Theory & Psychology

Article

Malignant side effects of null-hypothesis significance testing

Theory & Psychology
2014, Vol. 24(2) 256–277

DOI: 10.1177/0959354314525282
tap.sagepub.com
SAGE

Marc Branch
University of Florida, USA

Abstract

Six decades-worth of published information has shown irrefutably that null-hypothesis significance tests (NHSTs) provide no information about the reliability of research outcomes. Nevertheless, they are still the core of editorial decision-making in Psychology. Two reasons appear to contribute to the continuing practice. One, survey information suggests that a majority of psychological researchers incorrectly believe that *p* values provide information about reliability of results. Two, a position sometimes taken is that using them to make decisions has been essentially benign. The mistaken belief has been pointed out many times, so it is briefly covered because of the apparent persistence of the misunderstanding. The idea that NHSTs have been benign is challenged by seven "side-effects" that continue to retard effective development of psychological science. The article concludes with both a few suggestions about possible alternatives and a challenge to psychological researchers to develop new methods that actually assess the reliability of research findings.

Keywords

evolving science, generality, *p* values, reliability, replication

For over six decades irrefutable confirmations have been available showing that *p* values emanating from Null Hypothesis Significance Tests (NHSTs) do not provide probabilistic information about the reliability of research findings (e.g., Bakan, 1966; Carver, 1978; Cohen, 1994; Falk & Greenbaum, 1995; Gelman, Carlin, Stern, & Rubin, 1995; Gigerenzer, 1993; Greenwald, 1975; Kline, 2004; Lambdin, 2012; Nickerson, 2000; Oakes, 1986; Sohn, 1998; Thompson, 1999). Despite that, statistical significance remains, in most cases, a pre-requisite for publication in the psychological sciences. That is likely due to the fact that most practicing researchers in the psychological sciences

Corresponding author:
Marc Branch, Professor Emeritus, University of Florida, Box 112250, Gainesville, FL 32611, USA.
Email: branch@ufl.edu

believe, mistakenly, that *p* values do provide such information. The depth of the problem is illustrated by a survey conducted by Haller and Krauss (2002) using six questions developed by Oakes (1986). (See also, Kalinowski, Fidler, & Cumming, 2008; Mittag & Thompson, 2000.) Specifically, given a scenario in which a *t*-test has yielded $p < .01$, those surveyed were asked to state whether each of the following six statements was true or false: (a) you have disproved the null hypothesis (that there is no difference between the means); (b) you have found the probability (i.e., .01) of the null hypothesis being true; (c) you have proved your experimental hypothesis (that there is a difference between the population means); (d) you can deduce the probability of the experimental hypothesis being true; (e) you know, if you decide to reject the null hypothesis, the probability of making the wrong decision; and (f) you have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of the occasions. All of those statements are false. (For explanation of why, see below.) Haller and Krauss's (2002) survey included 30 academic psychologists who taught statistical methodology, and 80% of them got at least one of the items wrong! Among practicing psychologists in general over 90% of those queried got at least one wrong. It cannot be healthy for a field of investigation if its major criterion for publication is misunderstood by a majority of its practicing scientists.

Among some who understand that *p* values are not an indication of the probability of replication nor of the probability that results are due to chance or sampling error, a position is that at the very least their use is benign (e.g., Mulaik, Raju, & Harshman, 1997; Nickerson, 2000). In this paper, I argue that continued reliance on NHST has not been, nor does it continue to be, without harm. In this endeavor I echo the plea of Jones and Matloff (1986) who suggested that, "at its worst, the results of statistical hypothesis testing can be seriously misleading, and at its best it offers no informational advantage over its alternatives; in fact it offers less" (p. 1156). NHSTs are associated with what I shall term deleterious "side effects" that have contributed to the lack of cumulative, refined, and integrated development of psychological science. One of the side effects that has not been discussed previously is that their use leads to conflation of two separable, but related, subject matters (described in detail below), and, in so doing, has undermined the field's goal of understanding behavior or mind. In addition, acceptance of NHSTs has led to general neglect of the development of methods that actually assess the reliability of research results. (Although see Branch & Pennypacker, 2012; Johnston & Pennypacker, 2009; Sidman, 1960, for presentations of alternative approaches.)

Before discussing the side effects, however, it is necessary to review briefly the core reasons that *p* values do not provide quantitative information on whether research results will be replicable (that is, that the results are not a fluke), because of the apparent continuing widespread misunderstanding of what *p* values represent. They are, in fact, the probability of observing particular kinds of results given that the null hypothesis is true, that is, $P(\text{Data}|H_0)$; or more precisely, $P[T \geq t|H_0]$ where T is a value of a test statistic, and *t* is a criterion value). That is distinct from, and, more importantly, *unrelated to* the probability that the null hypothesis is true given the data, $P(H_0|\text{Data})$. That can be confirmed by taking any pair of conditional probabilities and reversing the conditionality. For example, $P(\text{Hanged}|\text{Dead}) \neq P(\text{Dead}|\text{Hanged})$ or $P(\text{Raining}|\text{Cloudy}) \neq P(\text{Cloudy}|\text{Raining}$; cf. Carver, 1978). An excellent example showing that P(positive mammogram|breast

cancer) ≠ P(breast cancer|positive mammogram) is provided by Gigerenzer, Gaissmeyer, Kurz-Milcke, Schwartz, and Woloshin (2008). (The reader is invited to try any reversible pair of conditional probabilities to confirm that the examples here are not unique.) As Falk and Greenbaum (1995) make clear, failure to understand that distinction leads to the illogic that too often accompanies NHSTs. As they note, the usual tactic is to believe that if a particular set of data is unlikely to occur if the null hypothesis is assumed, one can conclude that that is evidence that the null hypothesis is probably untrue. Or, stated more precisely, when the associated probability of *p* is smaller than the criterion adopted, chance is deemed unlikely to have produced the result obtained. The imprudence of that conclusion, even if the *p* value is very small, is made evident from the following application of exactly that logic. If the next person I meet is an American, it is very unlikely that it will be President Obama (P[Meet Obama|Meet an American] < .000000003). I just met Obama. Therefore it is unlikely I met an American (cf. Cohen, 1994; Falk & Greenbaum, 1995). This fallacy has been exposed for a long period. For example, Berkson (1942) noted,

> Consider [the argument] in syllogistic form. It says, "If A is true, B will happen *sometimes* [emphasis added]; therefore if B has been found to happen, A can be considered disproved." There is *no* [emphasis added] logical warrant for considering an event known to occur in a given hypothesis, *even if infrequently* [emphasis added], as disproving the hypothesis. (p. 326)

It is worth noting that this error does not depend on the characteristics of underlying distributions, random sampling, or other statistical issues. It is an error of logic. If a *p* value were $P(H_0|Data)$ then it would provide information about the truth of the null hypothesis. But, as the foregoing (and many other treatises) makes perfectly clear, that is not what a *p* value is. Knowing $P(Data|H_0)$ provides no information about $p(H_0|Data)$. This, of course, means that the phrase "statistically reliable" is a non-sequitur. The dispiriting fact is that most practicing psychologists apparently believe otherwise. We are left with a situation in which, even though a *p* value provides *no information* about the truth of the null hypothesis, it is used to make a judgment about its truth! As noted above, it has been argued that despite the fact that *p* values provide very little information, they are not necessarily completely without meaning (if, and only if, the null hypothesis is actually true, but see side-effect 4 below) and therefore do little harm (e.g., Nickerson, 2000), even though they dominate editorial review. In what follows, I present seven ways in which their use has been and continues to be deleterious. It should be noted at the outset that much of the information summarized below has been presented before. Nevertheless, presenting the problems *en masse* may help to initiate a reconsideration of the primary role NHSTs currently hold. Also, the list of side effects is not presumed to be exhaustive. For example, it does not include a discussion of how the misuse of statistical significance can compromise (and has compromised) the analysis of weighting parameters in multiple-regression analyses (see Ziliak & McCloskey, 2008). The seven are chosen because of the generality of their importance to the field of Psychological Science.

The main avenue by which the side effects work their negative effects into Psychology is via editorial practices. One of my goals in presenting this summary, therefore, is to convince editorial reviewers that reliance on statistical significance as any sort of

criterion for publication is a mistake. As Abelson (1997) noted, "Whatever else is done about null-hypothesis tests, let us stop viewing statistical analysis as a sanctification process … there are *no objective procedures* [emphasis added] that avoid human judgment and guarantee correct interpretations of results" (p. 13).

The side effects

Side effect 1: NHST promotes aimless, non-cumulative, non-integrated science

Meehl (1967, 1978, 1990) was the first, to my knowledge, to identify this issue. It arises from the common practice of pitting a null hypothesis, usually that there is no effect (what Cohen, 1994, dubbed the "nil" hypothesis), against an alternative hypothesis, which is usually what the scientist actually believes might be the case. NHSTs provide a *p* value, on the basis of which the scientist decides whether to argue that the null hypothesis is not the true state of affairs. If *p* is small enough, the null hypothesis is rejected, and the alternative hypothesis "gains support." It is not proven, of course, because there are infinitely many possible alternative hypotheses.

Meehl noted that *p* is determined by a test statistic, like a *t* or *F* value, that is computed by dividing variance attributable to the independent variable(s) under consideration by so-called error variance, variance attributable to other influences like measurement error or effects of uncontrolled variables (i.e., "chance"). That is,

$$\text{Test statistic} = \text{effect variance} / \text{error variance}.$$

As the value of the statistic increases, *p* decreases (as long as effect variance is not precisely zero). One of the goals of any scientific experiment is to minimize error, mainly by improving experimental methods. As methods improve, error variance is decreased, and therefore the value of the statistic is increased. That leads to a smaller *p*, and a greater likelihood that the null hypothesis will be rejected. Thus, better methods make it easier to meet the criterion for rejecting the null hypothesis, and thus to give support to the alternative hypothesis, *no matter what the alternative hypothesis is!* That is surely not a recipe for cumulative, integrative advancement in science, yet it has been the standard method of hypothesis evaluation in Psychology for more than 60 years.

This problem is exacerbated by the fact that in much actual practice deciding to reject the null hypothesis often might as well be based on the flip of a coin rather than collecting and analyzing any data at all (Cohen, 1994; Miller, 2009). If decisions between null and other hypotheses often approximate coin flipping, and it does not matter what the "alternative" hypothesis is, it is difficult to see how a science can advance (cf. Schmidt, 1996; Zakzanis, 1998, for evidence of the lack of cumulative, integrated knowledge in psychology).

Meehl suggested an alternative approach, one that has much in common with some methods of assessing goodness of fit of theoretical functions to data. Specifically, he recommended that instead of the null hypothesis being set at no difference or no effect,

the scientist's predicted effect should be set as the null. (There is no mandate that a null hypothesis has to be no effect or no relation.) In that case, if a result is determined to be statistically significant, what gets rejected is the prediction of the scientist's theory, a result likely to lead to theory modification, or perhaps even rejection. Thus, as experimental rigor and methods are improved, the predictions of a scientist's theory are subjected to an ever more rigorous test. This approach is clearly more in line with Popperian (Popper, 1959) falsification than is the standard, nil-hypothesis, approach of NHSTs.

This approach, however, arranges what might be an unfortunate contingency (J. Shepperd, personal communication, September 13, 2011). Specifically, if a scientist is invested in a theory, it is in the scientist's interest for a lack of statistical significance to be the result of the analysis, so that the theory's prediction is not rejected. Such a contingency might encourage less rigorous control with resulting larger error-variance values.

Meehl's suggested alternative approach does not avoid the fundamental logical problem outlined earlier (and another one to be described later). Nevertheless it is a step in a direction toward emphasizing the magnitudes of effects, which would help remedy the second side effect to be considered.

Side effect 2: NHST promotes "sizeless" science

As recently and exhaustively illustrated by Ziliak and McCloskey (2008), as typically employed NHSTs say nothing about the magnitudes of effects. They note that it is common, in reports of economics research, that no information whatsoever about effect magnitudes is presented. Regrettably, that is also true of much research in Psychology, although there seems to be a modest increasing trend in efforts to consider effect sizes in some sub disciplines. Given that any size of effect may be found to be statistically significant (see below, side-effect 4), ignoring the magnitudes of effects serves to retard the development of a science. Knowing the magnitude of any effect is essential to determining its likely importance, both practically and scientifically. For example, it might be of little interest to discover a variable that produces a 0.1% increase in respiration rate, but of great interest to find a 0.1% increase in the incidence of a fatal disease.

An important issue here is what is meant by effect size. There are statistical effect sizes, usually measured in units that vary from experiment to experiment, for example, mean differences in terms of variance or standard-error units (e.g., Cohen's *d;* Cohen, 1988). Although those measures are certainly an improvement over reporting whether a difference is statistically significant, they may not be sizes of effects of real significance. Effect sizes in a cumulative, effective science are directly measured in absolute units. Interest should be less in whether means differ by 0.5 or 3 units of standard error, and more in whether means are, for example, two correct math answers or ten, or a latency is 300 ms or 500 ms. That is not to say, of course, that the variability, either within subject or between subjects, is not important information, but I shall have much more to say about that later (Side effect 5).

Kline (2004) points out another aspect of the sizeless-science problem. When tests of low power are conducted (a common situation in Psychology; Cohen, 1962, 1990), and only statistically significant results merit publication, the effect sizes estimated from the published results are necessarily overestimates of the population effect sizes. Specifically,

to obtain a statistically significant result when there is low power the sample-effect size has to be larger than the population-effect size (see Kline's Table 3.2, 2004, p. 74, and Schmidt, 1996, for examples). This, of course, contaminates meta-analyses of the published literature. Schmidt (1996) noted this inaccurate estimation of effect sizes and argued that a cumulative science cannot develop from such data. Meta-analysis cannot solve the problem if only experiments that yield statistically significant results are analyzed. Thus, basing publication on statistical significance is likely to result in inaccurate estimates of population-effect sizes. That surely cannot be good for accurate, refined, cumulative knowledge.

Side effect 3: NHST blunts social processes that underlie successful science

In essence, science is the behavior of scientists, and one thing that sets science apart as a way of knowing from everyday approaches is that scientific knowledge is subject to empirical checking. The checking is what underlies the self-corrective characteristic of scientific knowledge. The main method of checking is via replication. If a result cannot be replicated (see cold fusion: Beaudette, 2002; Taubes, 1993), then it is placed on hold as something to be considered, or perhaps even ignored as failures to replicate mount. As noted earlier, statistical significance is silent with respect to whether a result should be considered replicable so when it is mistakenly thought to provide that information, claims of statistical significance replace actual replication.

The main point being made here, however, is that *p* values provide social safety for a scientist. For example, suppose a scientist conducts an experiment and obtains and reports a statistically significant result. Suppose in addition that sometime later someone else (or even the original scientist) repeats the experiment and does not replicate the statistical significance of the effect. Under the misunderstood rules of significance testing, the scientist who made the original report is "off the hook" with respect to being responsible for the error. That is, there are no negative social consequences of the erroneous initial report. The researcher can claim, "I played by the rules of significance testing that allow for a low probability of a Type I error; this must be one of those cases, so I bear no responsibility for the error." Contrast that with what prevailed before the advent of significance testing when a scientist's reputation rested to a significant extent on the reliability of what the researcher claimed. It is likely that such social pressure has a positive effect on science, in that when one's reputation rests on the reliability of what one reports, it will be less likely that unreliable results will be communicated. In essence, a scientist would perform and report on research that convinces the scientist her or himself that the results are "not a fluke." The goal would then be to convince reviewers that the results are reliable (ways to do that are discussed below).

Another important negative outcome is that requiring statistical significance be a prerequisite for publication makes difficult the publication of experiments that reveal the failure to replicate (the "File-drawer problem"). Given the pivotal role that replication plays in the evolution of scientific knowledge, it borders on unconscionable that failures to replicate be more difficult to publish than original reports.

Side effect 4: NHST is essentially a fool's errand

As noted by many accomplished statisticians, any sized difference can be found to be statistically significant (e.g., Hays, 1981; Meehl, 1978; Tukey, 1991) by increasing the number (*N*) of observations. Surprisingly, even this widely announced fact is apparently known by only about half of practicing researchers (Mittag & Thompson, 2000). One consequence of this fact is, as noted by Meehl, "As I believe is generally recognized by statisticians today and by thoughtful social scientists, the null hypothesis, taken literally, is always false" (1978, p. 822). The good news from this point of view is that the probability of a Type I error is essentially zero, so one need not worry about making one.

The fact that any point null can be rejected if *N* is large enough is another reason that Meehl's (1967) suggested remedy to the aimless-science problem is not fully satisfactory. If *N* is big enough, the point prediction of a scientist's theory will invariably be rejected. Here again, the sizeless-science problem rears its head. Whether or not a result is statistically significant is essentially a useless piece of information.

The fact that the probability that the null hypothesis is true is essentially zero undermines Nickerson's (2000; in the context of an outstanding and thorough review of the issues surrounding *p* values) main defense of the utility of *p* values. He shows (see his Table 3, p. 252) formally, based on Bayes Theorem, that if it is assumed that the prior probability of the null hypothesis and alternative hypothesis are *equal*, then a *p* value comes ever closer to the probability that the null hypothesis is true, given the data, as the true probability of the data given the alternative hypothesis approaches 1.0. The dubious assumption here is that $P(H_0)$ is the same as $P(H_A)$, which, even if true, leaves unexplained the infinite number of possible other values for the two probabilities. That notwithstanding, that the probability of the null hypothesis is essentially zero means that for Nickerson's analysis to have merit, the probability of the alternative hypothesis would have to be essentially zero, too. If they are both zero, then the entire approach becomes untenable because Bayes Theorem, from which Nickerson's calculations are derived, is indeterminate, with both the numerator and denominator approaching zero.

The reader may also have deduced another consequence of the fact that the null hypothesis is false, namely $P(Data|H_0)$ is *meaningless* because the "given" is not true. In the usual case, therefore, the *p* value is not only imprecise, it is invalid. It is difficult to justify using what is usually a meaningless number to make decisions about data. That is, expending scientific effort to answer the question, "Should I believe the null hypothesis true?" is a waste of time, time that could be more profitably spent developing methods and engaging in data analyses that actually get at the questions of reliability and magnitudes of results. Some of these are suggested later.

Finally, a little-discussed issue emanating from the fact that a point null hypothesis is always false, is the growing role of power analyses when NHSTs are employed. The adoption of such analyses has the admirable motivation to make tests sufficiently powerful to avoid problems associated with underpowered tests, such as those outlined by Cohen (1990). A standard approach is to determine the minimum *N* needed to result in a 20% or less chance of making a Type 2 error, given a specified effect size (often 0.5 standard-error units). The 20% criterion is what should be at issue. Why not a 1% chance,

or even less? The view that there are no Type 1 errors (i.e., the null hypothesis is always false) means that the only possible error is of the Type 2 sort.

Side effect 5: NHST promotes confusion of actuarial and behavioral science

This is the most subtle of the side effects, and to my knowledge has not been discussed before. Consequently, it is the one to which most attention will be directed in this review because consideration of it reveals not only a common confusion, but also points to ways of making psychology a cumulative and more effective science.

The popularity of NHSTs may rest to some degree on the fact that in most instances group means are compared. The means are from random samples from a population, and once a sample mean and its variance have been computed, inferences can be made about the population mean. What could be more general than something that applies to an entire population? That apparent generality is illusory, however, at least for a psychologist who is interested in understanding mind or behavior. As just noted, the mean from a sample provides an estimate of the mean of the entire population from which the sample is drawn, and that estimate can be bounded by confidence intervals that provide information—not the probability that another sample mean will fall in that interval (Cumming & Maillardet, 2006; Smithson, 2003)—as Nickerson (2000) states clearly, "A common misinterpretation of a confidence interval of *x%* around a statistic (e.g., sample mean) is that the probability is *x* that the parameter of interest (e.g., population mean) lies within the interval" (p. 279). To clarify, suppose 1000 iterations of a two-condition comparison were conducted yielding 1000 differences in means, and each provides a 95% confidence interval (CI). The population difference is not a random variable, so it is either within or outside any particular CI. A common misconception is that a single CI (e.g., the first one computed) is the range in which about 950 of the next 1000 sample-mean differences would fall. The correct interpretation is that the population mean would be captured by about 950 of the intervals. The sample mean, nevertheless, provides information about a parameter that applies to the entire population, so generality appears maximized. This raises two important issues.

First, there is the question of representativeness of the means with respect to the individual values that are used to compute the mean. Specifically, identical or similar means can result from substantially different distributions of scores. Two examples that illustrate this fact are given in Figures 1 and 2.

In Figure 1 (from Cleveland, 1994), four distributions of 20 scores are arrayed horizontally in the upper panel. The four plots in the lower panel show, with the top plot corresponding to the top distribution in the upper panel, and so on, the means (solid points) and standard deviations (bars) of the four distributions. The graphs show that identical means and standard deviations, the bases of most inferential statistics, can be obtained from very different distributions of values. This implies that when dealing with averages of measures, or averages across individuals, attention ought to be paid to the representativeness of the mean, not just its value and standard deviation (or standard error). Figure 2 contains what is known as Anscombe's Quartet (Anscombe, 1973), and it provides an even more dramatic illustration of how focusing only on the averages of

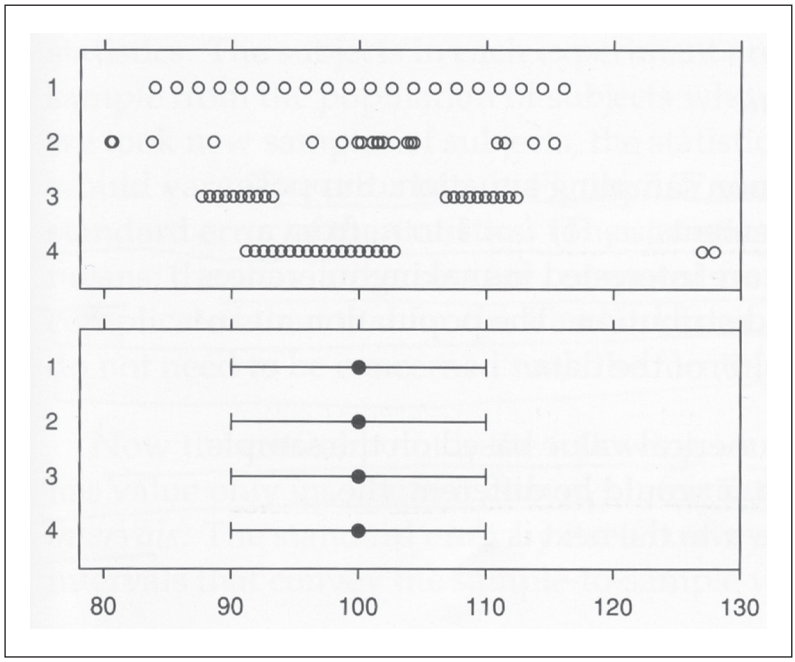

Figure 1. Upper panel: Four horizontal dot plots of 20 values. Lower panel: The corresponding means and standard deviations of the dot-plot distributions. (Reprinted from Cleveland, 1994, p. 215, with permission).

sets of numbers can lead one to miss important features of that set. The four graphs in Figure 2 show plots of 11 values in x/y coordinates, and also show the best fit (via the method of least squares) straight line. The distributions of points are quite different in the four sets. Yet the means for the x values are all the same, as are their standard deviations. The same is true for the y values (yielding eight instances of the sort shown in Figure 1). In addition, the slopes and intercepts of the straight lines are identical for all four sets, as are the sums of squared errors and sums of squared residuals, and all four yield the same correlation coefficient describing the relation between x and y. They are essentially identical in terms of common statistical analyses, but our eyes tell us otherwise.

The point of these illustrations is to indicate that a sample mean is *not* necessarily a good indicator of the generality across measures from *individuals*, which presumably is often the sort of generality in which a psychologist is actually interested. When the measures come from individual people (or other kinds of animals), it follows that the average from the group may not reveal, and may well conceal, much about individuals. Sample means from a group of individuals permit inferences about the population average, but these means do not permit inferences to individuals unless it is demonstrated that the mean is, in fact, representative of individuals. Surprisingly, it is rare in psychology to see the issue of representativeness of an average even mentioned, although recently, in the domain of randomized clinical trials in medicine, the limitations attendant to group averages have been gaining increased mention (e.g., Goodman, 1999; Kent & Hayward, 2007a, 2007b; Morgan & Morgan, 2001; Penston, 2005; Williams, 2010).

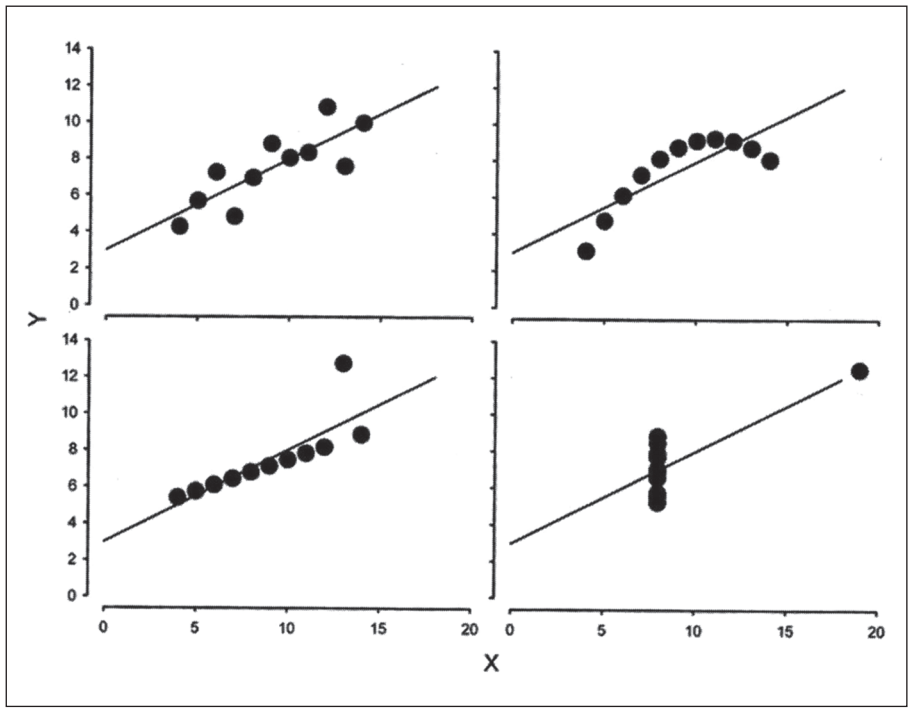

Figure 2. Anscombe's quartet. Each of the four graphs shows 11 x-y pairs and the best fitting straight line through the points. All standard statistical indicators, e.g., slope, intercept, distribution means and variances, etc. are identical for the four. (Reproduced from Anscombe, 1973, pp. 19-20, with permission).

The apparent generality promoted by group averages is an illusion because population means are, for lack of a better term, actuarial, not psychological, data. In psychology they are derived from behavioral data, but unless they are shown to be representative they can be quite misleading about generality across individuals. A confusion may arise here, occasioned by the use of the word "representativeness." In statistical analyses, a sample is representative of some population or subpopulation if it was randomly drawn from that population. That is not the meaning intended here where it refers to the degree to which a mean is depictive of individual measures summarized by the mean.

Therefore, widespread application of NHSTs has led to a field with two separable subject matters, behavior or mind and actuarial prediction (i.e., a science of population parameters). The argument here rests on the view that mind and behavior have meaning only at the level of an individual, not at the level of a group mean. It is meaningless (or at least boggles the imagination) to speak of group, or shared, mind. Your mind is yours, and it does not leak into anyone else's.

There are instances in which the difference between a population parameter, like the population average, and the activity of an individual is obvious. For example, suppose the average rate of pregnancy in women between 20 and 30 years of age is 5%. That, of course, is a useful statistic that can be used to predict how many women in that age category will be pregnant. More important for present purposes, however, is that the value,

5%, applies to *no* individual woman. That is, no woman is 5% pregnant. A woman is either pregnant or she isn't.

But what of situations in which an average is representative of the behavior of individuals? For example, suppose that a particular teaching technique results in a 10% increase in performance on an examination, and that the improvement is at or near 10% for every individual. Is that not a case in which a group average would permit estimation of a population mean that is, in fact, a good descriptor of the effect of the training for individuals, and because it applies to the population, has wide generality? The answer is yes and no.

It is "yes" because the representativeness (that is, the degree to which the average is a good description of what an individual will do) of the mean has been established, something that can be accomplished only by examining the data from the individuals. It may also be "no" for a more subtle reason that will be elaborated with an example. Consider a situation (modeled after one described by Sidman, 1960) in which a scientist is trying to determine the relation between amount of practice at solving 2-digit multiplication problems and subsequent speed of solving 3-digit problems in third-graders. Suppose that no practice, 10, 50, and 100 problems of practice, are to be compared. After the practice, children who have never previously solved 3-digit problems are given fifty 3-digit problems to solve, and the time-to-complete and accuracy are recorded. Because total practice might be a determinant of speed and accuracy, the scientist opts to use a between-groups design, with each group being exposed to one of the practice regimens. The hope is to extract the seemingly pure relation between amount of practice and later speed, uncontaminated by prior relevant practice. The scientist then averages the data from each group and uses those means to describe the function relating amount of practice to speed of solving the new, presumably more difficult, problems. In an actual case, there likely would be variability among individuals within each group, so a first issue would be how representative the average is of each member of each group. For our example, however, assume that the average is *perfectly* representative (i.e., every subject in a group gives exactly the same value). The scientist has generated a function, probably one that describes an increase in speed of correctly solving 3-digit multiplication problems as a function of amount of immediately prior practice. That function allows us to predict exactly what an individual would do if exposed to a certain amount of practice. Even though the means for each group are representative and therefore permit prediction about individual behavior, an important point is that the *function* has *no meaning* for an individual. That is, that function does not describe something that would occur for an individual because no individual can be exposed to different amounts of practice *for the first time.* The function is an actuarial account, not a description of a behavioral process. It is, of course, especially to the extent that the means are representative, a useful finding. It just is not descriptive of a behavioral/cognitive process in an individual. To examine the same issue at the level of an individual would require investigation of sequences of amounts of practice, and that examination would have to include experiments that factor in the role of repeated practice. Obviously, such an endeavor is considerably more complicated than the study that generated the actuarial curve, but it is the only way to develop a science of individual mind or behavior. The ontogenetic roots of mind or behavior cumulate over a lifetime.

The point here is not to diminish the value of actuarial, or population-parameter, data, nor to suggest that psychologists abandon the collection and analysis of such data. If means are highly representative, such data can offer predictions at the individual-subject level. Even if the means are not particularly representative, organizations like insurance companies and governments can and do make important use of such information in determining appropriate shared risk or regulatory policy, respectively. Using health policy as an example, even though the vast majority of people who smoke cigarettes do not get lung cancer, the incidence of lung cancer, on a relative basis, is substantially greater, *on average*, in that group. Because the group is large, even a low incidence rate yields a substantial number of actual lung-cancer cases, so it is in the government's, and the population's, interest to reduce the number of people who smoke cigarettes. Be that as it may, the point being made here is that in trying to establish a cumulative, evermore-accurate science of Psychology that can be effectively applied to individuals it will be important to distinguish mental/behavioral accounts from group-mean (actuarial) accounts. The two subject matters are related, but not the same. Some may argue that a science of actuarial effects is the best we can do because of the complexity of behavior, but subscribing to such an approach, an approach that is almost automatically the result of using NHSTs, guarantees that an advance from actuarial to individual prediction will be retarded.

The distinction being drawn between actuarial and psychological data should not be confused with the distinction between nomothetic and idiographic analyses (Allport, 1937), at least as usually conceptualized. Nomothetic processes are those that apply universally, and many such processes have been discovered using data obtained from individuals. Examples include developmental sequences (Piaget, 1928), classical-conditioning processes (Pavlov, 1927), operant-conditioning processes (Skinner, 1938), and aspects of memory (Ebbinghaus, 1885), and many perceptual processes, among others, so there is clear evidence that such an approach is not only possible, but can be highly fruitful and of considerable generality (Morgan & Morgan, 2001). Many applications of psychological knowledge, for example psychotherapy, involve individuals, so the field will be most effective practically if findings that apply to individuals serve as the basis for those kinds of applications. There are presumably also nomothetic processes that apply to actuarial data, but they refer to group-average effects, not necessarily to effects at the individual-subject level.

How is generality identified if group averages are not the focus? Easy; each studied participant in a research project is treated as a separate experiment, that is, an attempt at replication (more on this below). The consistency of effects across individuals in each condition, and of differences between individuals in differing conditions of a study, obviously provides direct information on reliability of effects that NHSTs cannot.

Side effect 6: NHST impedes the publication of "negative" results

Although this side effect is related to side effect 3, it is serious enough to merit its own treatment. Sometimes this problem is described as biasing the literature. If only effects that reach some criterion of statistical significance are published, much important research may go unreported. This issue, of course, is related to the NHST's negative

effects on publishing failures to replicate. But it can be a deeper problem. There are instances in which so-called negative effects can be vitally important to the growth of a science. A classic example is the famous Michelson–Morley experiment (Michelson & Morley, 1887). The experiment was conducted to validate the existence of the "Luminiferous ether." Because light had been shown to have wave properties, it was assumed that the waves needed a medium, just as water and air provide a medium for wave travel, and thus the idea of the luminiferous ether was that there exists in the universe this medium for light waves. It was reasoned therefore that the speed of light, because it is a wave, would differ depending on whether the light was traveling "upstream" or "downstream" with respect to the ether. It was also known that the earth is traveling at a high speed relative to space, so light traveling in the direction the earth is moving would be going upstream as the earth pierced the ether, and downstream in the other direction. The main result of the Michelson–Morley experiment is that the speed of light was the same no matter what direction it was traveling, that is, a null hypothesis that there is no difference in light speed could not be rejected. It turns out this is one of the most important discoveries in the history of physics. The fact that the speed of light is independent of its inertial frame is the foundation of special relativity theory. Any data-analysis method that makes difficult the possibility of reporting so-called negative results is not good for a science.

Some traditionalists might argue that it is a violation of Fisherian logic to publish a result that is not a statistically significant effect, because by that logic the null hypothesis cannot be accepted, only rejected. Therefore, not publishing a failure to achieve statistical significance is justified. That traditionalist must answer the question, then, how does one get such a failure published?

If statistical significance is not the criterion, what replaces it? A good, and non-frightening, answer is simple: expert, informed judgment. Failures to replicate and reports of "no-effect" need to be evaluated for experimental adequacy. How faithful is the replication attempt with respect to conditions of the original experiment. Are there any confounds? Is the experiment sufficiently rigorous to reduce variability? Does the attempt to replicate include internal replications? That is, exactly the same considerations that should go into evaluating any study should be applied. Some will balk. By what rules do we decide? The answer is also not frightening; it is that there are no fixed, generally applicable rules. Relying on expert judgment and experience served science well for centuries, and, as noted above, brings the appropriate social influences to bear. There is no reason that it cannot be so again for Psychology. Given that significance testing has provided no real advantages, and has yielded several important disadvantages (the "side effects"), a return to methods that rely on demonstrations of rigor and replication surely would not retard the development of psychological knowledge. Slavish adherence to a fixed, but deeply flawed, set of rules and conventions can only continue the problems.

Side effect 7: NHST inhibits the range of experimentation

This last side effect seems modest in scope, but it is insidious. In standard practice, NHSTs require that experiments be based on more or less formal hypotheses. That is too restrictive. There are many very good reasons to conduct experiments other than to test

hypotheses. Among them are examining the boundary conditions for a phenomenon, developing a new method, seeing if a phenomenon exists, characterizing the parameters of a phenomenon (e.g., determining if the relation between an independent and dependent variable is linear or logarithmic, information that could be crucial for theory generation), as well as testing hypotheses. Shoe-horning all experiments into a hypothesis-testing framework can limit the range of experimentation and therefore retard the advancement of science. Of course, some pay only lip service to the requirement for a hypothesis. Most have read papers that make it clear that the hypotheses were generated after the results had been obtained. Nevertheless, even hypotheses invented *post hoc* may serve to constrain thinking about the ramifications of research results.

It is not uncommon to see a grant application criticized because of a lack of hypotheses. That is, there is a trained generation of scientists who think that hypotheses are essential to good experimentation. It is difficult to believe, given that the history of great science is filled with experiments that were not based on hypotheses, that training in NHST has not contributed to the narrow view.

Some of the most important experiments in science have come from the "I wonder what would happen if …" approach. Some have even christened this approach "curiosity driven" science to contrast it with "hypothesis driven" science (e.g., Committee on CMMP 2010, 2007, p. 51). Reasons to conduct useful research therefore abound. It is well to remember the dictum, "*Hypotheses non fingo*" (I hold no hypotheses). That was the advice of Isaac Newton, perhaps anticipating research on experimenter bias (Rosenthal, 1966; Rosenthal & Fode, 1963). Most would agree that Newton was a scientist with more than a modicum of success.

Some recommendations for change

What can be done to rectify the many problems associated with NHSTs? Some have been mentioned in the foregoing, but the key locus for effecting changes lies in journal and grant reviewing. Peer review is one of the most important functions in science, so it certainly should not be grounded in a misconception. Editors could (and should) make it clear to reviewers that results need not be analyzed by conventional NHST methods to merit publication. Clear and tellingly tragic examples of the requirement of statistical significance exist in the health-care literature, where indications of lethal side effects of drugs, indications that did not reach statistical significance, have been ignored (exhibiting an all too common practice of equating lack of statistical significance with lack of effect; see for example, Nieuwenhuis, Forstmann, & Wagenmakers, 2011) only later to have been discovered to be real effects that resulted in numerous deaths (Ziliak & McCloskey, 2008). A fundamental issue to which reviewers should therefore direct initial attention is whether there is direct evidence that the effects seen are reliable. As noted by Thompson (1996) "If science is the business of discovering replicable effects, because statistical significance tests do not evaluate result replicability, then researchers should use and report strategies that *do* evaluate the replicability of the results" (p. 29). How compelling those indicators are, of course, will be influenced by what reviewers see as the rigor, importance, novelty, etc., of the research, as well as the sizes and consistency (across individuals) of effects. Nevertheless, it can be judged if the methods are suitably

rigorous and effects are of sufficient absolute, or statistical, magnitude to be worthy of additional consideration. It is completely clear that *p* values from NHSTs provide *no* information on these counts that cannot be determined from the data themselves, so reviewers will be asked to render expert judgment about whether the data provided indicate that the results are likely reliable (either at the individual or group-mean level, depending on whether research is cognitive/behavioral or actuarial), or worth getting into the literature so that replications will be attempted. What the requisite evidence will be will presumably vary depending on the phenomenon under study, thus making it especially important that reviewers have expertise in the relevant research domain. If significance tests were not used, researchers could report results that they consider important, and no one would automatically assume, on the basis of flawed logic, that the result had some probability of being reliable. That might very well lead, if others were also interested, in attempts to replicate the effect, thus achieving what significance testing cannot, an assessment of replicability. Focusing directly on reliability, therefore, would lead to more attempts to assess it within studies *and* across studies. It is hard to see how that could be bad for science.

In some cases, such as within-subject baseline-reversal (aka ABA) designs, replications are built in (see Branch & Pennypacker, 2012, for examples). In others, where a range of values of an independent variable is examined, orderliness of the relationship between independent and dependent variables provides evidence about the likelihood of successful replications. In most cases, nevertheless, expert judgment about the quality of the research, distributional characteristics of the data, and other factors will need to be weighed. Statistical significance can be thought of as a crutch used by editors and reviewers, but as the foregoing indicates, that crutch is a sham with respect to identifying the so-called reality of effects. The apparent, also illusory, objectivity of NHST (Berger & Berry, 1988) is outweighed by the misdirection that results from its use.

A second avenue should come from those who write introductory texts that are used to teach about NHSTs. Berger and Selke (1987) noted over two decades ago that "we know of no elementary textbooks that teach that $p = .05$ is *at best* [emphasis added] very weak evidence against H_0" (p. 114). It is regrettable that this is still the case (although some authors are getting closer to revealing the illogic, e.g., Motulsky, 2010).

What of value is likely to be lost if reviewers focus on reliability rather than statistical significance as an initial criterion? Most likely, nothing. Will there be an increase in publication of unreplicable results? Almost certainly not. Cohen (1990, 1994), Hung, O'Neill, Bauer, and Kohne, (1997), and Ioannidis (2005) have shown very high rates of publication of unreplicable work. It is hard to imagine that expert, informed judgment could do worse.

Some might argue that at least NHSTs provide for an objective approach, and, therefore, any replacement needs to be equally objective. It might be "objective" and "standardized," but is the production of a usually meaningless number, *p*, worth the costs? Some might argue that any replacement must include use of fixed rules or practices based on measures like effect sizes, but that too is unworkable. What constitutes an important effect size is going to depend on what the effect is. There is no practicable, cookie-cutter approach to deciding about data. Informed judgment is required.

The lack of replicability of statistically significant effects is increasingly being recognized as a clear negative effect of NHSTs in medical research (Ioannidis, 2005). A shift to a focus on evaluating the quality of an experiment, and on direct evidence of reliability, would likely result in *less* unreplicable work being published, not more.

What is also likely to be gained is a re-invigoration of psychological science to its stated primary goal, the understanding of mind and behavior of individuals, not being limited to the characteristics of population-level phenomena. If NHSTs are abandoned, or at least de-emphasized, as a gateway to publication, emphasis will likely be given to developing research designs and data-analysis methods that examine reliability directly.

Other useful methods to assess reliability have been suggested by Thompson (1993, 1994) and Loftus (1996), and they involve examining aspects of the data set. An example of that can be provided by performing some thought experiments with the data shown in Figure 1. Suppose that 10 is added to each value in the top row, and then the two sets of values are compared via a *t*-test. Ten is the standard deviation of the original distribution, so Cohen's *d* is 1.0, a so-called large effect. The resulting *p* value is less than .005, so the difference is statistically significant by most standards. What can be concluded, however, about how replicable the result is at the level of the individual? One way to do that is make all the possible individual comparisons. There are 20 scores in each of the two distributions, yielding 400 possible individual comparisons. For our example, the comparison will be simply of which of the two scores is larger. The result is that for 235 of the 400 comparisons a score taken from the second distribution is larger than a score from the original distribution. That is, at the individual level, the direction of the mean result is replicated 59% of the time. For that sample, therefore, 59% is a direct estimate of replicability of direction of effect at the individual level. (To get 100% replicability, about 30 would have to be added to each score.) Additional research on whatever topic generated the data could then use that number (59%) for comparison. Other comparisons can also be made. For example, the number of comparisons that meet or exceed a difference of 10 (the mean difference) could be computed relatively easily. That would provide an estimate of the representativeness of the mean difference. Note that in this example all the features of the distributions are identical, just displaced by one standard deviation.

To illustrate what can occur as a result of differences in distributions, consider what happens if we add 10 to each score in the bottom distribution and compare it to the top one. Again, all the standard statistical data are *exactly* the same: a mean difference of 1 SD, and $p < .005$. In this case, however, the 400 individual comparisons reveal that a score taken from the bottom (remember, shifted +10 to the right) distribution is larger than one from the top distribution 173/400 times. Here, individual comparisons show that the mean result is replicated only 43% of the time! That is, at the individual level, you are more likely *not* to replicate the mean effect than you are to reproduce it, despite the so-called large mean effect of 1 SD. Surely this kind of information is more likely to lead to accurate predictions at the individual level than are reports of means and statistical significance.

A practical issue arises when experiments involve many studied participants. Even though the goal is to understand generality across individuals, it is frequently impractical to show data from every participant studied. That is not as large a problem as it may seem. There are excellent methods for displaying and comparing data from distributions

in a minimum of space. For example, Tukey's (1977) stem-and-leaf plots and box-and-whisker plots are good examples. With some aggregation, quantile–quantile plots (e.g., Cleveland, 1994) can provide for illuminating comparisons. In addition, showing data from representative individuals is a time-honored tradition in the biological sciences, of which psychology is one. (Some might object that this could lead to so-called cherry picking, but that can be minimized by provision of criteria for selection.)

Another thorny, practical issue is that for many investigations in psychology it is difficult, if not impossible, to decide whether an effect size is worth pursuing. That occurs because the dependent variable has no fixed standards against which to judge it. Common among these kinds of measures are scores from rating scales or psychological tests, both of which are ubiquitous in psychology. For example, without NHSTs how is one to judge whether a rating of 5.2 on a 7-point Likert scale is *scientifically* or practically significantly different from 4.8? There are several approaches that can be used, depending on from where the two ratings came. Determining reliability is straightforward. If they are from two different items on a set of ratings, or from the same item on multiple occasions (e.g., before and after a treatment) then the first issue would be how representative the scores, and the difference between them, are for the several to many studied participants in the research, both with respect to absolute value and direction of the difference. Note that both of these indicators of reliability refer to activity at the level of the individual. If the ratings are from different groups, standard methods of comparing the distributions (e.g., like those developed by Tukey, 1977) can be employed to examine reliability at the group level. Once the reliability of the effect has been assessed and deemed convincing enough for further evaluation the problem gets more difficult, however. The meaning of the ratings needs to be assessed in some manner. For some psychological tests, a degree of meaning has been determined empirically. For example, certain scores on the Beck Depression Inventory have been related to the likelihood of attempted suicide (Brown, Beck, Steer, & Grisham, 2000; although only at the group-mean level). Suicide attempts are a countable entity, so the score can be related to (at least on average) countable episodes. For rating-scale data to be interpreted scientifically, they need to be related to real psychological outcomes (Baumeister, Vohs, & Funder, 2007). That is, they have to have size.

Final observations

A final argument for the retention of NHST is that it provides for uniformity of communication across experiments, research domains, and disciplines. That would be a logical defense if it were not for the fact that ordinarily what gets communicated is *misinformation*, most commonly about presumed reliability of findings. But, as has been made abundantly clear many times, *p* values do not provide information about reliability. That fact, coupled with the side effects summarized in this paper, argues strongly against the idea that a common language, based on NHSTs, provides a benefit to science. Science should be based on accurate information, not misinformation or what is at best relatively useless information.

Psychology as a scientific discipline can be seen as wallowing, perhaps slowly disintegrating. The American Psychological Association currently has 59 divisions, most of

which are completely independent of one another scientifically. They share no core of knowledge (except, ironically, how to employ NHSTs), the kind of knowledge that is generated by cumulative, evolving science. The typical introductory text has about 20–25 chapters, each of which can be read pretty much independently of any of the others. The order of topics, which is dictated more by tradition than logic (if the organization mimicked that of other, more mature and integrated sciences, it is likely that the basic psychological phenomena would appear first, and the reductionistic analyses of them would occur later), does not generally reflect any accumulation, refinement, or integration of knowledge. Instead, a student comes away with the view that there are many interesting things that psychologists study, but that they are pretty much unrelated. In my own department, which I believe to be typical in its training of students, there is not a single, substantive psychological fact or set of facts that every graduate student must know. A student can complete our graduate program without learning anything at all about basic learning processes, or basic sensory and perceptual processes, or memorial and other cognitive processes, or developmental processes, or social processes, or approaches to personality, and so on. Students, as in most graduate programs, can pick and choose among a few courses on those (and other) topics to provide them presumed breadth. But the only training *every* student must have is in NHST. I sometimes like to say, only partly in jest, that current graduate training in Psychology emphasizes learning a set of methods from which no basic facts (that is, facts that every psychologist should know) have emerged!

I am arguing that this state of affairs has developed because of the reliance on NHSTs as the dominant method for analyzing data and for deciding if results merit publication, thus retarding the development of cumulative, evolving, integrated knowledge. NHSTs have assumed this role in research to a large degree because their results are misunderstood by a majority of practicing psychologists, who mistakenly presume that statistical significance provides a quantitatively precise estimate and therefore protection against error. NHST simply does not do that, and, as emphasized in this treatise, it has led to practices that have retarded the development of the field of Psychology. For Psychology to progress, NHSTs will have to be de-emphasized and replaced by methods for assessing reliability and significance directly. As noted earlier, that has to start with editorial practices. Those who appreciate the negative impact of NHSTs and who also ascend to editorships can exercise top-down influence by urging their cadres of reviewers to de-emphasize NHSTs. Those who serve as reviewers can exert bottom-up, so-called grassroots, influence by suggesting that authors provide direct information about reliability, for example by illustrating the representativeness of group-averages. Combination bar and dot plots (replacing the ubiquitous standard error of the mean, which provides information about the population mean) are an excellent first step in providing evidence of reliability across subjects.

Using statistical significance as a pre-requisite for publication is simply a scientifically destructive ritual. It is time to move toward evidence-based methods, given that the evidence about the scientific irrelevance, and counter-productiveness, of NHSTs is clear, even to the lay press (e.g., Siegfried, 2010). Once editorial approaches are altered, it will likely be easier to implement instruction in data analysis to emphasize use of and development of methods that illuminate reliability, representativeness at the level of the

individual, and absolute magnitudes of effects. Failing to do so will permit the current trajectory of the disintegration of the field to continue.

Acknowledgements

Thanks to Jesse Dallery, Lawrence Kupper, James Shepperd, and Clive Wynne for helpful comments.

Funding

Preparation of this paper was supported by USPHS Grant No. DA004074 from the National Institute on Drug Abuse.

References

Abelson, R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science*, *8*, 12–15.

Allport, G. W. (1937). *Personality: A psychological interpretation*. New York, NY: Holt, Rinehart, & Winston.

Anscombe, F. J. (1973). Graphs in statistical analysis. *American Statistician*, *27*, 17–21.

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

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, *2*, 396–403.

Beaudette, C. G. (2002). *Excess heat: Why cold fusion research prevailed*. South Bristol, ME: Oak Grove Press.

Berger, J. O., & Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist*, *76*, 159–165.

Berger, J. O., & Selke, T. (1987). Testing a point null hypothesis: The irreconcilability of P values and evidence. *Journal of the American Statistical Association*, *82*, 112–122.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association*, *37*, 325–335.

Branch, M. N., & Pennypacker, H. S. (2012). *Generality and generalization of research findings*. In G. J. Madden (Ed.), *APA handbook of behavior analysis* (Vol. 1, pp. 151–175). Washington, DC: American Psychological Association.

Brown, G. K., Beck, A. T., Steer, R. A., & Grisham, J. R. (2000). Risk factors for suicide in psychiatric outpatients: A 20-year prospective study. *Journal of Consulting and Clinical Psychology*, *68*, 371–377.

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

Cleveland, W. S. (1994). *The elements of graphing data*. Murray Hill, NJ: AT&T Bell Laboratories.

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

Cohen, J. (1988). *Statistical power analysis for the behavior sciences*. Hillsdale, NJ: Erlbaum.

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

Cohen, J. (1994). The world is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Committee on CMMP 2010, *Solid State Sciences Committee, National Research Council. (2007)*. *Condensed-matter and materials physics: The science of the world around us*. Washington, DC: National Academies Press.

Cumming, G., & Maillardet, R. (2006). Confidence intervals and replication: Where will the next mean fall? *Psychological Methods*, *11*, 217–227.

Ebbinghaus, H. (1885). *Über das Gedächtnis* [On memory]. Leipzig, Germany: Verlag von Duncker & Humber.

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

Gelman, A., Carlin, J., Stern, H., & Rubin, D. B. (1995). *Bayesian data analysis*. London, UK: Chapman & Hall.

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., Gaissmeyer, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2008). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*, 53–96.

Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The *p* value fallacy. *Annals of Internal Medicine*, *130*, 995–1004.

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

Haller, S., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research*, *7*, 1–20.

Hays, W. L. (1981). *Statistics* (3rd ed.). New York, NY: Holt, Rinehart, & Winston.

Hung, H. M., O'Neill, R. T., Bauer, P., & Kohne, K. (1997). The behavior of the P-value when the alternative hypothesis is true. *Biometrics*, *53*, 11–22.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, 0696–0701.

Johnston, J. M., & Pennypacker, H. S. (2009). *Strategies and tactics of human behavioral research* (3rd ed.). New York, NY: Routledge.

Jones, D., & Matloff, N. (1986). Statistical hypothesis testing in biology: A contradiction in terms. *Journal of Economic Entomology*, *79*, 1156–1160.

Kalinowski, P., Fidler, F., & Cumming, G. (2008). Overcoming the inverse probability fallacy: A comparison of two teaching interventions. *Experimental Psychology*, *4*, 152–158.

Kent, D. M., & Hayward, R. A. (2007a). Limitations of applying summary results of clinical trials to individual patients: The need for risk stratification. *Journal of the American Medical Association*, *298*, 1209–1212.

Kent, D., & Hayward, R. (2007b). When averages hide individual differences in clinical trials: Analyzing the results of clinical trials to expose individual patients' risks might help doctors make better treatment decisions. *American Scientist*, *95*, 1016–1019.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: APA Books.

Lambdin, C. (2012). Significance tests as sorcery: Science is empirical—significance tests are not. *Theory & Psychology*, *22*, 67–90. doi: 10.1177/0959354311429854

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, *5*, 161–171.

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*, 806–834.

Meehl P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*, 195–244.

Michelson, A. A., & Morley, E. W. (1887). On the relative motion of the earth and the luminiferous ether. *American Journal of Science*, *34*, 333–345.

Miller, J. (2009). What is the probability of replicating a statistically significant effect? *Psychonomic Bulletin & Review*, *16*, 617–640.

Mittag, K. C., & Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher*, *29*, 14–20.

Morgan, D. L., & Morgan, R. K. (2001). Single-participant research design: Bringing science to managed care. *American Psychologist*, *56*, 119–127.

Motulsky, H. (2010). *Intuitive biostatistics: A nonmathematical guide to statistical thinking*. New York, NY: Oxford University Press.

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). *There is a time and place for significance testing*. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 65–116). Hillsdale, NJ: Erlbaum.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241–301.

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, *14*, 1105–1107.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York, NY: Wiley.

Pavlov, I. P. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex* (G. V. Anrep Ed. & Trans.). Oxford, UK: Oxford University Press.

Penston, J. (2005). Large-scale randomized trials: A misguided approach to clinical research. *Medical Hypotheses*, *64*, 651–657.

Piaget, J. (1928). *The child's conception of the world*. London, UK: Routledge and Kegan Paul.

Popper, K. (1959). *The logic of scientific discovery*. New York, NY: Basic Books.

Rosenthal, R. (1966). *Experimenter effects in behavioral research*. East Norwalk, CT: Appleton-Century-Crofts.

Rosenthal, R., & Fode, K. L. (1963). The effect of experimenter bias on the performance of the albino rat. *Behavioral Science*, *8*, 183–189.

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

Sidman, M. (1960). *Tactics of scientific research*. New York, NY: Basic Books.

Siegfried, T. (2010, March). Odds are, it's wrong: Science fails to face the shortcomings of statistics. *Science News*, *177*, 26–37. Retrieved from http://www.sciencenews.org/view/feature/id/57091

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. New York, NY: Appleton-Century-Crofts.

Smithson, M. (2003). *Confidence intervals*. London, UK: Sage.

Sohn, D. (1998). Statistical significance and replicability: Why the former does not presage the latter. *Theory & Psychology*, *8*, 291–311. doi: 10.1177/0959354398083001

Taubes, G. (1993). *Bad science: The short life and weird times of cold fusion*. New York, NY: Random House.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education*, *61*, 361–377.

Thompson, B. (1994). The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results. *Journal of Personality*, *62*, 157–176.

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. doi: 10.1177/095935439992007

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

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

Williams, B. A. (2010). Perils of evidence-based medicine. *Perspectives in Biology and Medicine*, *53*, 106–120.

Zakzanis, K. K. (1998). Brain is related to behavior ($p < .05$). *Journal of Clinical and Experimental Neuropsychology*, *20*, 419–427.

Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Ann Arbor: University of Michigan Press.

Author biography

Marc Branch is Professor Emeritus of Psychology at the University of Florida where his main research interests have been in behavioral pharmacology and learning and performance in animals. He has additional interest in scientific methods as exemplified by the chapter, M. N. Branch and H. S. Pennypacker (2012) Generality and Generalization of Research Findings (pp. 151–175), in G. J. Madden (Ed.), *APA Handbook of Behavior Analysis, Volume 1*. Address: Marc Branch, Professor Emeritus, University of Florida, Psychology, Box 112250, Gainesville, FL 32611, USA. Email: branch@ufl.edu