Special Section

NEEDED: A BAN ON THE SIGNIFICANCE TEST

John E. Hunter
Michigan State University

Abstract—*The significance test as currently used is a disaster. Whereas most researchers falsely believe that the significance test has an error rate of 5%, empirical studies show the average error rate across psychology is 60%—12 times higher than researchers think it to be. The error rate for inference using the significance test is greater than the error rate using a coin toss to replace the empirical study. The significance test has devastated the research review process. Comprehensive reviews cite conflicting results on almost every issue. Yet quantitatively accurate review of the same results shows that the apparent conflicts stem almost entirely from the high error rate for the significance test. If 60% of studies falsely interpret their primary results, then reviewers who base their reviews on the interpreted study "findings" will have a 100% error rate in concluding that there is conflict between study results.*

Consider a parable: A new young science has addressed significant problems and has attracted a large number of bright, tireless, empirical scientists. There is only one problem. This new science uses a defective decision-making technique that has been shown to have a 60% error rate. Worse yet, people in that science falsely believe that the error rate is 5%. Thus, people take the results of the decision-making device to be almost always right even though it is actually wrong more often than it is right.

The question is this: Do you expect a high or a low rate of progress in this new science? Before you answer, consider this ugly fact. Suppose two researchers in this new science are testing a new hypothesis. One does an empirical study and the other flips a coin. The person who does the study has a 60% chance of error and thus has a 40% probability of being right. The person who flips a coin has a 50% probability of being right. That is, the person who flips a coin will be right more often than the person who does a study.

My personal prediction is that progress in this new science will come at a glacial rate.

THE FACT

The new science in this parable is psychology, and the defective decision-making technique is the conventional significance test as it is currently used. That is, the decision-making technique that dominates today's journals has a 60% error rate.

The significance test as it is currently used in the social sciences just does not work. The significance test is a disaster and has been from the beginning. Whereas the typical researcher falsely believes that the significance test has an error rate of 5%, empirical studies show the average error rate across the field of psychology to be 60% (Sedlmeier & Gigerenzer, 1989). That is, the error rate for the significance test is 12 times higher than researchers think it to be. The error rate for inference using the significance test is greater than the error rate using a coin toss to replace the empirical study.

The current way of using the significance test has been a disaster for the research review process in psychology. Almost every comprehensive review in psychology cites conflicting results. Yet quantitatively accurate review of research results shows that the apparent "conflict" in results stems almost entirely from the high error rate for the significance test. If key results in 60% of studies are interpreted falsely, then reviewers who base their reviews on the study "findings" as interpreted using the significance test will have a 100% error rate in forming opinions as to the level of agreement between study results.

Frank Schmidt, Robert Rodgers, and I have documented case after case in which the use of the significance test has caused false reviews of the literature in industrial psychology, in organizational psychology, and in management. Similar documentation can be found in almost every meta-analysis now published, though few authors make this point in reporting their meta-analyses.

An otherwise well done review that falsely cites "conflicting findings" can delay progress in an area of research for decades. In personnel selection, the use of the significance test has caused a 50-year delay in progress in some research areas!

REACTIONS

Most psychologists initially reject my argument as "too radical." They cannot believe that there can be a fatal flaw in any technique that is so widely used now and has been so widely used for 50 years. That is, most scientists know that any technique with a 60% error rate must be abandoned, and so most current psychologists think that there must be some flaw in my argument.

The reactions to the statement that the significance test has a 60% error rate fall into four main categories:

- "This is no surprise. Everyone knows that far more than 60% of studies done in psychology are garbage studies with major methodological errors. However, my studies are methodologically top notch and so my studies do not make errors."
- "A 60% error rate is impossible; the error rate is only 5%."
- "Low power can't be that bad! It's true that when the treatment has an effect, a study may have less than 95% power to detect the effect. But the significance test does have only a 5% error if the

Address correspondence to John E. Hunter, Department of Psychology, 133 Snyder Hall, Michigan State University, East Lansing, MI 48824; e-mail: 06991jeh@msu.edu.

null hypothesis is true, and in fact the null hypothesis is almost always true. So the error rate cannot possibly be as high as 60%."

- "You're right, but there is nothing we can do about it. I know that the significance test has a 60% error rate, but I'm not going to tell an editor that. I include significance tests in all my papers because they'll get rejected if I don't."

Let me respond to each reaction in turn.

THE MYTH OF THE GARBAGE STUDY

Many laboratory researchers respond to my argument by saying that the significance test will have a high error rate only if the study is methodologically flawed. If the study is methodologically sound, it cannot have an error in results, and therefore the significance test will always come out right.

The question is, why are there conflicting results in the literature? Their answer is that many studies are "garbage studies," studies with fundamental design flaws. The key to knowledge is to determine which studies are garbage studies and which studies are good studies.

Are there a lot of garbage studies being conducted? Results from meta-analysis typically show that the differences in results across studies are largely explained by sampling error. Methodological differences—such as differences in the quality of measurement—typically explain only a small portion of the variation in results. Furthermore, no one has found a case in which differences in methodological quality created qualitative differences between study results. I have read thousands of studies across a wide range of psychological domains, and I have found very few garbage studies. The concept of "garbage studies" can be largely traced to hubris—my dog is better than your dog.

To be blunt, I have served on hundreds of graduate student committees, I am close friends with several hundred other researchers, and I have reviewed hundreds of manuscripts submitted for publication. Every person that I have ever known worked hard to make his or her study the best study it could be. Although scientists do make errors, they work very hard and very intelligently at their research. There are almost no "garbage studies." The claim of garbage studies is a defensive thought process used by researchers to protect themselves against admitting that other studies had results that conflict with their own results.

THE MYTH OF THE 5% ERROR RATE

Many researchers think that the error rate for the significance test cannot be as high as 60% because the error rate for the significance test is guaranteed to be 5%. The problem is failed statistics teaching; these researchers do not understand the low power of the significance test in current psychology. All the graduate statistics teachers that I know say that they teach the facts about power in their basic 1-year statistics courses. They then complain that most students do not understand the facts about power.

The problem is that most current textbooks present the facts about low power in such a terse and wishy-washy manner that students do not see the main point. Let me present those facts in bald language that no one will misunderstand.

- KEY FACT: The critical assumption for the significance test is that the null hypothesis must be true. If the null hypothesis is actually true in a research setting, then the probability of falsely concluding that there is a treatment effect is the well- known Type I error rate of 5%.

- KEY FACT: If the null hypothesis is false, then there is no reason to believe that the significance test will work right.

- KEY FACT: If a researcher is doing a study in which there is a treatment effect, then the potential error rate for the significance test is far higher than 5%, 10%, or 50%; it is even higher than 90%.
  —For the conventional two-tailed test (i.e., analysis of variance), the maximum error rate is 97.5%.
  —For the one-tailed test, the maximum error rate is 95%.

Many researchers do not realize that there are two kinds of error in using a significance test. If the treatment has no effect, then the only kind of error that can be made is a Type I error: falsely claiming that the treatment has an effect when it does not. The Type I error rate is indeed only 5%.

But if the treatment does have an effect, then a Type I error is impossible. The errors will be Type II errors: falsely concluding that the treatment has no effect when it actually does have an effect. The Type II error rate is rarely as low as 5% for the sample sizes typical of research in the social sciences. Indeed, there are many research areas where studies are virtually guaranteed to fail—areas where the Type II error rate for studies is very near the maximum 97.5%. Certainly there are many areas—especially in advanced research—where the error rates are over 90%.

Many researchers say that the error rate may be high in other areas of research, but it is not high in their own area of research. However, it is my experience that these researchers have not calculated error rates (i.e., power analysis) for research in their areas. I have yet to meet a researcher who has actually computed the statistical power in his or her own area of research and who claims that power is uniformly high. Among the researchers I know, everyone who has done power analysis in his or her own area of research has been astounded to find out how high the error rate is. Current empirical studies of power in our leading journals are consistent with my personal experience.

The first empirical studies of power in leading psychology journals were done by Cohen (1962). Sedlmeier and Gigerenzer (1989) reviewed 12 empirical studies that computed the error rate for the statistical significance test in leading psychological journals before conducting their own study. Close reading of their results shows that the error rate was about 60% at the time of their study, the same error rate as in the research of 30 years earlier. Sedlmeier and Gigerenzer rightly bemoaned the fact that the error rate had not improved between 1960 and 1989. These findings indicate not only that studies are as inaccurate now as more than 30 years ago, but also that there has been no significant improvement in understanding sampling error in more than 30 years. If researchers did understand the high error rate

for the significance test, they would be using better analytical methods, such as confidence intervals, the inference probability, and meta-analysis.

THE MYTH OF THE NULL HYPOTHESIS

Many researchers do understand that a study can have low statistical power to detect an effect (i.e., a high Type II error rate). However, they believe that this is not a major problem because the null hypothesis is true in most studies. That is, they believe that most studies are devoted to treatments that have no effect and that the relevant error rate is therefore the low 5% Type I error rate. If most studies have a 5% error rate, then the average error rate cannot be as high as 60%.

The big problem with this position is that it has been empirically disconfirmed. The empirical studies of power consistently find error rates of 50% or more (Sedlmeier & Gigerenzer, 1989). The average error rate across empirical studies is 60%.

If the null hypothesis were true in all or nearly all studies, then the error rate for the significance test would be a low 5%. Empirical studies instead find an average error rate of 60%. If the null hypothesis were true in 50% of studies, then the average error rate would be at most 51% (assuming the maximum 97.5% error rate for the 50% of studies in which the treatment has an effect). The actual average rate is 60%, so the null hypothesis must be true in fewer than 50% of studies.

How often is the null hypothesis true? An estimate of this frequency could not be made until recently; no survey had been done. However, Lipsey and Wilson (1993) recently reviewed 302 meta-analyses of treatment study domains. They tried to find all meta-analyses of treatment studies in psychology, and they interpreted psychology broadly (especially including many studies from education). Because the average number of studies in each meta-analysis is 60, the total number of studies considered in their review is 18,120.

How many of the meta-analyses found the null hypothesis to be true? I scanned their Table 1 and found only 3 cases out of 302 in which the effect size was 0. That would suggest that the null hypothesis was true in a little less than 1% of the research domains considered. However, this is an overestimate because these 3 cases were alternative meta-analyses for the same effect: the effect of the open classroom on students' achievement. Thus, in this first survey of research domains in psychology, the estimate is that the null hypothesis is true in less than 1% of domains.

Consider the implication of Lipsey and Wilson's (1993) results. Empirical studies have now shown that the null hypothesis is rarely true. Because the null hypothesis is rarely true, the error rate relevant to actual research is not the low 5% error rate for Type I errors but rather the high, potentially disastrous Type II error rate. The good news is that the error rate is well short of the potential 97.5% for studies with extremely small samples. The bad news is that the average error rate is 60%.

Many people, however, cannot believe that the null hypothesis is almost always false. Mere empirical findings are not sufficient to dissuade them; they want an explanation. The following explanation is my current hypothesis as to why the null hypothesis is rarely true. In essence, I argue that current studies are designed in such a way that the null hypothesis will almost never be true. The key to my argument is the distinction between *debunking studies* and *confirmatory studies*.

The significance test was not constructed for universal use. Mathematical statisticians use significance tests only when they have strong reason to believe that the null hypothesis is true. If they have any reason to suspect that the null hypothesis is false, they use more accurate ways to analyze the data so as to detect the pattern in the results.

Consider the kind of study for which the significance test was originally derived: the debunking study. Suppose a researcher has strong reason to believe that a certain theory is false or that a certain treatment will not work. The researcher then designs a study to show that there is no effect. If the population effect size is actually 0, sampling error will almost always cause the observed result to differ from 0 in one direction or the other. The significance test was developed to see if the deviation from 0 is too large to be credible.

To the extent that a debunking study is based on a hypothesis that has strong a priori empirical or theoretical support, it is likely that the null hypothesis is actually true for that study. If it is true, then the significance test will have only a 5% error rate. That is, if a researcher is doing a debunking study and the researcher is right, then use of the significance test for that study is proper. However, in actuality, very few debunking studies are done.

Most researchers work hard at their literature reviews. Most proposals cite considerable theoretical support for the hypotheses to be tested. Many proposals cite prior empirical studies that show positive effects. Thus, in almost all current research, a strong literature review shows that the treatment studied will have an effect. That is, in almost all current research, the literature guarantees that the null hypothesis will be false. I refer to such studies as confirmatory studies. Given the quality of current literature reviews, it is extremely unlikely that a confirmatory study will be done in a domain where the null hypothesis is true. This means that nearly every current study is using the significance test in a domain where the significance test will not work and will have a high error rate!

To summarize, empirical results on statistical power show that nearly all confirmatory studies start with a correct hypothesis: The direction of the effect is correctly predicted. The many, many "not significant" findings are not cases of hypothesis error, but rather cases of Type II error stemming from the use of the significance test in a domain where it should not be used.

IS IT SAFE TO BAD-MOUTH THE SIGNIFICANCE TEST?

Many people already know that the significance test does not work. However, they do not say this in public, and they still use significance tests in their research articles. The reason is fear—fear of social sanctions if they violate a social convention. Unfortunately, there is a real reason for fear, though sanctions are much less common now than when I started my research career.

When Cohen (1962) published his article showing the high Type II error rate in psychological research, I thought that people would listen. However, subsequent study has shown that the frequency of use of the significance test and dependence on it have increased since

Cohen's work was published. However, Cohen's work was not wasted; there are now many psychologists who have read his publications and who understand that the significance test can have a very high error rate.

In an invited tutorial at the 1979 American Psychological Association (APA) convention, I recommended a moratorium on the significance test. There was open snickering in the audience, but there were also many people who came up afterward and congratulated me for having the courage to speak out publicly.

The situation has now changed radically. There is a sizable minority of psychologists who know that the significance test does not work. There is now an APA committee looking into the question of whether the use of the significance test should be discouraged. Both Cohen (1994) and Schmidt (1996) have stated openly that the significance test should be abandoned, and both their articles were published in leading journals.

CAN THE SIGNIFICANCE TEST BE SAVED?

Many researchers know that the significance test is deeply flawed, but they think that there is no alternative. The usual argument for necessity is the need for scanning (i.e., for separating out studies that provide strong evidence for the direction of the population effect from those in which the data are too weak for us to be sure of this direction). But, as I pointed out earlier, almost all studies correctly predict the direction of the effect, so scanning is unnecessary. For researchers who think the significance test is necessary, the question is one of reform. Can we change the current use of the significance test so that we can reduce the high error rate? In this section, I briefly discuss two such programs: Cohen's (1962, 1990, 1994) longtime program called power analysis and Harris's (this issue) new program that derives from the three-tailed test concept put forward long ago by Kaiser (1960).

For more than 30 years, Cohen tried to save the significance test by requiring authors to perform a power analysis so that they would know the actual error rates for their tests. People who do this quickly learn to largely ignore significance test results—which is the only rational solution to a 60% error rate. However, we now have 30 years of experience showing that power analysis cannot be taught successfully in the present environment. Authors do not learn and do not use power analysis. Cohen (1994) himself has given up on the training for power analysis and now admits that the significance test must be abandoned.

Harris argues that the computations of the significance test can be saved by using a radically different interpretation scheme. However, his scheme was put forward 35 years ago by Kaiser and was adopted by no one. So even though his new scheme would be an improvement, we already know that it will not work in practice. There are also a number of technical errors in his article, but these are beyond the scope of my discussion here. I will be happy to e-mail my detailed critique of his article to any interested party.

In any case, note carefully that Harris does not argue for the significance test as it is currently used. To see the radical nature of his scheme, note first that he eliminates the null hypothesis, which he believes to be illogical and ridiculous on its face. In his scheme, the three outcomes are "significantly positive effect," "significantly negative effect," and "too close to call." That is, in his scheme, you never conclude there is "no effect."

ALTERNATIVES TO THE SIGNIFICANCE TEST

The older quantitative sciences such as physics, chemistry, and electronics do not now use and have never used the significance test as it is currently used in psychology. This fact strongly suggests that there are alternatives to the significance test that are used in the other sciences. The dominant technique used in the quantitative sciences and the dominant technique used by mathematical statisticians is the technique of confidence intervals. I have studied the use of confidence intervals in other fields, and I am convinced that these fields do not suffer the high error rate that we suffer in psychology.

I have taught both significance tests and confidence intervals in the 1st-year graduate statistics course in psychology, and the students find confidence intervals easier to learn and understand than significance tests. Indeed, those students who understand confidence intervals also understand Type II error and the high error rate for Type II errors. Among those students who are taught only significance tests, fewer than 10% understand statistical power and the high Type II error rate. Teaching confidence intervals teaches the significance test better than teaching the significance test directly!

Many very good quantitative psychologists have serious misunderstandings about confidence intervals. The problem is that in current statistics courses, so much time is devoted to the myriad of significance tests that there is not enough time left to discuss basic concepts and strategies.

GENERAL CONCLUSION

The significance test as currently used in psychology has an average error rate of 60%. The error rate using the significance test is worse than the error rate for deciding by coin flip. The false "conflicting results" found by objective comprehensive reviewers cause reviewers to conclude falsely that most hypotheses are still not properly tested even when dozens of studies have been done. At this point, progress in a research domain typically grinds to a halt. The significance test has been a disaster for modern psychology.

There are alternatives to the significance test that do not have a 60% error rate. For single studies, we can use confidence intervals to measure the potential sampling error in study results. A 95% confidence interval always has only a 5% error rate; it is not context dependent like the significance test and thus cannot have the super-high error rates that the significance test has when people consistently use it in a context for which it was not designed. Likewise, 50% and 68% confidence intervals have context-independent error rates—of 50% and 32%, respectively. Furthermore, the various confidence intervals are all compatible with each other; there is no need for a social convention such as "$\alpha = 5\%$."

When multiple studies provide estimates of a given result, we can use meta-analysis to greatly reduce the impact of sampling error. That is, meta-analysis provides both an improved estimate of the typical

study finding and an estimate of how much real variation there is in those results.

REFERENCES

Cohen, J. (1962). The statistical power of abnormal-social psychological research. *Journal of Abnormal and Social Psychology, 65*, 145–153.

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

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

Kaiser, H.F. (1960). Directional statistical decisions. *Psychological Review, 67*, 160–167.

Lipsey, M.W., & Wilson, D.B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*, 1181–1209.

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

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