*, and then compares *p* to α without appreciating that they do not actually have anything to do with one another. In our obsession with this "$p < \alpha$" gobbledygook, we collectively fantasize that our *p* value is akin to a temperature (call it *t*), that α is akin to knowing that water freezes at 32 degrees Fahrenheit, and that if $t < 32$ degrees, then the water is "statistically frozen." This is nonsense.

Fisher's *p* value is the probability of seeing results ≥ your own given the truth of the null. Neyman and Pearson's α is the probability of committing the error of falsely rejecting the null. Thus, α is an error probability; *p* is not (Hubbard & Armstrong, 2006). A commonly missed point is that α is not concerned with inductive inference at all. It is concerned with minimizing error in the long run. Fisher's *p* value, by contrast, has nothing to do with Type I error. It is an evidential probability intended to assist one in making a scientific inductive inference based on the results of experimentation.

Neyman and Pearson's α is the probability a rejected null is actually true. Since *p* already assumes the truth of the null, the probability of falsely rejecting the null is irrelevant. The calculated value for p is conditional on the probability of the null being true = 1. Therefore, to know the exact value of p and the probability of one's results, one must first assume the probability of falsely rejecting the null is zero. *Thus, stating that "*p *< α" is utterly meaningless*. Despite this, this easily achievable demonstration of baloney is all our peer-reviewed journals seem interested in. This state of affairs might seem more fitting if the presenters of such a tortured logic were Lewis Carroll or Groucho Marx and not the majority of social science researchers, textbook authors, and journal editors.

The etiology of this error is difficult to trace. What is known is that in psychology, the Fisher and Neyman–Pearson models somehow confusedly evolved into the amalgamated slapdash hodgepodge of NHST, which in the 1950s haphazardly became enshrined as the be-all and end-all approach to statistical analysis, institutionalized by professional associations and curricula alike (Gigerenzer, 1987, 1993, 2004), much to the dismay of Fisher himself (Fisher, 1955, 1956). This mindless, hybrid approach to statistical analysis quickly spread like a virus from psychology to other fields (including the medical and biological sciences), much to their loss (Gigerenzer, 2004; Ioannidis, 2005).

Much of the blame for NHST's popularity likely rests squarely with the APA. As Gigerenzer (2004) observes, the 1952 first edition of the *Publication Manual of the American Psychological Association* stressed significance testing and even went so far as to ignorantly dissuade authors from reporting nonsignificant results. The 1974 second edition famously states: "Caution: Do not infer trends from data that fail by a small margin to meet the usual levels of significance. Such results are best interpreted as caused by chance and are best reported as such" (APA, 1974, p. 19). Thus, the second edition of the *Manual* simultaneously achieves two embarrassing things: (a) it gives advice based on Fisher's early work while ignoring his later arguments; and (b) it encouraged belief in Carver's (1978) "odds-against-chance fantasy." Though the third edition of the *Manual* omitted this passage, your chances of finding a social science statistics text that points such things out is virtually zero (Gigerenzer, 2004).

The nil hypothesis

If your null hypothesis is a mean difference of zero, then your null is what is commonly referred to as "the nil hypothesis." This is the most common hypothesis used in psychology, and many respected methodologists have observed the imprudence of basing our inferences on such a dubious assumption (e.g., Kirk, 1996; Meehl, 1978). As Tukey (1991) states, "[T]he effects of A and B are always different, in some decimal place, and so to ask if they are different is foolish" (p. 100); and Bakan (1966), "There really is no good reason to expect the null hypothesis to be true in any population" (p. 426). Edwards (1965) makes clear why it should matter whether the null is plausible or scientifically preposterous: "If a hypothesis is preposterous to start with, no amount of bias against it can be too great. On the other hand, if it is preposterous to start with, why test it?" (p. 402)

Not everyone agrees with this assessment (e.g., Hagen, 1997; Mulaik et al., 1997). Hagen (1997), for instance, argues that though there is likely always some effect present that could theoretically be measured on some variable, there is no reason to expect this to be the dependent variable in question. But how can one expect it *not* to be the dependent variable? Further, how can one demonstrate a justification for this assumed immunity of the variable of interest? Hagen seems to be begging the question. Furthermore, as Jones and Tukey (2000) state in response to Hagen, “We simply do not accept that view” (p. 412).

> For large finite treatment populations, a total census is at least conceivable, and we cannot imagine an outcome for which $\mu_A - \mu_B = 0$ when the dependent variable (or any other variable) is measured to an indefinitely large number of decimal places. … For hypothetical treatment populations, $\mu_A - \mu_B$ may approach zero as a limit, but as for the approach of population sizes to infinity, the limit never is reached. The population mean difference may be trivially small, but will always be positive or negative. (p. 412)

Thus, according to Jones and Tukey, $\mu_A - \mu_B$ is either > 0, < 0 or not (yet) determined. It never $= 0$.

If Jones and Tukey (2000; among others, e.g., Kirk, 1996) are correct, then regardless of your null hypothesis, $\mu_A - \mu_B \neq 0$, which is why the nil hypothesis can always be rejected with a large enough sample size. Bakan (1966) empirically supported this by running statistical tests on 60,000 subjects. He writes that it didn’t matter whether subjects were divided into North vs. South, East vs. West, or Maine vs. the rest of the US—everything was significant. If this sounds to any reader like a demonstration of a technical point that is unrepresentative of “actual research,” I would then point out that Bakan’s demonstration is highly representative of much of what now passes for science in public health.

Such a discovery is also discussed in Nunnally (1960). Mulaik et al. (1997) object that it is contradictory to use a significance test to empirically support the proposition that significance tests are unempirical. I would suggest that, again, Mulaik et al. are missing the point. The point here is that it most certainly does matter whether the null hypothesis is empirically plausible or not. What both Bakan and Nunnally have done is shown us the circularity of NHST, of running lots of subjects and then testing to see whether we ran lots of subjects. This point was also made almost three-quarters of a century ago by Berkson (1938), who astutely observes:

> It would be agreed by statisticians that a large sample is always better than a small sample. If, then, we know in advance the *P* that will result from an application of the Chi-square test to a large sample, there would seem to be no use in doing it on a smaller one. But since the result of the former test is known, it is no test at all. (p. 526)

This brings up another point. If using a nil hypothesis you have no evidential reason to suspect is in any way plausible, why set up a statistical test to control for phantom Type I error? Why focus on the probability of falsely rejecting the implausible (Kirk, 1996)? Indeed, researchers often allow the real threat of Type II errors to remain extraordinarily high, typically as high as 50 to 80% (Cohen, 1962, 1994; Kirk, 1996; Sedlmeier & Gigerenzer, 1989). This implies that correcting for “alpha inflation” is often a particularly dubious procedure. Adjusting for alpha error across multiple tests typically only leads to an overestimation of the population effect size and reduces power (Cohen, 1994;

Sedlmeier & Gigerenzer, 1989). But if we are using a nil hypothesis, then power is simply a calculation of whether we have run enough subjects to detect what we empirically already suspect to be the case. Lykken (1968) observed that the odds of supporting a directional prediction through NHST—*even if the research hypothesis in question is false*—are typically 50–50. Why not just flip a coin instead? After all, all rejecting a nil hypothesis tells us is the direction of an effect (Granaas, 2002).

Armstrong (2007) argues that it only makes sense to test a null if it is a reasonable and likely conclusion. Science and knowledge, after all, advance more rapidly through the rejection of theories (Greenwald, 1975; Popper, 1935/1959). The hypothesis of interest, this implies, should be set up as the null. This is what Platt (1964), in his classic paper, calls "strong inference." Armstrong adds that even if the null is a reasonable hypothesis, effect sizes and confidence intervals should be the focus, not significance tests. He concludes that significance tests are unnecessary even when conducted and interpreted correctly and that ultimately all they really do is "take up space in journals" (Armstrong, 2007, p. 336). He is not alone in thinking this. Cohen (1994) states, "Even a correct interpretation of *p* values does not achieve much, and has not for a long time" (p. 1001); and Kirk (1996), "I believe that even when a significance test is interpreted correctly, the business of science does not progress at it should" (pp. 753–754).

We have seen that even when applied without laboring under common misconceptions, NHST still does not achieve much. I would now like to turn our attention to what I feel is the severest outcome of our abject failure to prevent NHST from running rampant for decades in the social sciences.

Cherry picking from the sea of happenstance

It was stated in the Introduction that the most unfortunate consequence of psychology's obsession with NHST is nothing less than the sad state of our entire body of literature. Our morbid overreliance on significance testing has left in its wake a body of literature so rife with contradictions that peer-reviewed "findings" can quite easily be culled to back almost any position, no matter how absurd or fantastic. Such positions, which all taken together are contradictory, typically yield embarrassingly little predictive power, and fail to gel into any sort of cohesive picture of reality, are nevertheless separately propped up by their own individual lists of supportive references. All this is foolhardily done while blissfully ignoring the fact that the tallying of supportive references—a practice which Taleb (2007) calls "naïve empiricism"—is not actually scientific. It is the quality of the evidence and the validity and soundness of the arguments that matters, not how many authors are in agreement. Science is not a democracy.

It would be difficult to overstress this point. Card sharps can stack decks so that arranged sequences of cards appear randomly shuffled. Researchers can stack data so that random numbers seem to be convincing patterns of evidence, and often end up doing just that wholly without intention. The bitter irony of it all is that our peer-reviewed journals, our hallmark of what counts as scientific writing, are partly to blame. They do, after all, help keep the tyranny of NHST alive, and "[t]he end result is that our literature is comprised mainly of uncorroborated, one-shot studies whose value is questionable for academics and practitioners alike" (Hubbard & Armstrong, 2006, p. 115).

In many fields in the social sciences, cherry picking is powerfully reinforced by the strong contingencies within which researchers operate. The incentives are strong. Many academics would not get funding or tenure without selectively supporting their body of work, even if it's nonsense (Andreski, 1972). Many doing research for companies are understandably reluctant to bite the hand that feeds them. Many corporate researchers are paid to find something specific and are typically statistically savvy enough to paint the picture they are hired to paint. A great many nonprofits could simply not survive without cherry-picked supportive findings. Public health initiatives often employ statisticians whose sole function is to sift through datasets fishing for numbers that support the funded program in question, ignoring the fact that the same spreadsheet also contains data that annihilates the culled finding. This practice is both fraudulent and unethical.

In clinical psychology, there is so much data available that evidence can be assembled to support almost any hypothesis. Published studies making competing claims recurrently have little overlap in their references. To take a now famous example, Sawyer's (1966) summary of the literature on clinical vs. mechanical (actuarial) prediction concludes that mechanical prediction is superior. Korman's (1968) review argues that clinical prediction is superior. Holt (1970) points out that though Sawyer and Korman's "reviews" of the scientific literature are only two years apart, there is zero overlap in their references.

Such cherry picking is not the type of synthesis scientific theories are supposed to represent. The marshalling of cherry-picked evidence to support pet claims is not science but an agenda. It replaces science with a parlor game that few will take seriously. We are supposed to be detectives after all, not just "advocates."

Because of this, perhaps the most important type of study that can presently be done in the social sciences is the meta-analysis of large and representative collections of studies (Schmidt, 1992). Sticking with the above example, the famous Grove, Zald, Lebow, Snitz, and Nelson (2000) meta-analysis quite convincingly demonstrates that mechanical prediction does indeed outperform clinical prediction.

As Armstrong (2007) points out, there are those who argue this is why we still need significance tests: meta-analyses require them, they say. This claim is false (Armstrong, 2007; Rosenthal, 1993; Schmidt & Hunter, 1997). Not only are significance tests unnecessary for meta-analyses, but as Schmidt and Hunter demonstrate, meta-analyses employing effect sizes and confidence intervals are superior. Broad meta-analyses of the studies conducted on a topic could perhaps aid in salvaging something empirical from more than half a century of fraudulent statistical assumptions. Given this fact, effect sizes should always be reported, if only so that they can be included in future meta-analyses, which need to be conducted in attempt to sift the gold from the humbug (Thompson, 1999).

The sheer enormity of the literature itself makes this task daunting. Twenty years ago it was estimated that in peer-reviewed psychology journals alone an article is published every 15 minutes. That's 35,040 articles a year (Thorngate, 1990)! That is ridiculous. It is perhaps unsurprising then that most published papers are likely never read, that one of the biggest predictors of whether an author will cite a paper is whether he has seen it cited by others, and that, because of such factors, papers that become well known do so largely owing to chance (Taleb, 2007). A partial solution to some of the problems noted above would simply be to quit requiring academics to publish in order to work and teach. I would much rather require that they stop teaching factually incorrect statistics than that they publish pseudo-original work. The unfortunate publish-or-perish reality in academia has led some to argue that much unscientific research is likely conducted by academics simply to ensure their own professional survival (Andreski, 1972).

Conclusion

This paper was meant not as a condemnation of psychology but rather as a call for reform. Our obsession with statistical tests of significance has made much of our research blatantly unscientific. Bakan (1966) reminds readers of Tukey's (1962) point that statistical tests often pull our attention away from the data themselves, which are ultimately what we are supposed to base inferences on. As Bakan (1966) notes, "When we reach a point where our statistical procedures are substitutes instead of aids to thought, and we are led to absurdities, then we must return to the common sense basis" (p. 436); and, "We must overcome the myth that if our treatment of our subject matter is mathematical it is therefore precise and valid. Mathematics can serve to obscure as well as reveal" (p. 436).

This is what Andreski (1972) referred to as "quantification as camouflage." The point of statistics is to clear the fog and help us identify patterns and relationships in cumbersome data that the naked eye cannot detect unaided. All too often, however, statistics instead serves—indeed, is all too often intentionally employed—to smoke-screen what the naked eye does indeed see unaided: that there's nothing there. If a p value is significant but confidence intervals are wide, effect sizes are minute, and corroborating, diverse evidence simply is not there, the p value is misleading you. *Ignore it*. To fail to do so is to mislead.

There is no substitute for eyeballing your data. Some surveyors of psychological research have concluded that .5 is the average effect size in some fields (e.g., Sedlmeier & Gigerenzer, 1989). Cohen (1992) argues that such an effect should be visible to the naked eye. There is no substitute for what Edwards, Lindman, and Savage (1963), in their classic article, call the "interocular traumatic test," which is when a relationship hits you right between the eyes. When your data show a clear relationship, *you typically do not need a significance test to see it*. How many times have researchers known an effect was present but failed to publish because of the tyranny of the p value? How many times have useless effects gained acceptance because N was sufficiently large to keep p sufficiently small, slipping shoddy results through the door utilizing the standard "$p < \alpha$" gobbledygook?

The making of scientific inferences is always a qualitative process. It is something that we must do ourselves. It can be helped by mathematics, but it cannot be *replaced by* mathematics. Math does not do the reasoning for us. The hard work will always take place between our ears. If we feel that our results, without being dressed up with excessive quantification, are not fancy enough, then we should remind ourselves that almost all great discoveries in science were qualitative in nature (Uttal, 2005). Furthermore, many of the genuine great discoveries in psychology were made by researchers who were not using significance tests (Gigerenzer, 2004).

Despite the (paltry and counterproductive) efforts of the APA to right its wrongs, the tyranny of significance testing continues to reign. And reign it will until the incentives in place begin to change. The ritual observance of this statistical folly will continue unabated so long as it serves as the near-exclusive rite of passage into the undeservingly revered halls of social science peer-reviewed journals. It should by now not sound overly cynical to observe that the business of most social scientists is to stay in business. Most psychologists are not sycophantic myrmidons of the church of NHST. As many economists well realize, process and incentives are what matter most, and the fact of the matter is that social scientists must make a living and they must publish to do so. It should in no way surprise us then that most social scientists are far more interested in publishing articles than they are in statistical thinking or producing science (Gigerenzer, 2004).

As Roseanne Conner's father says in the sitcom *Roseanne*, "You can lead a horse to water, but you can't make him think." Paraphrasing Fidler et al., (2004), "You can lead a social scientist to a confidence interval, but you can't make him think about statistics." The recommendations of the APA's "task force" (Wilkinson & The TFSI, 1999) will undoubtedly go entirely unheeded until our peer-reviewed journals lead the way and advocate change by themselves changing the requirements for publishing.

I do not harbor the vain illusion that the present article will change current practice. This article is only a reminder of the far better ones referenced herein. I merely hope it to further raise awareness. In conclusion, it is time that we put our *p* values away and get around to the business of science.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Notes

1. Meehl (1978) found Andreski's "hatchet job" of the social sciences so effective that he argued it should be required reading for all social science Ph.D. candidates.
2. Andreski (1972) observes that in the social sciences, "[p]retentious and nebulous verbosity, interminable repetition of platitudes and disguised propaganda are the order of the day" (p. 11). As noted above, Andreski warns us that the disguising of propaganda as research is shamanism, not science. And today things are worse, not better. Indeed, in this day and age, it would behoove all social scientists to bear in mind that the wrapping of advocacy in the cloak of the unfettered pursuit of truth is, as Andreski calls it, sorcery. The more political the social sciences become, the more fraught with propaganda (and shamanism) are its journals. This disturbing practice can today be seen taking over pockets of the social sciences, such as community psychology, where inherently anti-scientific causes such as "social activism" threaten to suffocate any attempt at intellectually honest and open-minded scientific inquiry (for an interesting discussion, see Wright & Cummings, 2003).
3. Indeed, many will likely dismiss this paper by flippantly observing that I have said nothing "original." That is irrelevant. What matters is this: Am I right or wrong?

References

American Psychological Association. (1952). *Publication manual* (1st ed.). Baltimore, MD: Garamond/Pridemark Press.

American Psychological Association. (1974). *Publication manual* (2nd ed.). Baltimore, MD: Garamond/Pridemark Press.

Anastasi, A. (1976). *Psychological testing* (4th ed.). New York, NY: Macmillan.

Andreski, S. (1972). *Social sciences as sorcery*. London, UK: André Deutsch Limited.

Armstrong, J.S. (2007). Statistical significance tests are unnecessary even when properly done and properly interpreted: Reply to commentaries. *International Journal of Forecasting*, *23*, 335–336.

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

Bakan, D. (1974). *On method: Toward a reconstruction of psychological investigation*. San Francisco, CA: Jossey-Bass.

Berkson, J. (1938). Confidence curves: An omnibus technique for estimation and testing statistical hypotheses. *Journal of the American Statistical Association*, *33*, 526–542.

Blinkhorn, S., & Johnson, C. (1990, December 27). The insignificance of personality testing. *Nature*, *348*, 671–672. doi: 10.1038/348671a0

Boring, E.G. (1919). Mathematical versus scientific significance. *Psychological Bulletin*, *16*, 335–338.

Bradburn, N.M. (2007). A tribute to Bill Kruskal. *Statistical Science*, *22*, 262–263.

Brunswik, E. (1952). The conceptual framework of psychology. In O. Neurath (Ed.), *International encyclopedia of unified science* (Vol. 1, pp. 655–760). Chicago, IL: University of Chicago Press.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley: University of California Press.

Campbell, J.P. (1982). Some remarks from the outgoing editor. *Journal of Applied Psychology, 67*, 691–700.

Carver, R.P. (1976). Letter to the editor. *Educational Psychologist*, *12*, 96–97.

Carver, R.P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*, 378–399.

Carver, R.P. (1993). The case against statistical significance testing, revisited. *The Journal of Experimental Education*, *61*, 287–292.

Cattell, R.B. (1978). *The scientific use of factor analysis in behavioral and life sciences*. New York, NY: Plenum.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *69*, 145–153.

Cohen, J. (1990). Things I have learned so far. *American Psychologist*, *12*, 1304–1312.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *12*, 997–1003.

Dawes, R.M. (1994). *House of cards: Psychology and psychotherapy built on myth*. New York, NY: Free Press.

Dawes, R.M. (2001). *Everyday irrationality: How pseudo-scientists, lunatics, and the rest of us systematically fail to think rationally*. Boulder, CO: Westview.

Dixon, P. (1998). Why scientists value *p* values. *Psychonomic Bulletin & Review*, *5*, 390–396.

Edgeworth, F.Y. (1886). Progressive means. *Journal of the Royal Statistical Society*, *49*, 469–475.

Edwards, W. (1965). Tactical note on the relation between scientific and statistical hypotheses. *Psychological Bulletin*, *63*, 400–402.

Edwards, W., Lindman, H., & Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Falk, R. (1998). In criticism of the null hypothesis statistical test. *American Psychologist*, *53*, 798–799.

Falk, R. (2008). Probabilistic reasoning is not logical. *Mathematics Magazine*, *81*, 268–275.

Falk, R., & Greenbaum, C.W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology*, *5*, 75–98.

Faust, D., & Ziskin, J. (1988, July 1). The expert witness in psychology and psychiatry. *Science, 241*(4861), 31–35.

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think. *Psychological Science*, *15*, 119–126.

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the "Journal of Applied Psychology": Little evidence of reform. *Educational and Psychological Measurement*, *61*, 181–210.

Fisher, R.A. (1925). *Statistical methods for research workers*. London, UK: Oliver & Boyd.

Fisher, R.A. (1929). The statistical method in psychical research. *Proceedings of the Society for Psychical Research*, *39*, 185–189.

Fisher, R.A. (1935). *The design of experiments*. Edinburgh, UK: Oliver & Boyd.

Fisher, R.A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society. Series B (Methodological)*, *17*, 69–78.

Fisher, R.A. (1956). *Statistical methods and scientific inference*. Edinburgh, UK: Oliver & Boyd.

Gerrig, R.J., & Zimbardo, P.G. (2002). *Psychology and life* (16th ed.). Boston, MA: Allyn & Bacon.

Gigerenzer, G. (1987). Probabilistic thinking and the fight against subjectivity. In L. Krüger, G. Gigerenzer, & M. Morgan (Eds.), *The probabilistic revolution: Vol. II. Ideas in the sciences* (pp. 11–33). Cambridge, MA: MIT Press.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311–339). Hillsdale, NJ: Erlbaum.

Gigerenzer, G. (2000). *Adaptive thinking: Rationality in the real world*. Oxford, UK: Oxford University Press.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*, 587–606.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Kruger, L. (1989). *The empire of chance*. Cambridge, UK: Cambridge University Press.

Gill, J. (1999). The insignificance of null hypothesis significance testing. *Political Research Quarterly, 52*, 647–674.

Granaas, M. (2002). Hypothesis testing in psychology: Throwing the baby out with the bathwater. *ICOTS 6 Proceedings*. Retrieved from http://www.stat.auckland.ac.nz/~iase/publications/1/3m1_gran.pdf

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

Grove, W.M., Zald, D.H., Lebow, B.S., Snitz, B.E., & Nelson, C. (2000). Clinical vs. mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*, 19–30.

Hagen, R.L. (1997). In praise of the null hypothesis test. *American Psychologist*, *52*, 15–24.

Hebb, D.O. (1966). *A textbook of psychology*. Philadelphia, PA: Saunders.

Holt, R.R. (1970). Yet another look at clinical and statistical prediction: Or, is clinical psychology worthwhile? *American Psychologist*, *25*, 337–349.

Hubbard, R. (2004). Alphabet soup: Blurring the distinctions between *p*'s and α's in psychological research. *Theory & Psychology, 14*, 295–327.

Hubbard, R., & Armstrong, J.S. (2006). Why we don’t really know what statistical significance means: Implications for educators. *Journal of Marketing Education*, *28*, 114–120.

Hubbard, R., & Lindsay, R.M. (2008). Why *p* values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology*, *18*, 69–88.

Hubbard, R., & Ryan, P.A. (2000).The historical growth of statistical significance testing in psychology—and its future prospects. *Educational and Psychological Measurement*, *60*, 661–681.

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

Jones, L.V., & Tukey, J.W. (2000). A sensible formulation of the significance test. *Psychological Methods*, *5*, 411–414.

Kirk, R.E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746–759.

Kirk, R.E. (2003). The importance of effect magnitude. In S.F. Davis (Ed.), *Handbook of research methods in experimental psychology* (pp. 83–105). Oxford, UK: Blackwell.

Korman, A.K. (1968). The prediction of managerial performance. *Personnel Psychology*, *21*, 295–322.

Lindsay, R.M. (1995). Reconsidering the status of tests of significance: An alternative criterion of adequacy. *Accounting, Organizations and Society*, *20*, 35–53.

Loftus, G.R. (1991). On the tyranny of hypothesis testing in the social sciences. *Contemporary Psychology*, *36*, 102–105.

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

Lykken, D.T. (1991). What’s wrong with psychology anyway? In D. Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology: Matters of public interest* (Vol. 1, pp. 3–39). Minneapolis: University of Minnesota Press.

Mahoney, M. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, *1*, 161–175.

Meehl, P.E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103–115.

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*, 103–115.

Meehl, P.E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*, 108–141.

Mulaik, S.A., Raju, N.S., & Harshman, R.A. (1997). There is a time and a place for significance testing. In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 65–115). London, UK: Erlbaum.

Murray, D., Schwartz, J., & Lichter, R. (2001). *It ain’t necessarily so: How the media make and unmake the scientific picture of reality*. Lanham, MD: Rowman & Littlefield.

Nunnally, J. (1960). The place of statistics in psychology. *Educational and Psychological Measurement*, *20*, 641–650.

Paulos, J.A. (2003). *A mathematician plays the stock market*. New York, NY: Basic Books.

Petrinovich, L. (1979). Probabilistic functionalism: A conception of research method. *American Psychologist*, *34*, 373–390.

Platt, J.R. (1964, October 16). Strong inference: Certain methods of scientific thinking may produce much more rapid progress than others. *Science*, *146*, 347–353.

Popper, K.R. (1959). *The logic of scientific discovery*. London, UK: Hutchinson. (Original work published 1935)

Rosenberg, A. (1988). *Philosophy of social science*. Boulder, CO: Westview.

Rosenthal, R. (1993). Cumulating evidence. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 519–559). Hillsdale, NJ: Erlbaum.

Rosnow, R.L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*, 1276–1284.

Rozeboom, W.W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*, 416–428.

Salmon, W.C. (1984). *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.

Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century*. New York, NY: Holt.

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin*, *66*, 178–200.

Schmidt, F. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist*, *47*, 1173–1181.

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

Schmidt, F., & 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. 37–64). London, UK: Erlbaum.

Schopenhauer, A. (2004). *Essays and aphorisms*. London, UK: Penguin. (Original work published 1851)

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*, 309–316.

Shermer, M. (2000). *How we believe: Science, skepticism and the search for God*. New York, NY: Holt.

Skinner, B.F. (1972). *Cumulative record: A selection of papers* (3rd ed.). New York, NY: Appleton-Century-Crofts.

Stigler, S. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Harvard University Press.

Tabachnick, B., & Fidell, L. (2001). *Computer-assisted research design and analysis*. Needham Heights, MA: Allyn & Bacon.

Taleb, N.N. (2007). *Black swan: The impact of the highly improbable*. New York, NY: Random House.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, *25*, 26–30.

Thompson, B. (1999). Statistical significance tests, effect size reporting and the vain pursuit of pseudo-objectivity. *Theory & Psychology*, *9*, 191–196.

Thompson, B. (2002). “Statistical,” “practical,” and “clinical”: How many kinds of significance do counselors need to consider? *Journal of Counseling and Development*, *80*, 64–71.

Thompson, B. (2003). Guidelines for authors reporting score reliability estimates. In B. Thompson (Ed.), *Score reliability: Contemporary thinking on reliability issues* (pp. 91–102). Thousand Oaks, CA: Sage.

Thompson, B. (2006). Research synthesis: Effect sizes. In J. Green, G. Camilli, & P.B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 583–603). Washington, DC: American Educational Research Association.

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools*, *44*, 423–432.

Thorngate, W. (1990). The economy of attention and the development of psychology. *Canadian Psychology*, *31*, 262–271.

Tukey, J.W. (1962). The future of data analysis. *Annals of Mathematical Statistics*, *33*, 1–67.

Tukey, J.W. (1991). The philosophy of multiple comparison. *Statistical Science*, *6*, 100–116.

Uttal, W.R. (2005). *Neural theories of mind: Why the mind–brain problem may never be solved*. Mahwah, NJ: Erlbaum.

Wilkinson, L., & The Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

Wright, R.H., & Cummings, N.A. (Eds.). (2003). *Destructive trends in mental health: The well-intentioned path to harm*. New York, NY: Taylor & Francis.

Yates, F. (1951).The influence of statistical methods for research workers on the development of the science of statistics. *Journal of the American Statistical Association*, *46*, 19–34.

Charles Lambdin is a human factors engineer at Intel. Address: Intel Corporation-Ronler Acres, 2501 Northwest 229th Avenue, Hillsboro, OR 97124-5506, Mailstop RA1-222, USA. Email: charles.g.lambdin@intel.com