

# **Statistical Significance Tests and Effect Magnitude Measures Within Quantitative Research Manuscripts Published in the Journal Of Agricultural Education During 1996-2000**

Matthew T. Portillo, Oklahoma State University

## Abstract

Manuscripts published in the *Journal of Agricultural Education* are expected to comply with criterion established for mathematical and statistical copy set forth by the American Psychological Association (1994) and the editorial policies of the *Journal of Agricultural Education*. This paper described the reporting practices of 141 quantitative research manuscripts published in the *Journal of Agricultural Education* during the five year period, 1996-2000, concerning statistical significance tests and effect magnitude measures. Findings indicated quantitative research designs permeated most manuscripts. Findings also indicated statistical significance tests were utilized in almost half of the manuscripts to determine differences among variables where as, over half of these manuscripts failed to report any effect magnitude measures. Further findings indicated the proportion of effect magnitude measures were reported by less than one-third of all the manuscripts. The proportion of manuscripts utilizing statistical significance tests and reporting any effect magnitude measure were reported by less than one-fourth of all the manuscripts. It was recommended that researchers utilizing statistical significance tests include effect magnitude measures, power analysis, confidence intervals, and the adoption and enforcement of more strict editorial policies regarding statistical significance testing and effect magnitude measures. It was further recommended that Agricultural Education researchers review the debate between hypothesis testing versus effect size and to review these two statistical methods.

## Introduction

Scholars have been conducting statistical testing for research purposes since the early 1700s, (McLean & Ernest, 1998). Descriptive and inferential statistics are the tools researchers use to analyze their research. The role of statistical significance testing in educational research has been the subject of much controversy recently (Kaufman, 1998; Knapp, 1998; Levin, 1998; McLean & Ernest, 1998; Nix & Barnette, 1998; Thompson, 1998). As early as 1931, R. W. Tyler noted the misuse of statistical significance, “. . . we are prone to conceive of statistical significance as equivalent to social significance. These two terms are essentially different and ought not to be confused” (cited in Daniel, 1998a, p. 24). Berkson (1942), Yates (1951), Kish (1959), and Kerlinger (1979) also lamented that too much emphasis was placed upon statistical significance tests as the end all product (Daniel, 1998a). Not until Cohen’s (1962) first inquiry into how much power ( $1-b$ ) did typical published research studies contain result in any serious examination of statistical significance versus practical significance. Since Cohen (1962), a cacophony of use and misuse of statistical significance tests has become a major methodological paradigm in journals of the social sciences (Brewer, 1972; Daniel, 1998a; Fern & Monroe, 1996; Thompson, 1999, March; Thompson & Snyder, 1997; Wilkerson & Olson, 1997).

The legacies of Sir Ronald Fisher (concerning differences of between and within groups using probability levels) and Karl Pearson (concerning correlation analyses providing indices of association) are two important approaches of statistical testing and how statistical analyses have developed (McLean & Ernest, 1998; Nix & Barnette, 1998). This paper explains the cogency in reporting effect magnitude measures along with statistical significance tests and examines this relationship in quantitative research manuscripts published within the *Journal of Agricultural Education* during 1996-2000.

### Editorial Policies

Utilizing the *Publication Manual of the American Psychological Association (1994)* as a source of clear communication for authors submitting manuscripts to the *Journal of Agricultural Education*, the American Association for Agricultural Education accepted an agreed upon style for standards of content and form when reporting statistical and mathematical copy. The American Psychological Association (1994) “encouraged” that authors of manuscripts using inferential statistics “include sufficient information to help the reader corroborate the analyses conducted” (p. 16). Moreover, when reporting inferential statistics, authors of manuscripts should “include information about the obtained magnitude or value of the test” (p. 15).

In order to determine if the statistical significant tests are of any practical significance, Vasquez, Gangstead, and Henson (2000) reiterated that journals in the education field require authors to report the relative treatment magnitude along with the statistical significance test. The *Journal of Agricultural Education* along with *Educational and Psychological Measurement*, *Journal of Applied Psychology*, *Journal of Consulting and Clinical Psychology*, *Journal of Experimental Education*, *Journal of Learning Disabilities*, *Language Learning*, *The Professional Educator*, and *Research in the Schools*, have adopted editorial policies regarding statistical significance tests and effect sizes (Thompson, 2000). Specifically, the guidelines for authors set forth by Kotrlik (2000) on the inside cover of the *Journal of Agricultural Education* reads, “Authors should report effect sizes in the manuscript and tables when reporting statistical significance.” For consumers to interpret research within the *Journal of Agricultural Education*, it is reasonable to provide evidence that an event did not happen by chance. Moreover, is it not also reasonable to desire research that is meaningful or practical and an event that is replicable?

### Statistical Significance Tests

Thompson (1994) asked what does the concept of statistical significance testing mean? “Too few researchers understand what statistical significance testing does and doesn’t do, and consequently their results are misinterpreted” (p. 1). Statistical significance tests determine whether or not a difference exists between variables (Rea & Parker, 1997).

To fully understand the concept of statistical significant testing, a review of Fisher’s single binary null hypothesis is warranted. The null hypothesis ( $H_0$ ) implies that there is no difference in the two population means. Researchers such as Bakan (1966), Cohen (1988), Hinkle, Wiersma, & Jurs (1994) have called this difference, the hypothesis of no relation or no difference (cited in Nix & Barnette, 1998). At this point, it should be noted that Fisher did not develop or support the alternative hypothesis (Nix & Barnette, 1998). Further development of

Fisher's null hypothesis resulted in the null hypothesis indicating direction (e.g.,  $\mu_1 \leq \mu_2$  or  $\mu_1 \geq \mu_2$ ). Conversely, the research question or the alternative hypothesis ( $H_1$ ) indicates that there is a difference between two population means (e.g.,  $\mu_1 \neq \mu_2$ ) and this hypothesis may also be directional (e.g.,  $\mu_1 > \mu_2$  or  $\mu_1 < \mu_2$ ).

To correctly interpret the results of null hypothesis significance testing, an understanding of the two types of inferential error that might occur, based on Fisher's p-value as the strength of the statistic developed by Neyman and Pearson, is needed (Nix & Barnette, 1998). A Type I error involves rejecting the null hypothesis when the null hypothesis is in fact true (a false positive, e.g., the treatment was effective when it was not). The probability of making a Type I error is equal to the value of the researcher's selected alpha ( $\alpha$ ) level. If the researcher chooses an alpha level equal to .05, the probability of committing a Type I error is five times out of one hundred. Therefore, as the researcher lowers the alpha level, the probability of committing a Type I error is lowered. However, as the researcher lowers the probability of committing a Type I error, the researcher then sacrifices the power of the test.

The power of the test ( $1-b$ ) "is the probability that a test statistic will find statistical significance" (Rossi, 1997, p. 177, cited in Nix & Barnette, 1998). Pearson and Hartley (1951) developed power charts to aid the researcher (cited in Hinkle & Oliver, 1983). A test with power of .80 indicates the researcher would have an 80 percent chance of finding statistical significance. Since power is defined as  $1-b$ , beta ( $b$ ) represents Type II error. When the researcher accepts the null hypothesis and the null hypothesis is false, a false negative results (e.g., no treatment effect present when there was).

### Effect Magnitude Measures

Awash in a sea of terminology, researchers use different terms to refer to effect magnitude measures as effect size, percent of variance accounted for, strength of association, measure of association, relative treatment magnitude, or magnitude of effect (Plucker, 1997). Effect magnitude measures (Nix & Barnette, 1998) can be classified first as measures of strength of association. "Measures of association reflect the strength of the relationship between two or more variables. They are single-summary statistics that augment the analysis of contingency tables and provide information to supplement the results of statistical significance tests" (Rea & Parker, 1997; see also Hinkle & Oliver, 1983). Furthermore, the magnitude of the effect statistic tells the researcher the degree to which the "dependent variable is controlled, predicted, or explained by the independent variable" (Mahadevan, 2000, p. 19). Secondly, effect magnitude measures can be classified as measures of effect size involving differences between group means. "Any mean difference index, estimated effect parameter indices, or standardized difference between means qualify as measures of effect size" (Nix & Barnette, 1998, p. 8). Together, measures of strength of association and measures of effect size provide the consumer of the research with the practical significance of the research. Robinson and Levin (1997) succinctly stated "First convince us that a finding is *not due to chance*, and only then, assess how *impressive* it is" (cited in McLean & Ernest, 1998, p. 18, italics in original).

## Statistical Significance Tests vs. Effect Magnitude Measures

Objections to null hypothesis statistical testing (NHST) “have provided compelling evidence that NHST has serious limiting flaws that many educators and researchers are either unaware of or have chosen to ignore” (Nix & Barnette, 1998). Debate over the value of statistical significance tests center around three areas of criticism: 1) the logic of null hypothesis testing; 2) the interpretation of null hypothesis statistical tests; and 3) the use of alternative and/or supplementary methods of inference testing (Ernest & McLean, 1998, see also Daniel, 1998a,b; Knapp, 1998; Levin, 1998; Nix & Barnette, 1998; Thompson, 1998).

Arguing vehemently against the logic of NHST ( $H_0: \mu_1 - \mu_2 = 0$ ), Bakan (1966) stated, “A glance at any set of statistics on total populations will quickly confirm the rarity of the null hypothesis in nature” (p. 5, cited in Nix & Barnette, 1998). Nix and Barnette (1998) reiterated this point, “The test of differences in NHST posits an almost impossible situation where the null hypothesis differences will be exactly zero” (p. 5). If in a study, failure to reject the null hypothesis results, the researcher is faced with a double-edged sword. One edge of the blade says either Bakan and Cohen are correct about NHST or the researcher must return to examine errors in NHST, where as these errors may include treatment differences, measurement error, and/or sampling error. As a result, many researchers have found that by increasing sample size, their findings have a greater chance of resulting in statistical significance and thus, the likelihood of a published manuscript. Here is where the researcher gets cut on the other side of the blade. If the researcher increases the power of the test, it becomes increasingly more difficult to detect statistical significance. However, if the researcher increases the sample size to achieve a higher level of power, any differences thus become statistically significant no matter how small. To counterbalance this dilemma, proponents of effect magnitude measures encourage reporting measures of association or effect size to reveal whether the results yield a practical significance.

Maxwell, Camp, & Arvey (1981) suggested the “primary advantage of measures of strength of association is that they have the potential to reveal whether a statistically significant result reflects a meaningful rather than a trivial experimental effect” (p. 525). Critics see statistical significance testing as nothing more than a numbers game where researchers are only concerned with reporting only statistically significant results even when the results were not of any practical importance (Daniel, 1997; see also Fan, 1999; Hess & Olejnik, 1997; Hinkle & Oliver, 1983; McLean & Kaufman, 1998; Thompson, 1987; Vacha-Haase & Nilsson, 1998).

When researchers solely rely on statistical significance testing, either using the observed significance level (p-value) or test statistics like  $F$ ,  $t$ , or  $\chi^2$ , the researcher may be distracted from more important considerations like result importance or value, result replicability, and result magnitude or effect (McLean & Ernest, 1998; Thompson, 1999, March). Thompson and Snyder (1997) described researchers use language like “significance” when they meant “statistically significant” resulting in misleading uses of the wording. Brewer (1972) found that journals in behavioral sciences tended to overwhelmingly report “significant” results to mean a rejection of the null hypothesis even with a small effect size. “The implication of this response is that regardless of how small the effect is, they want to detect it, i.e., small ES”

(p. 394). Thompson (1987) reported reliance on statistical significance testing has inadvertently led to a bias against reporting statistical non-significant results thereby creating misinterpretations of statistical significant results (see also Hetrick, 1999).

Addressing the interpretation of NHST, Plucker (1997) conceptualized the misinterpretations of statistical significance testing as analogous to standing on the edge of a deep chasm. If an individual desires to cross the chasm (*e.g.*, the p-level), it is therefore important to the individual to find out the size of the chasm (*e.g.*, the effect size) before crossing. Is the chasm 10 inches or 100 feet? Plucker explained that determining the chasm's existence is important, but by doing so provides no information about the size of the chasm. Therefore, researchers reporting the relationship between the independent and dependent variables will allow the consumer of the research to determine the "practical significance" of jumping over the chasm (see also Daniel, 1997; Fan, 1999; Keppel, 1991; Kieffer & Thompson, 1999).

Lastly, the use of alternative and/or supplementary methods of inference testing can be best described by the statistician's motto, "In God we trust. All others bring data" (Claypool, 2001). Researchers arguing against statistical significance tests state that not enough information is being provided to the consumer of the research. Nix and Barnette (1998) reported researchers failed to tell readers if the assumptions of a statistical test have been satisfied or tested. "For research to be valuable it must be precise and as unambiguous as possible so that is (sic) can be comprehended (sic) by practitioners as well as other researchers" (p. 56). Thompson (1995, November) reported that when statistical significance was obtained, many researchers simply concluded the analysis. However, the analysis should continue to determine if the statistical significance was due to sampling error or effect size (cited in Nix & Barnette, 1988).

In fairness to proponents of statistical significance tests, Levin (1998) shouted, "Show me the data!" with respect to reporting effect sizes (p. 46). As Levin pointed out, if reporting effect sizes are going to change the world, then the researcher is remiss not to report any biases inherent in the researcher when reporting statistics.

### Purpose/Objectives

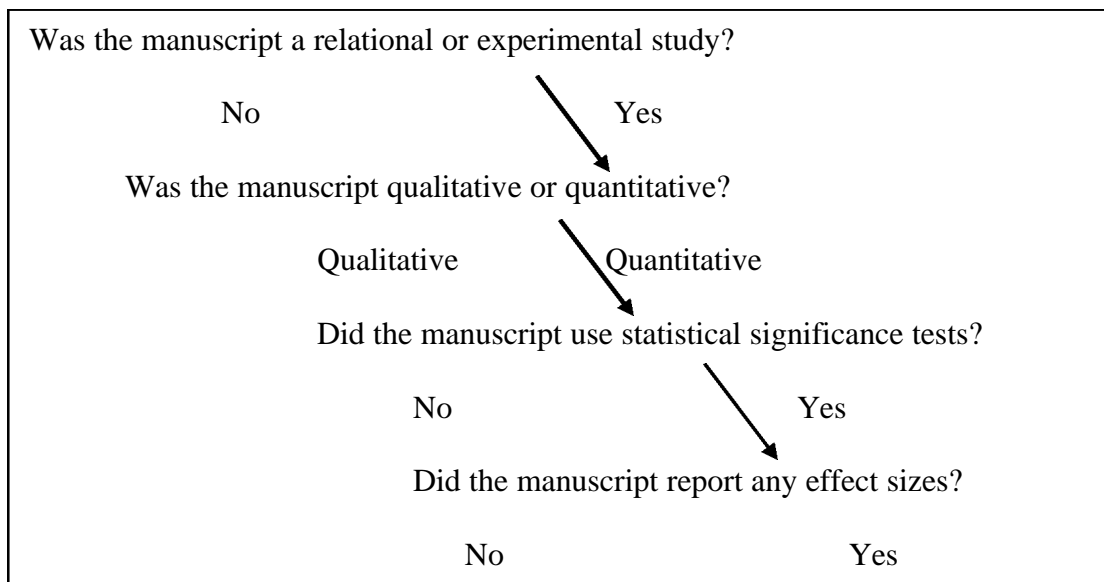
The purpose of this study was to describe the reporting practices of effect magnitude measures and statistical significance testing within quantitative research manuscripts published in the *Journal of Agricultural Education* during 1996-2000. It was also the purpose of this study to determine the relationship between effect magnitude measures and statistical significance testing within quantitative research manuscripts published in the *Journal of Agricultural Education* during 1996-2000. To accomplish the purpose of this study, the following objectives were established:

- 1) To describe reporting practices of effect magnitude measures within quantitative research manuscripts published in the *Journal of Agricultural Education* during 1996-2000.
- 2) To describe reporting practices of statistical significance testing within quantitative research manuscripts published in the *Journal of Agricultural Education* during 1996-2000.

- 3) To determine the relationship between reporting practices of effect magnitude measures and statistical significance tests within the *Journal of Agricultural Education* during 1996-2000.

### Methods

The population of this descriptive study included all 171 manuscripts published in the *Journal of Agricultural Education* during 1996-2000. All published manuscripts were classified via a dichotomous key (see Figure I). Each manuscript was content analyzed with respect to the type of research. Subsequently, manuscripts classified according to relational (sample survey) or experimental research were solely used. In addition, only quantitative manuscripts were germane to this study (n = 141). If a manuscript consisted of qualitative and quantitative research, the manuscript was coded as quantitative. Further examination of the manuscripts involved an analysis of statistical significance and a measure of strength of the association.



**Figure I.** Dichotomous key used to classify quantitative research manuscripts in the *Journal of Agricultural Education*.

### Analysis of Data

All data were analyzed using the Statistical Package for the Social Sciences (SPSS) for Windows version 8.0. A 2 x 2 Chi square ( $\chi^2$ ) test of significance and the Yates' correction for continuity were used to test the frequency differences. An unbiased estimator for measures of association, phi ( $\phi$ ), was used to measure the strength of the association due to each variable containing only two categories. Interpretation of phi ( $\phi$ ) was based on Rea and Parker (1997). Measures of strength of the association were interpreted as negligible (.00 to .09), weak (.10 to .19), moderate (.20 to .39), relatively strong (.40 to .59), strong (.60 to .79), and very strong (.80 to 1.00). The probability of committing a Type 1 error was set at .05, *a priori*.

## Results

### Objectives 1 and 2

From the 171 manuscripts published in the *Journal of Agricultural Education* during 1996-2000, 149 manuscripts consisted of a relational or experimental design, which accounted for 87.1 percent of the total manuscripts published. The remaining 22 manuscripts consisted of distinguished lectures, research syntheses of literature, or philosophical concerns which accounted for 12.9 percent. From these 149 manuscripts, eight manuscripts utilized solely qualitative research methods, which accounted for 5.4 percent. The remaining 141 manuscripts utilized quantitative research methods or a combination thereof, which accounted for 94.6 percent. From the 141 quantitative research manuscripts, 65 manuscripts utilized statistical significance tests, which accounted for 46.1 percent. Thus, the remaining 76 manuscripts utilized no statistical significance tests, which accounted for 53.9 percent. At 95 percent confidence, the data indicated that the proportion of all manuscripts will utilize statistical significance tests was between 38.7 percent to 53.5 percent ( $t_{(.05, 141)} = 1.645$ ). From the 141 quantitative research manuscripts again, 41 manuscripts reported one or more effect magnitude measures, which accounted for 29.1 percent (see Table I for the frequency and type of effect magnitude measures reported). Therefore, the remaining 100 quantitative research manuscripts reported no effect magnitude measures, which accounted for 70.9 percent. At 95 percent confidence, the data indicated that the proportion of all manuscripts will report effect magnitude measures was between 21.7 percent to 36.5 percent ( $t_{(.05, 141)} = 1.645$ ).

Table I

#### Frequency and type of effect magnitude measures reported in the *Journal of Agricultural Education* during 1996-2000

Type	Frequency
Spearman Rho	5
Pearson Product Moment	24
R <sup>2</sup>	5
R <sup>2</sup> adjusted	10
eta <sup>2</sup>	1
phi (φ)	4
Cramer's V	1
canonical correlation	4
point biseral	11
Hodges' g	1

### Objective 3

The total numbers of manuscripts utilizing statistical significance tests that reported one or more type of effect magnitude measure were 29 manuscripts. The total number of manuscripts utilizing statistical significance tests, but did not report any type of effect magnitude measure

were 36 manuscripts. The total numbers of manuscripts not utilizing statistical significance tests, but reported one of more type of effect magnitude measures were 12 manuscripts. The total numbers of manuscripts not utilizing statistical significance tests that reported no type of effect magnitude measures were 64 manuscripts (see Figure II). A 2x2 Chi square ( $\chi^2$ ) test of significance and a Yates' correction for continuity showed a statistical significant difference between effect magnitude measures and statistical significance tests [ $\chi^2(1, N = 141) = 12.753, p < .000, f = .316$ ]. The amount of variation accounted for between variables was 31.6 percent, representing the strength of association as moderate between effect magnitude measures and statistical significance tests. Furthermore, the data indicated at 95 percent confidence that the proportion of all manuscripts utilizing statistical significance tests was greater than the proportion of all manuscripts reporting effect magnitude measures by eight percent to 26 percent ( $t_{(.05, 141)} = 1.645$ ). Lastly, the data indicated at 95 percent confidence that the proportion of all manuscripts utilizing statistical significance tests and reporting effect magnitude measures was 15.3 percent to 25.9 percent ( $t_{(.05, 141)} = 1.645$ ).

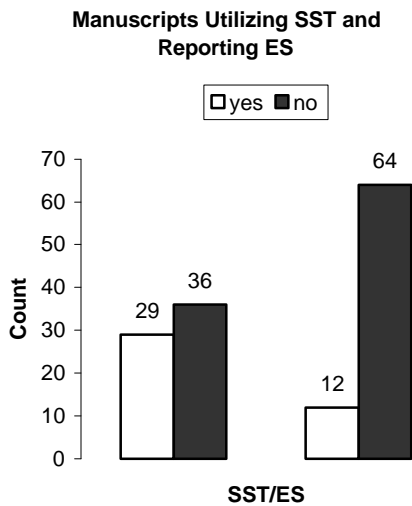


Figure II. Manuscripts utilizing statistical significance test and reporting effect sizes.

### Conclusions

Conclusions of the findings indicated most manuscripts published in the *Journal of Agricultural Education* during 1996-2000 involved relational or experimental methods. In addition, quantitative research designs permeated most manuscripts. Statistical significance testing was utilized in almost half of the manuscripts to determine differences among variables. However, over half of manuscripts utilizing statistical significance tests did not include any type of effect magnitude measures to support a practical significance of their findings to the reader. This finding supported Plucker's (1997) findings about the absence of effect size estimates when statistical significance tests were used within three different research journals in gifted educational research. In addition, this finding also contrasted Robinson and Levin (1997) when these manuscripts first convinced the reader that the finding was not due to chance and then failed to show the reader how impressive the study was (cited in McLean & Ernest, 1998).

An overwhelmingly proportion of manuscripts, reporting one or more types of effect magnitude measures, utilized no statistical significant tests. This finding concurred with Carver (1978), Meehl (1978), Schmidt (1996), and Shulman (1970) who advocated the complete abandonment of statistical significance testing as a method of evaluating statistical results (cited in Daniel, 1998). Furthermore, the findings of this study showed that the proportion of all manuscripts utilizing statistical significance tests was greater than the proportion of all manuscripts reporting effect magnitude measures. The findings also indicated that the proportion of all manuscripts utilizing statistical significance tests in conjunction with reporting effect magnitude measures was limited to a quarter of the manuscripts published or less.

### Recommendations

The following recommendations were based on the results of this study:

1. If the goal of scientific inquiry is to determine if the results of a test have any practical importance, it is recommended that all quantitative research utilizing statistical significance testing report an effect magnitude measure to highlight the distinction between statistical and practical significance.
2. The adoption and enforcement of more strict editorial policies regarding the reporting of the results of statistical significance testing and effect magnitude measures will perhaps eventually move the field toward improved practice. Editorial policies in the *Journal of Agricultural Education* should (a) require authors to index results of statistical significance tests to sample size, (b) require effect magnitude measures with statistical significance tests, (c) encourage Type II error analyses and confidence intervals, and (d) in cases of statistically non-significant results researchers should consider conducting statistical power analyses (Daniel, 1998b).
3. It is further recommended that readers of this paper review many of the citations made to achieve a full understanding between the debate of statistical significance tests and effect magnitude measures. Numerous effect magnitude measure formulas are available in Fern & Monroe (1996), Hetrick (1999), and Thompson (1999, in press).
4. Because “researchers are slow to adopt approaches in which they were not trained originally” (McLean & Ernest, 1998, p. 16), it is recommend that Agricultural Education researchers periodically review statistical methods. Miller (1998) suggested that if statistics are the tools of the researcher, we, as researchers then need to know our tools. “Tractor mechanics, artists, and masons have their tools and they must know how to use them. We, likewise, need to know how to use ours. You are challenged to get “checked-out” again on your tools; that is, devote some of your personal in-service or professional development time to renewing, maintaining, and improving your skills” (p. 1).

### References

Brewer, J. K. (1972). On the power of statistical tests in the American Educational Research Journal. *American Educational Research Journal*, 9, 391-401.

Claypool, P. L. (2001, Spring). Statistics 5013: Statistical experimenters 1. Lecture notes. Oklahoma State University: Stillwater.

Cohen, J. (1962). The statistical power of abnormal social psychology research. Journal of Abnormal and Social Psychology, 65 (3), 145-153.

Daniel, L. G. (1997). Statistical significance testing in “Educational and Psychological Measurement” and other journals. Paper presented at the Annual Meeting of the National Council on Measurement in Education. Chicago: IL.

Daniel, L. G. (1998a). Statistical significance testing: A historical overview of misuse and misinterpretation with implication for the editorial policies of education journals. Research in the Schools, 5 (2), 23-32.

Daniel, L. G. (1998b). The statistical significance controversy is definitely not over: A rejoinder to responses by Thompson, Knapp, and Levin. Research in the Schools, 5 (2), 63-65.

Ernest, J. M. & McLean, J. E. (1998). Fight the food fight: A response to Thompson, Knapp, and Levin. Research in the Schools, 5 (2), 59-62.

Fan, X. (1999). Statistical significance and effect size: Two sides of a coin. Paper presented at the Annual Meeting of the American Evaluation Association. Orlando: FL.

Fern, E. F. & Monroe, K. B. (1996, September). Effect-size estimates: issues and problems in interpretation. Journal of Consumer Research, 23 (2), 89-106.

Hess, B., & Olejnik, S. (1997). Top ten reasons why most omnibus ANOVA F-test should be abandoned. Journal of Vocational Education Research, 22 (4), 219-232.

Hetrick, S. (1999). A primer on effect sizes: What they are and how to compute them. Paper presented at the Annual Meeting of the Southwest Educational Research Association. San Antonio: TX.

Hinkle, D. E. & Oliver, J. D. (1983). How large should the sample be? A question with no simple answer? Or . . . Educational and Psychological Measurement, 43 (4), 1051-1060.

Kaufman, A. S. (1998). (ed.). Introduction to the special issue on statistical significance testing. Research in the Schools, 5 (2), 1.

Keppel, G. (1991). Design and analysis: A researcher's handbook (3rd ed.). Upper Saddle River, NJ: Simon & Schuster Company.

Kieffer, K. M. & Thompson, B. (1999). Interpreting statistical significance test results: A proposed new “What If” method. Paper presented at the Annual Meeting of the Mid-South Educational Research Association. Point Clear: AL.

Knapp, T. R. (1998). Comments on the statistical significance testing articles. Research in the Schools, 5 (2), 39-41.

- Kotrlik, J. W. (2000). (ed.). Journal of Agricultural Education. 41 (1).
- Levin, J. R. (1998). What if there were no more bickering about statistical significance tests? Research in the Schools, 5 (2), 43-53.
- Mahadevan, L. (2000). The effect size statistic: Overview of various choices. Paper presented at the Annual Meeting of the Southwest Educational Research Association. Dallas: TX.
- Maxwell, S. E., Camp, C. J., & Avery, R. D. (1981). Measures of strength of association: A comparative examination. Journal of Applied Psychology, 66, 525-534.
- McLean, J. E. & Ernest, J. M. (1998). The role of statistical significance testing in educational research. Research in the Schools, 5 (2), 15-22.
- Miller, L. H. (1997). Appropriate analysis. Journal of Agricultural Education, 39 (2), 1-10.
- Nix, T. W. & Barnette, J. J. (1998). A review of hypothesis testing revisited: Rejoinder to Thompson, Knapp, and Levin. Research in the Schools, 5 (2), 55-57.
- Plucker, J. A. (1997). Debunking the myth of the “Highly Significant” result: Effect sizes in Gifted Education Research. Roeper Review, 20 (2), 122-126.
- Publication manual of the American Psychological Association. (1994). (4<sup>th</sup> ed.). Washington, DC: American Psychological Association.
- Rea, L. M. & Parker, R. A. (1997). Designing and conducting survey research: A comprehensive guide. (2<sup>nd</sup> ed.). San Francisco, CA: Jossey-Bass Publishers.
- Thompson, B. (1987). The use (and misuse) of statistical significance testing: Some recommendations for improved editorial policy and practice. Paper presented at the Annual Meeting of the American Educational Research Association. Washington: DC.
- Thompson, B. (1994). The concept of statistical significance testing. Office of Educational Research and Improvement, U.S. Department of Education. ERIC/AE Digest: ED366654. [On-line]. Available: <http://ericae.net/edo/ED366654>
- Thompson, B. (1998). Statistical significance and effect size reporting: Portrait of a possible future. Research in the Schools, 5 (2), 33-38.
- Thompson, B. (1999, March). Why ‘encouraging’ effect size reporting is not working: The etiology of researcher resistance to changing practices. The Journal of Psychology, 133 (2), 133-141.

Thompson, B. (1999, Spring). Improving research clarity and usefulness with effect size indices as supplements to statistical significance tests. Exceptional Children, 65 (3), 329.

Thompson, B. (1999, in press). Common methodology mistakes in educational research, revisited, along with a primer on both effect sizes and the bootstrap. [On-line]. Available: <http://www.coe.tamu.edu/~bthompson/aeraad99.htm>

Thompson, B. (2000). Various editorial policies regarding statistical significance tests and effect sizes. [On-line]. Available: <http://www.coe.tamu.edu/~bthompson/journals.htm>

Thompson, B. & Synder, P. A. (1997). Statistical significance testing practices in The Journal of Experimental Education. The Journal of Experimental Education, 66 (1), 75-79.

Vacha-Haase, T. & Nilsson, J. E. (1998, April). Statistical significance reporting: Current trends and uses in MECD. Measurement and Evaluation in Counseling and Development 31 (1), 46-57.

Vasquez, L. M., Gangstead, S. K., & Henson, R. K. (2000). Understanding and interpreting effect size measures in general linear model analysis. Paper presented at the Annual Meeting of the Southwest Education Research Association. Dallas: TX.

Wilkerson, M. & Olson, M. R. (1997, November). Misconceptions about sample size, statistical significance, and treatment effect. The Journal of Psychology, 131 (6) 627-632.