ED391990 1995-00-00 Inappropriate Statistical Practices in Counseling Research: Three Pointers for Readers of Research Literature. ERIC Digest.

ERIC Development Team

www.eric.ed.gov

Table of Contents

If you're viewing this document online, you can click any of the topics below to link directly to that section.

ERIC Identifier: ED391990
Publication Date: 1995-00-00
Author: Thompson, Bruce
Source: ERIC Clearinghouse on Counseling and Student Services Greensboro NC.

Inappropriate Statistical Practices in Counseling Research: Three Pointers for Readers of Research Literature. ERIC Digest.

The research literature provides important guidance to counselors working to keep abreast of the latest thinking regarding best practices and recently developed counseling tools. However, in my work as a former editor of Measurement and

Evaluation in Counseling and Development, and as Editor of Journal of Experimental Education and of Educational and Psychological Measurement, I have noticed some errors that seem to recur within the research literature read by counselors. The purpose of this digest is to highlight a few of these errors, and to provide some helpful references that further explore these problems. In "buying" the ideas presented within publications, as in buying more tangible products, the old maxim of caveat emptor does indeed remain useful.

1. "Insufficient Attention to Score Reliability" Pedhazur and Schmelkin (1991, pp. 2-3) recently noted that, "Measurement is the Achilles' heel of sociobehavioral research... [I]t is, therefore, not surprising that little or no attention is given to properties of measures used in many research studies." In fact, empirical studies of the published literature indicate that score reliability is not considered in between 40 and 50 percent of the published research. And, similarly, in doctoral dissertations we occasionally even see scores being analyzed that have reliability coefficients that are less than negative one (Thompson, 1994)!

The failure to consider score reliability adequately in substantive research is very serious, because effect sizes and power against Type II error are both attenuated by measurement error. Thus, prospectively we may plan and conduct studies that could not possibly yield noteworthy effect sizes, given that score unreliability inherently attenuates effect sizes. Or, retrospectively, we may not accurately interpret the effect sizes in completed studies if we do not consider as part of our interpretation the reliability of the scores we are actually analyzing.

Consumers of published research should generally expect authors to analyze the reliability of the scores in their own data. It is not sufficient even to report reliability coefficients from test manuals or from other research, because tests are NOT themselves reliable (i.e., tests are not imprinted both with ink and with reliability during the various stages of the printing process). Score reliability is influenced by various facets of the measurement process, including when, how, and to whom the test was administered. Thus, it becomes an oxymoron to speak of "the reliability of the test," because such a telegraphic shorthand way of speaking is also an incorrect way of speaking, i.e., makes an inherently untrue assertion.

Partly because this shorthand way of speaking is so common, too few researchers recognize that "reliability is a characteristic of scores" and not of tests. Because "scores" possess or lack these characteristics, different sets of scores generated by even the same measure may each have different reliabilities.

These telegraphic ways of speaking become problematic if we come unconsciously to ascribe literal truth to our shorthand, rather than recognizing that our jargon is

sometimes literally untrue. As noted elsewhere:

"This is not just an issue of sloppy speaking--the problem is that sometimes we unconsciously come to think what we say or what we hear, so that sloppy speaking does sometimes lead to a more pernicious outcome, sloppy thinking and sloppy practice (Thompson, 1992, p. 436)."

Readers of published research should expect authors only to offer assertions that they reasonably believe are true, and thus we should not condone use of the language, "the test is reliable." Furthermore, we should expect authors of published research to offer empirical evidence that the scores they are actually analyzing have reasonable measurement integrity.

2. "Overreliance on Tests of Statistical Significance"

The business of science is identifying relationships that recur under stated conditions. Unhappily, too many researchers at least unconsciously incorrectly assume that the p values calculated in statistical significance tests evaluate the probability that results will recur (Carver, 1993).

To get a single estimate of the p(robability) of the sample statistics, the null hypothesis is posited to be exactly true in the population. Thus, statistical significance testing evaluates "the probability of the sample statistics for the data in hand, given that null hypothesis about the related parameters in the population is presumed to be exactly true." This is "not" a test of result replicability, i.e., is "not" a test of whether roughly equivalent effect sizes would be detected in subsequent studies conducted under similar conditions!

In fact, the requirement that statistical significance testing must presume an assumption that the null hypothesis is true in the population is a requirement that an untruth be presumed. As Meehl (1978, p. 822) notes, "As I believe is generally recognized by statisticians today and by thoughtful social scientists, the null hypothesis, taken literally, is always false." Similarly, Hays (1981, p. 293) points out that "[t]here is surely nothing on earth that is completely independent of anything else [in the population]. The strength of association may approach zero, but it should seldom or never be exactly zero."

And positing an untruth about the population has a very important implication. Whenever the null is not "exactly" true in the sample(s), then the null hypothesis will "always" be rejected at some sample size. As Hays (1981, p. 293) emphasizes, "virtually any study can be made to show significant results if one uses enough subjects."

Although statistical significance is a function of several different design features, sample size is a basic influence on statistical significance. Thus, statistical significance testing can create a tautology in which we invest energy to determine that which we already know, i.e., our sample size.

Consumers of published research should expect authors to never say "significant" when they mean "statistically significant." Since statistical significance does not evaluate result importance, "always" using the phrase "statistically significant" when referring to statistical tests helps somewhat to avoid confusing statistical significance with the issue of importance. As Thompson (1993) emphasized:

"Statistics can be employed to evaluate the probability of an event. But importance is a question of human values, and math cannot be employed as an atavistic escape (a la Fromme's Escape from Freedom) from the existential human responsibility for making value judgments. 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. Like it or not, empirical science is inescapably a subjective business. (p. 365)"

Second, it is important to expect authors reporting statistical significance to supplement these tests with analyses that do focus on result importance and on result replicability. With respect to result importance, authors should be expected to report and interpret effect sizes. Even the recently published fourth edition APA style manual acknowledges that probability values reflect sample size, and thus encourages all authors to provide effect-size information.

With respect to result replicability, authors should be expected to report actual, so-called "external" replication studies, or to conduct "internal" replicability analyses. (Thompson, 1993, 1994b). The latter include cross-validation, the jackknife, and the bootstrap. These analyses, unlike statistical significance tests, do inform judgment about whether detected relationships replicate under stated conditions.

3. "Stepwise Methods Should Not Be Used"

Stepwise analyses are used with some frequency in published research, almost always to bad effect (cf. Thompson, 1994a). There are three problems. First, the computer packages use the wrong degrees of freedom in computing statistical significance in these analyses, and the incorrect degrees of freedom systematically bias the tests in favor of yielding statistical significance that is bogus. Second, not only does doing k steps of analysis "not" yield the best predictor set of size k, it can occur that "none" of the predictors entered in the first k steps are even among the best predictor set of size k. Third, because the linear sequence of entry decisions can be radically influenced by

sampling error, thus throwing the whole sequence of decisions off track at any step, and because so many decisions are made along the way of a stepwise analysis, stepwise analyses often produce results that are very unlikely to replicate!

REFERENCES

Carver, R. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61(4), 287-292.
Hays, W.L. (1981). Statistics (3rd ed.). New York: Holt, Rinehart and Winston.

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.

Pedhazur, E.J., & Schmelkin, L.P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Erlbaum.

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

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. Journal of Experimental Education, 61(4), 361-377.

Thompson, B. (1994a, April). Common methodology mistakes in dissertations, revisited. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. ED forthcoming.

Thompson, B. (1994b). The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results. Journal of Personality, 62(2), 157-176.

Bruce Thompson is Professor of Education and Distinguished Research Fellow at Texas A&M University, and Adjunct Professor of Community Medicine at Baylor College of Medicine in Houston.

This publication was funded by the U.S. Department of Education, Office of Education Research and Improvement, Contract No. RR93002004. Opinions expressed in this report do not necessarily reflect the positions of the U.S. Department of Education, OERI, or ERIC/CASS.

Title: Inappropriate Statistical Practices in Counseling Research: Three Pointers for Readers of Research Literature. ERIC Digest.
Document Type: Information Analyses---ERIC Information Analysis Products (IAPs)

(071); Information Analyses---ERIC Digests (Selected) in Full Text (073);
Descriptors: Counseling, Educational Researchers, Evaluation Methods, Evaluation Problems, Research Design, Research Methodology, Research Problems, Scoring, Statistical Analysis, Statistics, Test Interpretation, Test Reliability, Test Use
Identifiers: ERIC Digests
###

—

[Return to ERIC Digest Search Page]