Psychology in the Schools, Vol. 44(5), 2007
Published online in Wiley InterScience (www.interscience.wiley.com).

DOI: 10.1002/pits.20233

LACK OF STATISTICAL SIGNIFICANCE

THOMAS J. KEHLE, MELISSA A. BRAY, AND SANDRA M. CHAFOULEAS

University of Connecticut

TAKUJI KAWANO

The University of Tokushima

Criticism has been leveled against the use of statistical significance testing (SST) in many disciplines. However, the field of school psychology has been largely devoid of critiques of SST. Inspection of the primary journals in school psychology indicated numerous examples of SST with nonrandom samples and/or samples of convenience. In this article we present an argument against SST and its consequent *p* values in favor of the use of confidence intervals and effect sizes. Further, we present instances of common errors that impede cumulative knowledge in the literature related to school psychology. © 2007 Wiley Periodicals, Inc.

Within school psychological research, there is a common insistence that data be subjected to statistical significance testing (SST). It is assumed that the consequent *p* value is an essential component for the proper reporting of results. Literally hundreds of articles have been published outlining the numerous defects inherent in SST. However, in the field of school psychology, these arguments are either unknown or simply ignored. Most egregiously, the flawed logic of SST is rarely even discussed in the school psychology journals. SST remains popular and widely embraced. The purpose of this article is to present an argument against the use of SST in that, in reality, there is little chance of drawing a random sample of human subjects from a defined population and, further, even if random selection from a population is realized, SST remains of little or no value.

Do We Really Need Statistical Significance Testing

As early as 1978, Carver argued against the use of SST. However, such advice does not often resonate in psychology in general and, particularly, not in school psychology. Fidler, Thomason, Cummings, Finch, and Leeman (2004) stated that although SST has been criticized in many disciplines, within psychology, "there has rarely been mention of such criticisms, or of the reform efforts in ecology, medicine, and other disciples" (p. 119). In their article, Fidler et al. noted that the 1996 APA Task Force on Statistical Inference (TFSI) was charged to investigate a proposal to ban SST. However, the TFSI did not even discuss Shrout's (1997) argument suggesting that the editorial decision to ban SST in the *American Journal of Public Health* resulted in "dramatically improved statistical practices" (p. 119). Fidler et al. also stated, "For a discipline that claims to be empirical, psychology has been strangely uninterested in evidence relating to statistical reform" (p. 119).

Johnson's (1999; p. 767) litany of quotations cited in his award-winning publication "The Insignificance of Statistical Significance Testing" presented a history of harsh criticism of SST:

> In 1963, Clark noted that it was "no longer a sound or fruitful basis for statistical investigation" (p. 466).
>
> Bakan (1966) called it "essential mindlessness in the conduct of research" (p. 436).
>
> Deming (1975) commented that the reason that students have problems in understanding hypothesis tests is that they may be trying to think.

Correspondence to: Thomas J. Kehle, 249 Glenbrook Road, Unit 2064, Storrs, CT 06269-2064. E-mail: kehle@uconn.edu

> Carver (1978) recommended that statistical significance testing should be eliminated; it is not only useless, it is also harmful because it is interpreted to mean something else.
>
> Guttman (1985) recognized that, "In practice, of course, tests of significance are not taken seriously" (p. 7).
>
> Loftus (1991) found that it was difficult to imagine a less insightful way to translate data into conclusions.
>
> Cohen (1994) noted that statistical testing of the null hypothesis "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).
>
> Barnard (1998) argued, "simple *p*-values are not now used by the best statisticians" (p. 47).

Additionally, Rozeboom (1997) noted that, "Null-hypothesis significance testing is surely the most boneheadedly misguided procedure ever institutionalized in the rote training of science students" (p. 335). SST "is one of the worst things that ever happened in the history of psychology" (Meehl, 1978, p. 817), and that it is a "wonderment that this statistical practice has remained so unresponsive to criticism" (Rozeboom, p. 335).

It is puzzling to explain the continued, irrational, and widespread support of SST. Perhaps the continued use of SST, regardless of the subjectivity involved or the lack of meaningful information provided, is because it appears to provide, but in reality does not, an objective and valid procedure to control for chance. A second plausible reason for the popularity of SST is perhaps the lack of understanding of what the *p* value means.

The American Psychological Association (APA; 2001) strongly recommended that results derived from inferential statistics include information about the obtained magnitude and direction of the effect and confidence intervals, which, taken together, clearly convey the importance of the research finding. Regardless of recommendations of the APA, the use of effect sizes to report research findings is not that frequent, and the use of confidence intervals is even less so (Fidler et al., 2004; Volker, 2006). Even with the clarity and interpretability afforded by effect sizes and confidence intervals to the extent that they preclude the need for SST, they are still most often used as supplements to the *p* values.

There is a belief in school psychology that single-subject designs are seriously limited in that they do not afford generalization. This is true. However, similarly, group designs should not be considered devoid of this limitation in that most school psychological research employs participant samples drawn from nonrandom populations of convenience or total populations (e.g., classrooms, etc.). In both single-subject and group designs, ultimately the external validity of the results can be determined primarily by whether or not the results can be subsequently and consistently replicated. The relevant questions should be: how big is the effect? How important is it? Can it be independently replicated? And often, particularly in school psychological research: does the effect endure? If these questions are answered, the notion of statistical significance is irrelevant.

Do Not Misinterpret the Meaning of the P Value

Simply, SST is conducted to determine whether the null hypothesis can be rejected. That is to say, if the null hypothesis is rejected, one can then conclude that the difference noted between sample means reflects a similar difference between population means. A test of statistical significance should only be conducted, if at all, to determine how probable that it is that the differences found between the samples would also be found in populations from which they were randomly selected (Gall, Gall, & Borg, 2003). This dichotomous statistical conclusion is quite limited; nevertheless, the *p* value is often granted an illusory power in the information it conveys (Volker, 2006).

There are several common misinterpretations of *p* values. Four of these are particularly salient erroneous assumptions (Gall et al., 2003):

1. The *p* value is an indication of "the probability that the differences found between groups can be attributed to chance" (p. 138).
2. The *p* value indicates the probability that the research hypothesis is correct.
3. The *p* value indicates the "probability of finding the same research results if a replication study was conducted" (p. 139).
4. The *p* value is an indication of the practical or theoretical importance of the results.

What Should Replace SST

We have argued that effect sizes and confidence intervals should replace SST, particularly in school psychological research, where it is improbable to employ truely random selection from defined populations. Further, even when employing inferential statistics in instances where correct random selection procedures are realized, we see no need to use SST. It is, at best, of a minuscule value that is eclipsed by information afforded by effect sizes and confidence intervals, which allow for some understanding of the range of the practical effect. When we report confidence intervals, we need to be aware of the difference between confidence intervals for means and confidence intervals around effect sizes (Thompson, 2002). Smithson's (2003) monograph contains many empirical examples of how to construct and interpret confidence intervals for a fairly wide range of statistical techniques, including effect sizes, which require noncentral distributions.

Some statisticians continue to believe that SST has utility and should be retained. For example, Onwuegbuzie and Levin (2003) recommend that both SST and effect sizes be used "in tandem to establish a reported outcome's believability and magnitude respectively" (p. 133). We assume that these authors define "believability" as the degree of certitude that the inferences obtained from the sample represent the underlying population, "leading to meaningful conclusions in which as many rival explanations as possible are eliminated" (p. 146). Consequently, they think inferential statistical procedures, such as SST, should be "retained as a gatekeeper for determining whether or not effect sizes should be interpreted" (p. 133). Therefore, Onwuegbuzie and Levin argue for the use of both statistical and substantive significance. Substantive significance is addressed by calculating effect sizes, confidence intervals, and, in concert with Onwuegbuzie and Levin, independent replications. This is a curious argument because, obviously, replication alone would negate the need for the establishing of statistical significance (Shaver, 1993). Substantive significance remains informative in that it is an economical means of unambiguously communicating the magnitude of effect and its practical range. But the heart of the matter is replication (Shaver) and extension (Swaminathan, personal communication, April, 2006). In a very real sense, science is built upon replication and extension, allowing for the accumulation and evolvement of knowledge and its application. Such is woefully absent in school psychological research.

In conclusion, Johnson's (1999) cautionary summary to biologists is clearly relevant to school psychologists. "Our work is important, so we should use the best tools we have available. Rarely, however, is that tool statistical hypothesis testing" (p. 771).

Other Problems in School Psychological Research

There is the common assumption that after conducting a carefully designed experiment that resulted in a pronounced effect, we can legitimately posit that the treatment caused the effect. In fact, all that we really can conclude is a reduction in the number of possible alternative explanations for the effect. Misattribution of what caused the effect can seriously impede theory development that is the foundation of the scientific advancement. Examples of this can be illustrated in research on children with selective mutism and the efficacy of psychotherapy.

Although several theoretical explanations have been offered to explain selective mutism, ranging from psychodynamic to behavioral theories, the definitive causes remain unknown (Kehle, Bray, & Theodore, 2006). However, contemporary treatments tend to be either behavior-based, pharmacological, or a combination of these. With respect to behavior theory, the child's mutism is assumed to be learned behavior that is shaped and maintained by the consequences of being selectively mute, and, consequently, treatments typically employ interventions that include contingency management, stimulus fading, shaping, self-modeling, and combined treatment strategies (Kehle et al.). The assumption is that the child learned to be selectively mute and therefore treatments should attempt to facilitate the child relearning appropriate speech in varied formerly problematic settings, particularly in the school.

To this end, an augmented self-modeling approach that incorporated numerous behavior-based learning strategies has been successfully employed in several cases to affect substantial and enduring change resulting in a complete cessation of the child's selective mutism (see Kehle, Madaus, Baratta, & Bray, 1998; Kehle, Cressy, & Owen, 1990, for a complete description of the self-modeling intervention). We assumed that the behavior/learning-based intervention caused the child to talk in settings where he or she was formerly mute. We may have been wrong. The self-modeling intervention employed edited videotapes depicting the child supposedly exhibiting expected and normal speech in formally problematic settings. The child views these brief 2–3-min tapes on five or six occasions. This procedure is almost identical to the procedures used in studies on the alteration of children's memories (e.g., Braum & Loftus, 1998; Garry & Gerrie, 2005; Loftus, 1997). A very tenable alternative explanation is that having children repeatedly view edited videotapes of themselves engaging in expected verbal behavior may have created false memories that they are not, or have not been, selectively mute. Preliminary results of one of our student's dissertation research currently being conducted appear to support this alternative explanation. Of course, the results of this research need to be independently replicated. If such is confirmed, the efficacy of treatment strategies that are designed to positively affect children's academic and social competence may be related to the extent that such treatments can alter the children's memories of their maladaptive behaviors and promote memories of adaptive behaviors.

Many other misattributions of treatment effects undoubtedly exist that affect and impede theory development that consequently have extremely important ramifications on student preparation, design of treatments, and policy. Perhaps one of the most significant examples is Wampold's (2001) eloquent and comprehensive argument supporting the efficacy of the contextual model over the medical model of psychotherapy. In brief, Wampold provided convincing evidence that all the myriad psychotherapeutic treatments, supposedly supported by widely disparate theoretical assumptions, achieve virtually indistinguishable positive effects. Consequently, he argued that there exists some other unacknowledged common core of curative factors that are efficacious. According to Wampold, "the evidence is clear, that the type of treatment is irrelevant, and adherence to a protocol is misguided, but yet the therapist, within each of the treatments, makes a tremendous difference" (p. 202). What this means is that the empirically supported treatment movement that demands specificity and adherence to treatment protocols, in concert with the medical model of psychotherapy, is not of practical value. Even with inspection of the specific components of the widely popular cognitive behavior therapy, there are no known treatment elements that have been shown to be efficacious (Wampold). Nevertheless, psychotherapy is effective for reasons other than the type of therapy that is employed. Wampold's strong belief in his arguments is mirrored in his statement, "it would be difficult to imagine how a scientist could examine these data and come to a different conclusion" (p. xiii). He noted that countless psychotherapy studies have resulted in a dearth of "important psychotherapeutic principles that have been scientifically established and generally accepted. . . . How is it that so much research has

yielded so little knowledge?" (p. 1). It is the result of misattributing psychotherapy's effectiveness to some specific theoretical approach that for all practical purposes masked the important but yet unacknowledged common elements that are inherent in all varied types of therapy.

Other rather common errors in school psychological research are the use correlational data to suggest causality and conduction of correlational research that omits relevant variables. The results of correlational studies, such as path analysis and structural equation modeling, that supposedly are intended to ascertain hypothetical causal relationships between the variables are often presented as confirmation of a theoretical position. The term "hypothetical" often fades in the discussion of the results. Correlational data, regardless of how it is analyzed, cannot evolve into statements of causality. There can be "No causation without manipulation" (Holland, 1986, p. 959). As Holland stated, "Put as bluntly and as contentiously as possible, . . . I take the position that causes are only those things that could . . . be treatments in experiments" (p. 954).

Correlational research can be of importance in that it affords testing of theories of hypothesized causal relationships between variables. However, its worth is related to the degree to which the variables are derived from psychometrically sound indices and relevant variables are not overlooked or omitted (Gall, Gall, & Borg, 2003).

However, it is atypical that theoretical models include all relevant variables. This, we believe, is practically impossible due to both an unawareness of all relevant variables and a reluctance to investigate the predictive power of variables that are politically sensitive. When relevant variables are omitted, conclusions are necessarily erroneous, particularly the relationships between variables.

Finally, there is a reluctance to determine whether or not an experimental effect endured. Rarely do studies published in the school psychology journals examine the longitudinal effect of interventions, reported in effect sizes and confidence intervals, beyond a single year.

Conclusion

It is quite apparent to us that school psychological practice has not evidenced any noticeable enduring change in promoting children's academic and social competencies. There is scant evidence of a cumulative knowledge base that can reliably inform the design and implementation of efficacious treatments that result in substantial and enduring betterment in children's learning or behaving. Can one honestly answer in the affirmative that children are better learners and more civil than they were 50 years ago? Similarly, Wampold's (2001) criticism of the psychotherapy literature, the voluminous heap of educational and school psychological studies, has resulted in very few important principles and effective interventions that have been scientifically established and generally accepted. How can so much research result in so little cumulative knowledge? In part, this may be due to ineffectual methods of inquiry and analyses, most appreciably statistical significance testing. Or perhaps Wampold's arguments supporting the contextual model of psychotherapy are also relevant to educational and school psychological practice. Are we also missing important but yet unacknowledged common elements that are inherent in all of the varied types of educational and school psychological practices?

To paraphrase Johnson (1999), the work of school psychologists is of utmost importance and we must use the best tools we have available. "Rarely, however, is that tool statistical hypothesis testing" (p. 771), and furthermore it "has created considerable damage as regards the cumulation of knowledge" (Thompson, 1992, p. 436).

References

American Psychological Association (2001). Publication manual of the American Psychological Association (5th ed.). Washington, DC: Author.

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

Barnard, G. (1998). Pooling populations. New Scientist, 157, 47.

Braum, K.A., & Loftus, E.F. (1998). Advertising's misinformation effect. Applied Cognitive Psychology, 12, 560–591.

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

Clark, C.A. (1963). Hypothesis testing in relation to statistical methodology. Review of Educational Research, 33, 455–473.

Cohen, J. (1994). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Deming, W.E. (1975). On probability as a basis for action. American Statistician, 29, 146–152.

Fidler, F., Thomason, N., Cummings, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. Psychological Science, 15, 119–126.

Gall, M.D., Gall, J.P., & Borg, W.R. (2003). Educational research: An introduction (7th ed.). Boston: Person Education.

Garry, M., & Gerrie, M.P. (2005). When photographs create false memories. Current Directions in Psychological Science, 14, 326–330.

Guttman, L. (1985). The illogic of statistical inference for cumulative science. American Stochastic Models and Data Analysis, 1, 3–10.

Holland, P.W. (1986). Statistics and causal inference. American Statistical Association Journal, 81, 945–960.

Johnson, D.H. (1999). The insignificance of statistical significance testing. Journal of Wildlife Management, 63, 763–772.

Kehle, T.J., Bray, M.A., & Theodore, L.A. (2006). Selective mutism. In G. Bear & K. Minke (Eds.), Children's needs III (pp. 293–302). Washington, DC: National Association of School Psychologists.

Kehle, T.J., Cressy, E.T., & Owen, S.V. (1990). The use of self-modeling as an intervention in school psychology: Case study of an elective mute. School Psychology Review, 19, 113–119.

Kehle, T.J., Madaus, M.M.R., Baratta, V.S., & Bray, M.A. (1998). Augmented self-modeling as a treatment for children with selective mutism. Journal of School Psychology, 36, 377–399.

Loftus, E.F. (1997). Memories for a past that never was. Current Directions in Psychological Science, 6, 60–65.

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

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.

Onwuegbuzie, A.J., & Levin, J.R. (2003). Without supporting statistical evidence, where would reported measures of substantive importance lead? To no good effect. Journal of Modern Applied Statistical Methods, 2, 133–151.

Rozeboom, W.W. (1997). Good science is abductive, not hypothetic-deductive. In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), What if there were no significance tests? (pp. 335–392). Mahwah, NJ: Erlbaum.

Shaver, J.P. (1993). What statistical testing is, and what it is not. Journal of Experimental Education, 61, 293–316.

Shrout, P. (1997). Should significance tests be banned? Introduction to a Special Section exploring the pros and cons. Psychological Science, 8, 1–2.

Smithson, M. (2003). Confidence intervals. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-140. Thousand Oaks, CA: Sage.

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. Journal of Counseling and Development, 70, 434–438.

Thompson, B. (2002). What future quantitative social science research should look like: Confidence intervals for effect sizes. Educational Researcher, 31, 25–32.

Volker, M.A. (2006). Reporting effect size estimates in school psychological research. Psychology in the Schools, 43, 653–672.

Wampold, B.E. (2001). The great psychotherapy debate: Models, methods, and findings. Mahwah, NJ: Erlbaum.