THE CONTROVERSY OF SIGNIFICANCE TESTING:
MISCONCEPTIONS AND ALTERNATIVES

By Dale N. Glaser, PhD. From Pacific Science & Engineering Group, San Diego, Calif.

The current debate about the merits of null hypothesis significance testing, even though provocative, is not particularly novel. The significance testing approach has had defenders and opponents for decades, especially within the social sciences, where reliance on the use of significance testing has historically been heavy. The primary concerns have been (1) the misuse of significance testing, (2) the misinterpretation of P values, and (3) the lack of accompanying statistics, such as effect sizes and confidence intervals, that would provide a broader picture into the researcher’s data analysis and interpretation. This article presents the current thinking, both in favor and against, on significance testing, the virtually unanimous support for reporting effect sizes alongside P values, and the overall implications for practice and application. (American Journal of Critical Care. 1999;8:291-296)

In the history of science, specifically since 1900, many researchers have resorted to tests of significance when testing hypotheses. Karl Pearson laid the foundation for significance testing as early as 1900. Ronald Fisher, as well as the team of Jerzy Neyman and Egon Pearson, formulated the practice of null hypothesis significance testing, involving such crucial components as type I and type II errors, as well as power. Many refinement of significance testing occurred during the 1950s (for a brief history of significance testing, see Cipriani and Huberty). Even Fisher made clear that the P value of .05 was at best a convention, with many advising against the rigid adherence to “.05” (some scholars dropped it as early as the late 1950s). For the most part, however, the schools of alarm have been summarily dismissed or minimized.28 Doubts about the misapplication of significance testing have again risen within the social sciences, with the psychological sciences particularly sounding the alarm. Part of this increased debate may be due to the long-lasting reliance on and misuse of significance testing within the behavioral and educational sciences, in conjunction with the documented ploy of complaints about significance testing. In brief, the controversy involves the sole use (and misinterpretation) of the P value without taking into account other descriptive statistics, such as effect sizes and confidence intervals, statistics that provide a broader glimpse into the data analysis. Hence, a primary objective of this article is to summarize the controversy surrounding the use of significance testing, solutions that are being proffered, and what implications the suggestions have for applied researchers in the health sciences. This article is geared for both consumers of clinical research and applied researchers, thus, statistical arguments are kept to a minimum. Even though significance testing has many critics, it still has a place for clinical researchers so long as it is used judiciously.
A Brief History
As we know from Kahn, paradigmatic changes in science take place slowly, especially when the change involves modifying or eliminating a practice that has been firmly entrenched, such as significance testing. In an oft-cited paper from 1966, Bakan1 asserted that "the test of significance does not provide the information concerning psychological phenomena characteristically attributed to it; and that, furthermore, a great deal of mischief has been associated with its use." In 1960, Rozobroa vigorously attacked the faulty inferences used in the test of significance and in 1957 asserted: "Null-hypothesis significance testing is surely the most bone-headedly misguided procedure ever institutionalized in the routine training of science students." In even more recent terms, Frank Read and John Hunter, two leading proponents of meta-analysis, have claimed that the use of significance testing actually retards the ongoing development of the research enterprise.13

Especially instrumental in the resurgence of interest in the topic of the misuse of significance testing was the 1964 article "The Earth Is Round (p<.05)" by Jacob Cohen, which appeared in the American Psychologist. Some readers will recognize Cohen as the author of a text on power analysis14 that is a mainstay of many researchers when determining sample sizes for their studies. In the American Psychologist article, Cohen cited errors that many researchers were still committing in the use of significance testing; most had to do with the flawed interpretation of the P value. Partially because of this provocative article, as well as Schmidt's address at the 1964 American Psychological Association (APA), debates on the misapplication of significance testing were reinvigorated, with many articles and convention forums focusing on this topic.15 Because of the controversy, the APA Board of Scientific Affairs appointed the Task Force on Statistical Inference, comprising some of the most influential researchers in psychology, statistics, and the social sciences. The task force first met in December 1966, and their initial report can be located at the following Web site: http://www.apa.org/sci-ence/ful.html. To the relief of many, the task force did not recommend the abandonment of significance testing, but rather advised that supplementary tools (e.g., effect size, confidence intervals) should be used with the same frequency that researchers currently use significance testing.16 Following in the footsteps of the task force, in 1997 two researchers asked a similar question: What if There Were No Significant Results?17

Defining the P Value
Regardless of the statistical technique used (e.g., parametric vs nonparametric) or type of design (e.g., experimental vs quasi-experimental), P values are often reported in clinical research. The P value (the applied researcher will recognize this as "Sig." in most statistical printouts) is compared with the a priori alpha level (α), which serves as the basis for rejecting or failing to reject the null hypothesis. However, misinterpretations of the P value and what it entails are not unusual. Part of this misinterpretation may stem from the lack of uniformity across texts in the definition of the P value. Even though the exact terminology may differ, one generally agreed-upon definition of the P value is that we are ultimately testing the null hypothesis against a level of significance (α) designated by the researcher.

Significance and Importance
Indiscriminate use of certain statistical principles occurs early in the training of clinical researchers. An example in statistics is the frequent exhortation that "correlation does not imply causality" (for an interesting discussion of correlation and causality in the context of structural equation modeling, see Bullock et al18). Another principle that has been part of the statistical training of applied researchers is that "statistical significance does not mean the same as importance." Even though many authors have clearly differentiated statistical significance from practical significance, it is still not unusual for researchers to herald the P value as being synonymous with importance.19 Bruce Thompson,20 editor of the influential Educational and Psychological Measurement, states in regard to the issue of practical versus statistical importance: "Improbable events are not intrinsically interesting. Some highly improbable events, in fact, are completely inconsequential." The following example illustrates this problematic misconstruing of significance testing.

A dietary change tested in a randomized clinical trial results in a mean decrease in total cholesterol level, from 243 mg/dL to 236 mg/dL, in the experimental group and no change in the control group. It
this mean difference of 5 mg/dL enough to modify clinical practice? Most clinicians would judge the dif-
ference insignificant. But what if the difference is statistically significant (i.e., P<.05)? Many might be
tempted to view this difference as more important. If the pooled standard deviation of this example was 25
and the groups had 400 patients each, a difference of 5 mg/dL would indeed be statistically significant.

At this juncture, it becomes critical to distinguish the phenomenon of statistical significance (deter-
mined in part on the basis of the chosen α level) from clinical significance. If obtaining significance (i.e.,
rejecting the null hypothesis) is the predominant objective of the researcher, (even diminishing the proba-
ability of type II error (α, failing to reject the null hypothesis when it is false, or saying no differ-
ence exists when a difference may exist) is warranted. Hence, many clinical researchers are
savvy enough to do a power analysis before the start of the study. Given a postulated effect size, direction-
ality of effect (i.e., 1-tailed vs 2-tailed), α level, and desired power (power of a statistical test being the proba-
nability of yielding statistically significant results), a sample size for the study at hand can be deter-
dined. What is clearly evident, as illustrated in the cholesterol example, is that, given a large enough
sample size, it becomes increasingly easy to reject the null hypothesis. But rejecting the null hypothesis is
not synonymous with import of effect. Hence, there has been a strong movement of late (even though
Cohen has been discussing the issue of examining effect sizes for years) to report not only P values but
also effect sizes.1

Effect sizes, which can be calculated for virtually any type of statistical analysis, assist researchers in
arriving at a measure of magnitude or strength.2 In the cholesterol example, if the absolute value of the dif-
ference between the means is divided by the pooled standard deviation (5/25), I arrive at Cohen's d, an effect
size that is expressed in units of a standard devi-
ation. So, the "significant" finding in the cholesterol analysis reveals an effect size that is a standardized
score of only .20. To put α in context, a small d is considered to be 0.2, according to Cohen's guidelines
of small, medium, and large effect sizes. However, and as Cohen warns, these effect sizes should be used
only as guidelines and not as blind conversion, because sometimes small effects may indeed be clin-i-
cally important (e.g., decreased mortality).3 Reporting the effect size, in conjunction with the result of signif-
icance testing, provides a broader perspective as to what the data mean.4 It is not enough to state that "a
significant and positive correlation was obtained."5 P

.01"; hence, this result's P value may be associ-
ated with a =.20, accounting for just 4% of the vari-
ance. Now, 4% may be impressive in some research
contexts, especially if the manipulation of the inde-
pendent variable was weak or the research team was
investigating unexplored terrain, but it still leaves
96% of the variation unexplained.5 Hence, the onus
should be placed on researchers to report effect sizes
with the same regularity as they report P values.6

P as an Indicator of Strength

In those cases in which the clinical researcher
obtains a P value that is about, 0.05 (e.g., P = .048), he or
she may choose to suspend judgment, invoking the
paraphrased truism: "Surely, God loves .048 as much as .049!"7 The researcher makes a predetermined
decision about the level of significance (α) necessary to reject the null hypothesis. This α level (generally set at .05) is also the type I error (α, probability of wrongly rejecting the null hypothesis when it is true—say ing a difference exists when it does not).8,9 It is essential to keep in mind that signifi-
cance testing is inherently dichotomous: the ultimate decision is to reject the null hypothesis or to
fail to reject the null hypothesis. When criticisms are put aside, significance testing still has an important role in the sciences for making ordinal claims (reject vs fail to reject).10 Significance testing "provides us with the criteria by which provisionally to distinguish results due to chance variation from results that represent systematic
errors in data available to us."11 However, what is most egregious is when the researcher interprets the
P value in 1 of the 2 following ways (1) "the obtained results (P = .048) approached significance" or
"the results were marginally significant" and (2) "the obtained P value of .0001 indicates that the results were very significant." Both of these depictions are fallacious and misleading. Regarding the first interpretation, Thompson12 provides an example of an ap retort: "How did you know your results were not trying to avoid being statistically signifi-
cant?" In an effort to understand (or enthrall) their findings, researchers turn to P values as indices of
importance, a logic that is deeply flawed. It is in specious to suggest that a study with a P value of .001 is "more important" than a study with a resultant
P value of .049.13 P values say nothing about magni-

dude or import. Those types of apple-and-orange con-

trasts lead researchers and their audience to draw conclu-
sions about a magnitude of effect that may be

mathematically infinitesimal.

Many editors require that authors provide exact P
values in the results section, but the process of significance testing is dichotomous: you either reject or fail to reject the null hypothesis, given your sample data, when comparing the resultant $P$ value with the nominal level. So does the $P$ value, $p < 0.05$ in the first interpretation warrant the author's use of support for "marginal significance," given the inherently dichotomous nature of significance testing? It is true that with a larger sample size (or a stronger manipulation—an issue generally not factored in when power analysis is done), the observed $P$ might be less than 0.05. So, instead of claiming marginal significance on an a posteriori basis, researchers should report the effect size and let the results carry their own weight. Unfortunately, and as many researchers are painfully aware, null results rarely see the light of day in the published literature, a phenomenon that has been referred to as the "file drawer problem." Thus, the incentive is great to claim marginal significance in a study that is void of "significant" effects; again, this situation is one in which the publishing of effect sizes can be most beneficial.

Now, what about the case of the ubiquitous asterisks? In the following scenario, the researcher sets the $a$ at a predetermined level (generally $a = 0.05$), obtains a $P$ value of 0.01, and then, with unintentional trickery, uses asterisks: $**P < 0.05$; $$$P < 0.01$; $$$$P < 0.001$. This practice goes researchers into conclusions that their findings are "more significant" and hence "more powerful" than the individuals truly are, a conclusion that is deceptive and misleading. Cronbach and Snow,3 as cited in Pedhazur and Schmelkin,4 assert: "A $P$ value chatched by classical methods is not a summary of the data. Nor does the $p$ attached to a result tell how strong or dependable the particular result is ... Writers and readers are all too readily to read ... as $p (H: E)$, the probability that the Hypothesis is true given the Evidence." As textbooks on statistics reiterate almost in vain, is $p (H: E)$, the probability that this Evidence would arise if the null Hypothesis is true." Cohen called this type of confounding the inverse probability error. It is incumbent that researchers realize the dubious practices, as well as faulty logic, of ascribing strength or magnitude to a $P$ value. Unfortunately, the requirement of many journals to report asterisked $P$ values, absent of effect sizes, perpetuates the myth that the $P$ value denotes importance.

The Null and the NII

The language of significance testing, when introduced at the undergraduate level, can indeed be for-bidding, if not somewhat contrary. When we reject the null hypothesis, we encourage language such as "prove" or "confirm, " when referring to the alternative hypothesis. Given methodological artifacts such as sample specificity and measurement error, in the spirit of the Popperian notion of falsification, we can only hope to disconfirm the null hypothesis.6 We use double negatives when we assert that no difference exists—this is, we fail to reject the null hypothesis—or we claim that the results are "inconclusive." We do not encourage the student to "accept" the null hypothesis in the case of a decision to fail to reject the null hypothesis, for failing to reject the null hypothesis does not prove that the null hypothesis is true.7 Even though the irreproducible student may see this wordplay as a basisometer of the researcher's unwillingness to take a theoretical stand, this line of reasoning, tenuous as it may appear, characterizes the scientific pursuit. Claiming a significant result does not prove the veracity or the durability of the theory. In fact, one misconception often associated with significance testing is that $1 - P$ denotes the probability that the findings will be replicated in future studies. This misconception has been tested the replicability failure.8 Every study is unique, with all its attendant idiosyncrasies: sample characteristics, environmental variables, and other factors that may contribute to error variance. Hence, rejecting the null hypothesis ($P < 0.05$) for 1 study, even though that rejection provides a guidance to future research, does not prove anything.

Furthermore, failing to reject the null hypothesis ($P > 0.05$) does not provide corroborating evidence for the nonexistence of the phenomenon. As stated earlier, failing to reject the null hypothesis does not prove the null hypothesis: For any given study, countless factors may have culminated in a fail-to-reject decision: faulty manipulation, of the independent variable, insufficient power, unreliable measures, and threats to internal validity. Accepting the null hypothesis has led many a researcher to conclude that the test statistic equals the parameter, that is, the sample mean minus the population mean equals zero. Part of this logic may be attributed to the default settings in many of the popular statistical software programs and to the discussion of significance testing in introductory statistics texts. For example, the test of significance for the Pearson product moment correlation coefficient ($r$) is compared with that of the population correlation ($p$), with the latter assumed to be zero. However, how realistic is this point estimate of zero? If a researcher were to correlate nursing satisfaction and job performance, and on the basis of meta-analysis findings the average
effect size is r = 0.32, does comparing this sample esti-
mate to a population parameter of r = 0 seen as all
logical? A more defensible tactic might be to compare
the sample estimate against a non-zero effect size—an
effect size that makes sense given the content of the
study. Cohen, influenced by Meehl, has had provocative
views about the probabilistic application of
testing a point estimate against a parameter of
zero.6,7 This mode of significance testing has been
widened the null hypothesis, referring to the postulate
that an effect size for a given null hypothesis is zero.
The argument is that there will always be some differ-
beneath an estimate and a parameter (or differ-
ence between 2 samples in the context of a t test, in
which the null hypothesis revolves around a zero dif-
ference, so failing to reject the null hypothesis (or by
software settings, the nil hypothesis) does not confirm
that the parameter equals the hypothesized value. In
fact, some extreme factions have asserted that the null
hypothesis is always false, at least to some degree.
Even though this line of reasoning can result in the
deaths of statistical absolutism, the prevailing
message is that we do not accept the null hypothesis
or prove the null hypothesis. Some difference will
always be present; only by investigating the effect
sizes can researchers answer the question, How much
difference and does it matter?

Suggestions and Alternatives
Positioning oneself as the naysayer is easy; it does
not take great facility to degrade practices that have
been ingrained in the mainstream. The larger challenge
is to provide suitable alternatives that can withstand
reasonable debate and be instituted either alongside or
in place of hitherto conventional practices.
As a first recommendation, and without excep-
tion, effect sizes should be reported by clinical
researchers with the same frequency that P values are
reported. As a matter of practice, all pertinent descrip-
tive statistics should be reported (eg, means, standard
deviations, and any graphical displays such as box
plots). Even the Publication Manual of the American
Psychological Association,17 recognizing the integral role of effect sizes, encourages researchers to
provide effect-size information.7 Whereas some edi-
tors have maintained other emphatic points of view
about the reporting of effect sizes, with Bruce
Thompson, editor of Psychological and Educational
Measurement, requiring such reporting, the frequency
of reporting effect sizes remains abysmally low.18,19
Perhaps instead of "encouragement," the APA should
"require" the reporting of effect sizes.20 Consen-
murate with this philosophy, the editor of the
Journal of Applied Psychology, Kevin Murphy,21 has
stated the following: "If an author decides not to
present an effect size estimate along with the out-
come of a significance test, I will ask the author to
provide specific justification for why effect sizes are
not reported. So far, I have not heard a good argument
against reporting effect sizes. Therefore, unless there
is a real impediment to doing so, you should routinely
include effect-size information in the papers you sub-
mit." The preceding guidance did receive approval
from the APA Publications Committee.22
A second recommendation is that the reporting of
point estimates and confidence intervals should be a
regular practice.23 Confidence intervals, which pro-
vide a margin of error around a sample estimate, not only
are helpful for assessing the extent of error around the estimate but also reveal some of the informa-
tion that significance testing yields.21 As Cohen has
implied, however, the lack of reporting of confidence
intervals may be due to the existence of the "embar-
assing large intervals that transpire in many research-
ers' results."23 As for point estimates, it may
serve researchers well to consider testing the null
hypothesis against an estimate that is logically defi-
sible (eg, a normative correlation between nurses' satis-
faction and performance), as opposed to the default
null (ie, zero effect) hypothesis.

Another option, even though relatively ambitious
in application, is the use of meta-analysis. Meta-anal-
ysis integrates findings across multiple studies (eg,
studies focusing on nursing interventions), with the
effect size of the individual study generally being the
unit of analysis. Franz Schmidt and John Hunter are
longtime proponents of the use of meta-analysis,
largely motivated by their concern about the deficient
power associated with most individual studies.24,25 In
case the author's assertion that the average individu-
ual study has a power of .50, hence, a probability of
Type II error of 50, then one must be astounded that
50% of studies are not sufficiently sensitive to the
experimental or correlational effect of interest. Thus,
are the results of most individual studies, even if all
have fail-to-reject decisions (ie, nonsignificant
findings) will end up in essential! mathematical
"pie drawer," never to see the light of day in pub-
lished journals.25 Meta-analysts have argued that the
strength in their technique is that the process of com-
paring numerous studies, across various samples and
settings, offsets the deficient power associated with
most individual studies. This affords researchers the
opportunity to draw more general conclusions that
would, by necessity, be concluded in the context of
any single study.26 Conversely, some criticisms are

295
associated with meta-analysis (eg, only studies with nonsmall findings are considered, possibly resulting in overestimation of the true effect size), so meta-analytic findings should be used as a guide rather than a data-driven panacea. Use of meta-analysis entails a potentially immediate extension of resources that may not be realistic for clinical research.

Further, given the high number of unreported studies, clinical researchers must pay more attention to the power of their studies. As frequently emphasized in many studies, the most sophisticated statistical mechanisms cannot salvage a flawed design. It behooves researchers not only to consider sample size in their efforts to enhance the power of their study but also to evaluate the manipulation of the independent variable.

Despite the ramifications and potential overtones of the proposed testing of the significance tests, a more compendious solution has been offered by a wide array of researchers.87,88 Significance testing will always have its advocates and opponents. At this time, however, most that any other researchers are considering the import of increasing effect sizes, confidence intervals, and power analysis alongside the traditional notion of significance testing. The recommendation is not that clinical researchers disavow significance testing, but rather that they incorporate additional information that will supplement their findings.

REFERENCES


