The “significance” crisis in psychology and education

Bruce Thompson*

Department of Educational Psychology, Texas A&M University, College Station, TX 77843-4225, USA

Abstract

The present article explores various reasons why psychology and education have not proceeded further with adopting reformed statistical practices advocated for several decades. Initially, a brief statistical history is presented. Then both psychological and sociological barriers to reform are considered. Perhaps economics can learn from some of the mistakes made within psychology and education, and invent its own new pitfalls, rather than becoming mired in the same crevasses discovered by other disciplines.

Keywords: Significance; Psychology; Education

1. Introduction

As noted elsewhere within this special issue, criticisms of statistical significance tests are almost as old as the methods themselves (e.g., Boring, 1919; Berkson, 1938). These criticisms have been voiced in disciplines as diverse as psychology, education, wildlife science, and economics, and the frequency with which such criticisms are published is increasing (Anderson et al., 2000).

Some of the most widely cited exemplars of these critiques were provided by Carver (1978), Cohen (1994), Schmidt (1996), and Thompson (1996). Some of these critics have argued that statistical significance tests should be banned from psychology/education journals. For example, Schmidt and Hunter (1997) argued that “Statistical significance testing
retards the growth of scientific knowledge; it never makes a positive contribution” (p. 37). Rozeboom (1997) was equally direct:

Null-hypothesis significance testing is surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students . . . [I]t is a sociology-of-science wonderment that this statistical practice has remained so unresponsive to criticism . . . (p. 335)

Harlow et al. (1997) provided a balanced and comprehensive treatment of these arguments in their book, “What if there were no significance tests?” Hubbard and Ryan (2000) and Huberty (1999, 2002) presented the related historical perspectives on the uptake of statistical testing within psychology and education.

The purpose of the present article is not to summarize these numerous criticisms, nor the counterarguments to some of them. Instead, some reasons why psychology and education have not proceeded further with adopting reformed practices are explored. Because the field has moved inexorably, albeit glacially, the question could also be framed as, “What barriers have slowed the recent progress of the field in adopting more informative methods?”

First, however, a brief history is presented. Perhaps economics can learn from some of the mistakes made within psychology and education, and invent its own new pitfalls, rather than becoming mired in the same crevasses discovered by other disciplines.

2. A brief history

In 1994, the fourth edition of the Publication Manual of the American Psychological Association (APA) first mentioned statistics that characterize magnitude of effect (e.g., “effect sizes” such as Cohen’s , , , ). Effect sizes quantify by how much sample results diverge from the null hypothesis. There are dozens of effect size statistics (Kirk, 1996). The estimation and interpretation of effect sizes are discussed by Snyder and Lawson (1993), Rosenthal (1994), Hill and Thompson (2004), Kirk (in press), and Thompson (2002a,b, in press).

The 1994 APA Publication Manual noted that, “Neither of the two types of probability values reflects the magnitude of effect because both depend on sample size . . . You are encouraged to provide effect-size information” (APA, 1994, p. 18, emphasis added). However, 12 studies of 23 different journals showed that this “encouragement” had little, if any, effect on reporting practices (Vacha-Haase et al., 2000).

In 1996, the APA Board of Scientific Affairs created a Task Force to recommend whether or not statistical significance tests should be banned from journals. The Task Force published its suggestions three years later (Wilkinson and APA Task Force on Statistical Inference, 1999), and did not recommend such a ban, but did present numerous related recommendations.

Included were admonitions about effect sizes, such as “Always [emphasis added] provide some effect-size estimate when reporting a p value” (Wilkinson and APA Task Force, 1999, p. 599). The APA Task Force further emphasized, “reporting and interpreting effect sizes in the context of previously reported effects is essential [emphasis added] to good research” (p. 599). The Task Force also strongly emphasized the value of reporting confidence intervals.
In 2001, APA published the fifth edition of the *Publication Manual*, which stated that “For the reader to fully understand the importance of your findings, it is *almost always necessary* [emphasis added] to include some index of effect size or strength of relationship in your results section …” (pp. 25–26). The new Manual also labelled “failure to report effect sizes” as a “defect in the design and reporting of research” (APA, 2001, p. 5). And as regards confidence intervals, the 2001 manual suggested that CIs represent “in general, the *best* reporting strategy. The use of confidence intervals is therefore *strongly recommended*” (p. 22, emphasis added).

Clearly, over the past decade, some progress has been realized. Nevertheless, the revised manual is not without its critics, as revealed in the interviews conducted by Fiona Fidler (2002), and summarized in her penetrating essay.

Today, researchers in psychology and education seemingly now recognize that “surely, God loves the 0.06 [level of statistical significance] nearly as much as the 0.05” (Rosnow and Rosenthal, 1989, p. 1277). Changed views are reflected in the increasing uptake of three premises.

First, researchers in these disciplines increasingly accept that magnitudes of effect size are at least as important as (and perhaps more important than) indices of the probability of sample results. As psychologist Roger Kirk (1996) explained,

\[ \ldots \text{a rejection means that the researcher is pretty sure of the direction of the difference.} \]

Is that any way to develop psychological theory? I think not. How far would physics have progressed if their researchers had focused on discovering ordinal relationships?
What we want to know is the size of the difference between A and B and the error associated with our estimate; knowing that A is greater than B is not enough. (p. 754)

For this reason, today 24 journals in psychology and education have gone beyond the admonitions of the APA *Publication Manual*, and explicitly require effect size reports. Included are journals with more than 50,000 subscriptions!

Second, researchers in these disciplines have increasingly recognized that evidence of result replicability is important, but that statistical significance tests do not evaluate whether or not results are serendipitous. Instead, these fields have turned toward “interpretation of new results, once they are in hand, via explicit, direct [emphasis added] comparison with the prior effect sizes in the related literature” (Thompson, 2002b, p. 28).

Third, researchers are increasingly recognizing the value of reporting and interpreting effect sizes by graphically presenting confidence intervals for effects from both current and prior related studies (Thompson, in press). New software (e.g., Algina and Keselman, 2003; Altman et al., 2000; Cumming and Finch, 2001; Kline, 2004) has greatly facilitated these endeavors.

3. Barriers to reform

Journal editor Loftus (1994) lamented that repeated publications of concerns about statistical significance testing “are carefully crafted and put forth for consideration, only to just kind of dissolve away in the vast acid bath of our existing methodological orthodoxy” (p. 1). Another editor commented that “p values are like mosquitoes” and apparently “have
an evolutionary niche somewhere and [unfortunately] no amount of scratching, swatting or spraying will dislodge them” (Campbell, 1982, p. 698). Nevertheless, as we have seen, some progress on these matters has been achieved at least in psychology and education.

In formulating models of barriers to reform, there may be a natural tendency to frame explanations that draw on the language and theories of one’s own discipline. For example, in the present journal issue economic explanations are cited. It is argued that statistical significance testing remains unscathed because the methods are cheap, or because market inefficiencies have occurred.

To compliment these views through the lens of economics in the current issue, here two alternative perspectives are offered. Of course, the explanations are not mutually exclusive, and in fact these various barriers to reform may be comorbid.

4. Psychological barriers

In his classic 1994 article, Cohen concluded that the statistical significance test “does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!” (p. 997). Similarly, Rozeboom (1960) observed that “the perceptual defenses of psychologists are particularly efficient when dealing with matters of methodology, and so the statistical folkways of a more primitive past continue to dominate the local scene” (p. 417). Such views suggest that psychological dynamics must be considered as part of the etiology of resistance to change, and indeed various psychological theories have been offered.

For example, Schmidt and Hunter (1997) seemingly argued that statistical testing arose without thoughtful consideration. Once the adoption began, it became self-sustaining. They suggested that “logic-based arguments [against statistical testing] seem to have had only a limited impact . . . [perhaps due to] the virtual brainwashing [emphasis added] in significance testing that all of us have undergone” (pp. 38–39).

An alternative model speaks of a “psychology of addiction to significance testing” (Schmidt and Hunter, 1997, p. 49). The nature of the addiction is not fully explored in their treatment. But Thompson (1993) offered a possible model, pointing to the atavistic desire to escape the burdens of responsibility. He noted that researchers presenting work when questioned about import could merely say, “the result is statistically significant ($p < 0.05$)," and the answer was universally and thoughtlessly accepted. A seemingly objective and accessible mechanism for routinely adjudicating result value had seemingly been isolated.

Unfortunately, as Thompson (1993) noted, $p$ values cannot contain information about result value. He explained that valid deductive argument may not contain in conclusions any information completely absent from premises. He noted, “If the computer package did not ask you your values prior to its analysis, it could not have considered your value system in calculating $p$’s, and so $p$’s cannot be blithely used to infer the value of research results” (p. 365).

Yet another psychological explanation invokes the theory of cognitive dissonance. During the 1950s, psychologist Leon Festinger (1957) thought it would be intriguing to observe people who believed they had received messages that the world would end on a given date, and who abandoned their worldly possessions and/or families to congregate together in a
given place at the appointed time. Festinger then observed these people once the appointed times for the apocalypse had come and gone, to see how they rationalized their decisions and the consequences.

In essence, Festinger’s theory says that when beliefs or expectations are perceived to conflict with other beliefs, or reality, people seek congruence by adjusting perceptions to be less dissonant with each other. Festinger (1961) did a study with rats, published in APA’s flagship journal, suggesting that the need to resolve dissonance may even be biological.

The psychological theory of cognitive dissonance can be applied at two levels to the phenomenon of continued emphasis on statistical significance testing. First, students typically find the mastery of the logic of these tests convoluted, and consequently, painful. One aspect of cognitive dissonance theory may be summarized as: we come to love that for which we have suffered. Students may resolve their dissonance over painful exposure to statistical testing by exaggerating the value of statistical significance tests.

Second, dissonance may impact faculty as well. Faculty who had published over the course of careers, always with p values, may seek to avoid the dissonance of acknowledging the errors of past ways by stubbornly denying the validity of arguments favoring reformed analytic practices.

5. Sociological barriers

Professionals organize themselves into discipline-related professional associations. Scholars tend to do whatever is viewed as normatively correct within their disciplines, and professional associations tend to enforce these expectations at their meetings and in their journals.

Professional associations create sociological barriers to reformed statistical practices in two aspects. First, the bureaucracies of professional associations inevitably create time delays in responding to reform initiatives. The committees (e.g., Publications Committee) of professional associations only meet and conduct their business periodically (e.g., once or twice a year). Furthermore, some professional associations are hierarchically organized, and reform initiatives must be approved in a sequential order. Such structures create additional delays, and provide opportunities to stop reform initiatives. For example, editorial policies requiring effect size reporting in journals of the American Educational Research Association (AERA) would first have to be approved by the AERA Publications Committee, and then by the AERA Governing Council.

Consider the example of the APA Task Force on Statistical Inference. The APA Board of Scientific Affairs took two years to discuss the creation of the Task Force. Then the primary report of the Task Force was presented after an additional three years had passed.

Second, each professional organization has its own culture. The last few years have seen paradigm shifts in psychology and education as regards quantitative as against qualitative methods. Some professional associations have responded in an avoidant manner by creating cultures to sidestep these conflicts by adopting nonjudgmental perspectives. In such environs anything goes, and the maintenance of academic rigor is left to the conscience of each individual scholar. Statistical reforms are disadvantaged if the efforts are advanced in cultures that are hostile to any exercise of collective judgment.
6. Some concluding thoughts

Notwithstanding psychological and sociological (and economic, and other) barriers, statistical reform is occurring within psychology and education. Researchers working here may still ask, “how probable are the sample results?”, and may still report \( p \) values.

But the focus of inquiry in psychology and education has now shifted to addressing two questions: (1) how big are detected effects? and (2) are the effects replicable? To address the first question, researchers are evaluating effect sizes (cf. Kirk, 1996; Thompson, in press). They are evaluating the magnitude of effects by considering the context of the research (e.g., medical life-or-death outcomes, educational instructional effectiveness), and by explicit, direct comparison of study results with those in prior related studies (cf. Thompson, 2002b).

As regards the second question, scholars in psychology and education have moved away from accepting statistical significance tests as evaluating the replicability of results (see Cohen, 1994; Thompson, 1996). Instead, the replicability of effects is increasingly being judged by evaluating how stable effects are across a related literature, using what some have come to call “meta-analytic thinking” (Cumming and Finch, 2002).

In other words, don’t tell me your results are improbable, or highly improbable. Tell me they have practical import, reflected in effect sizes. Tell me explicitly why you think a given effect size, given what you are studying, is important. And give me the evidence that effects across studies are reasonably comparable, so that I have some confidence that your results are replicable and not serendipitous. Do not use statistical significance tests as a warrant for replicability, because that is not what the tests do (Cohen, 1994; Thompson, 1996). Instead, use replicability as evidence of result replicability!

References

Cohen, J., 1994. The earth is round \( (p < 0.05) \). American Psychologist 49, 997–1003.
Harlow, L.L., Mulaik, S.A., Steiger, J.H. (Eds.), 1997. What if There were no Significance Tests? Erlbaum, Mahwah, NJ.


Loftus, G.R. 1994. Why psychology will never be a real science until we change the way we analyze data. Paper Presented at the Annual Meeting of the American Psychological Association, Los Angeles.


