DOCUMENT RESUME

ED 454 409 CE 081 918

AUTHOR Gordon, Howard R. D.
TITLE American Vocational Education Research Association Members' Perceptions of Statistical Significance Tests and Other Statistical Controversies.
PUB DATE 2001-03-08
NOTE 27p.; Paper presented at the Annual Community of Scholars Symposium in Workforce Development and Education (2nd, Columbus, OH, March 8, 2001).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Educational Research; Postsecondary Education; Predictor Variables; Research Methodology; Researchers; Scores; Secondary Education; *Statistical Analysis; *Statistical Significance; *Statistical Studies; Statistical Surveys; Test Interpretation; *Test Validity; *Testing Problems; Vocational Education
IDENTIFIERS American Vocational Education Research Association; Stepwise Regression

ABSTRACT

A random sample of 113 members of the American Vocational Education Research Association (AVERA) was surveyed to obtain baseline information regarding AVERA members' perceptions of statistical significance tests. The Psychometrics Group Instrument was used to collect data from participants. Of those surveyed, 67% were male, 93% had earned a doctoral degree, 67% had more than 15 years of experience in educational research, and 82.5% were employed at the university level. The respondents generally disagreed with the proposition that statistical significance tests should be banned. Stepwise methods were more likely to be perceived as acceptable for identifying the best variable set and importance, which suggested that some AVERA researchers are not aware that stepwise methods do not identify the best predictor set of a given size. Overall views regarding score reliability appeared to be "neutral." The respondents' general views regarding statistical testing were consistent with previous research. The responses suggested that the controversy over statistical testing has raised some consciousness among AVERA researchers' perceptions on the general views of statistical testing. It was recommended that future AVERA researchers be encouraged to always interpret effect sizes and conduct empirical investigations of the replicability of results. (Contains 63 references and 9 tables.) (MN)

American Vocational Education Research Association Members' Perceptions of Statistical Significance Tests and Other Statistical Controversies

Howard R.D.Gordon

Professor of Occupational Leadership/

Educational Research

Marshall University

Department of Adult and Technical Education

Huntington, WV 25755

Email: gordon@marshall.edu

This article was presented at The Ohio State University second annual Community of Scholars Symposium in Workforce Development and Education, March 8, 2001.

Abstract

The purpose of this study was to establish baseline information regarding AVERA members' perceptions of statistical significance tests. A simple random sample was utilized to select 113 AVERA members. The Psychometrics Group Instrument was used to collect data from participants. The findings showed that 67% of the respondents were males, 93% had earned a doctoral degree, 67% had more than 15 years of experience in educational research, and 82.5% were employed at the university level. There was general disagreement among respondents concerning the proposition that statistical significant tests should be banned. Views pertaining to stepwise methods were more likely to be perceived as acceptable for identifying the best variable. Overall, the findings suggest that the controversy has raised some consciousness among AVERA members' perceptions on the "general views" of statistical significance tests. Future AVERA researchers should be encouraged to always interpret effect sizes, and the replicability of the results should be empirically investigated.

American Vocational Education Research

Association Members' Perceptions of

Statistical Significance Tests and Other Statistical Controversies

Historically, vocational education (career and technical/workforce education) at the secondary and postsecondary levels has suffered from a "second-class citizen" image. This image has carried over into higher education. Departments of vocational teacher education at the university level have not always been held in the highest esteem. Whether merited or not, this stigma has been attached to research in vocational education. Research conducted in vocational education at the university often has been view as less than first-rate (Moore, 1992).

According to Moore (1992):

> We place too much emphasis on statistical significance and not enough emphasis on practical or applied significance of the research. We need to pay more attention to selecting problems for study. (p.11)

Educational research is an ongoing process, which starts at the determination of the problem followed by execution of research procedures (Gay, 1996). The subsequent stages of the process, including statistical analysis, are logically influenced by the nature of the research problem and the methodological strategy of a study.

During the past two decades, there has been an increase in vocational education research. The growth in vocational education research has been accompanied by an increase in the use of statistical techniques, with both positive and negative results. In a 1981 study by Oliver, some of the positive effects are described as: (a) more complex problems are being investigated, (b) the information produced is becoming more meaningful, and (c) the efficiency of the research is increasing. The negative effects primarily are that some problems and issues have arisen. Oliver

(1981) noted that "Statistical techniques are being used in cases where the assumptions are not being met and there is generally a failure to distinguish between statistical significance and practical importance" (p.9).

Conceptual Framework and Related Literature

The empirical-analytic paradigm of research in vocational education heavily relies on the use of statistics (Smith, 1984). The impact of statistical methods on vocational education research was recognized by many researchers in the field (Zhang, 1993; Cheek, 1988; Warmbrod, 1986; Oliver, 1981).

The use of statistics in educational research can be traced back as early as 1901 when Edward L. Thorndike published his Noted on Child Study (Walker, 1956). However, it was amend 1949 that "the era of empirical generalization" finally arrived in educational research (West & Robinson, 1980). In spite of frequent calls from many researchers in vocational education to broaden paradigms for inquiry (Zhang 1993), quantitative research still prevailed in the field during the 1980s (Lynch, 1983; Hillison, 1989).

Previous studies concurred that ANOVA, correlations, t-tests, regression, chi-square tests, and multivariate techniques were among the most frequently used technique in educational research (Zhang, 1993). The use of variations on statistical significance tests was popularized in the social sciences by Sir Ronald Fisher, Jerzy Neyman, and Egon Person (Huberty, 1987). Today, most researchers implicitly employ some hybrid of the logics suggested by these three figures (Thompson, 1996).

The etiology of the propensity to conduct statistical significance tests can be traced to two dynamics. The first involves an unrecognized error in logic when consciously trying to be

scientific, whereas the second dynamic occurs as a frankly irrational process. These two dynamics undergirding continued emphasis on statistical tests must be understood is reform efforts are to be effective (Thompson, 1996).

Statistical significance testing has existed in some form for approximately 300 years (Daniel, 1998) and has served an important purpose in the advancement of inquiry in the social sciences. The controversy about the use or misuse of statistical significance testing that has been evident in the literature for the past 10 years has become the major methodological issue of our generation (Kaufman, 1998).

Bracey (1988) reminded us that "Statistical significance has nothing to do with meaningfulness" (p.257). Kupfersmid (1988) observed that "A... problem related to the meaningfulness of 'statistically significant' findings is that what is 'significant' in a meaningful sense may be contradictory" (p. 636). Tests of statistical significance are overused and misused in an attempt to make a poor or mediocre study appear good (Moore, 1992).

Why do educational researchers place such emphasis on statistical significance? Soltis (1984) provided a clue:

> Much of the social and behavioral sciences have developed their present forms by consciously seeking to imitate the methods and forms of the natural sciences, many educational researchers have tried to travel the same royal road to knowledge, legitimacy and status. (p. 6)

Shaver (1992) maintained that educational researchers insist on tests of statistical significance because they "provide a façade of scientism in research. For many in educational research, being quantitative is equated with being scientific... despite the fact that some scientists and many psychologists... have manage very well without inferential statistics." (p. 2)

Few researchers understand what statistical significance testing does and doesn't do, and consequently their results are misinterpreted. Even more commonly, researchers understand elements of statistical significance testing, but the concept is not integrated into their research (Thompson, 1994). For example, the influence of sample size on statistical significance may be acknowledge by a researcher, but this insight is not conveyed when interpreting results in a study with several thousand subjects. Because statistical significance tests have been so frequently misapplied, some reflective researchers (Carver, 1978; Meehl, 1978; Schmidt, 1996; Shulman, 1970) have recommended that statistical significance tests be completely abandoned as a method for evaluating statistical results.

Biskin (1998) argues that practical or clinical significance can be noteworthy even when results are not statistically significant. Conversely, he argues that even results are or would be statistically significant, at least in some such cases "the researcher's prime consideration should be effect size." Cohen (1977) has offered the following definitions of effect size for the behavioral sciences:

Small effect size: $r^2 = .01$

Medium effect size : $r^2 = .09$

Large effect size: $r^2 = .25$ (pp. 79-80)

Reporting effect sizes has three important benefits. First, reporting effects facilitates subsequent meta-analyses incorporating a given report. Second, effect size reporting creates a literature in which subsequent researchers can more easily formulate more specific study expectations by integrating the effects reported in related prior studies. Third, and perhaps most importantly, interpreting the effect sizes in a given study facilitates the evaluation of how a study's results fit into existing literature, the explicit assessment of how similar or dissimilar

results are across related studies, and potentially informs judgment regarding what study features contributed to similarities or differences in effects (Vacha-Haase, Nilsson, Reetz, Lance & Thompson, 2000).

Biskin (1998) reported that "as a research area matures," effect size should be deemed more important than statistical significance. Recent empirical studies of articles published since 1994 in psychology, counseling, special education, and general education suggest that merely "encouraging" effect size reporting (APA, 1994) has not appreciably affected actual reporting practices (Vacha-Haase & Thompson, 1998). Kotrlick (2000) reported that authors should report effect sizes in the manuscript and tables when reporting statistical significance in the Journal of Agricultural Education (the only career and technical/workforce education journal with this requirement).

According to Tryon (1998), the fact that statistical experts and investigators publishing in the best journals cannot consistently interpret the results of these analyses is extremely disturbing. Seventy-two years of education have resulted in minuscule, if any, progress toward correcting this situation. It is difficult to estimate the handicap that widespread, incorrect, and intractable use of a primary data analytic method has on a scientific discipline, but the deleterious effects are doubtless substantial...(p. 796).

Several empirical studies have shown that many researchers do not fully understand the statistical tests that they employ (Mittag & Thompson, 2000; Nelson, Rosenthal, & Rosnow, 1986; Oakes, 1986; Rosenthal & Gaito, 1963; Zuckerman, Hodgins, Zuckerman, & Rosenthal, 1993). In their AERA study on "statistical significance tests", Mittag and Thompson (2000) recommended that other national research associations conduct similar studies to resolve conflicting views related to the use of statistical tests.

At present, there is a dearth of information in the literature about the perceptions of career and technical / workforce education researchers toward "statistical significance tests." The significance of this study is to serve as a framework for promoting further discussion of controversial statistical issues among career and technical / workforce education researchers.

Purpose and Objectives

The primary purpose of this study was to establish baseline information regarding AVERA members' perceptions of statistical significance tests. The following objectives guided the study:

1. To explore current perceptions of AVERA members regarding statistical significance tests.
2. To determine perceptions of AVERA members regarding selected statistical issues, such as score reliability and stepwise methods.

Methodology

This study utilized quantitative descriptive research methodology. According to Gay and Airasian (2000), "quantitative descriptive studies are carried out to obtain information about the preferences, attitudes, practices, concerns, or interests of some groups of people" (p. 11).

Population and Sample

The population consisted of current AVERA members ($\underline{N}$ = 160) during the 2000-01 school year. The AVERA membership directory was used to identify the population. Using a formula suggested by Krejcie and Morgan (1970), a sample size of 113 AVERA members was

needed, based upon a 5% degree of accuracy and a 95% confidence level. A simple random sample was selected from the population using the random number generator in Microsoft Excel.

Instrumentation

The Psychometrics Group Instrument (Mittag, 1999) was used to determine participants' perceptions of statistical significance tests and other statistical issues. Likert-scale response categories for the 29-items ranged from disagree (1) to agree (5). The questionnaire was pilot tested for content validity and reliability. The reliability coefficient of the questionnaire (part II) was .89. Appropriateness and permission of the use of this instrument for the study was discussed with the author. Some items were reverse-worded so as to minimize response set influences. Mittag and Thompson (2000) recommend the recoding of reverse-worded items, so that higher scores have a consistent meaning.

Data Collection

Elements of Dillman's (2000) mail and internet surveys were utilized to achieve optimal return rate. Data collection began in October, and was concluded in December, 2000.

To control nonresponse error and maintain validity, early and late respondents were compared statistically (Ary, Jacobs & Razavieh, 1996). Research shows that nonrespondents are often similar to late respondents (Miller & Smith, 1983). A late respondent was classified as one who returned his or her questionnaire during December. Statistical tests revealed no differences between respondents. Respondents' data were compiled, yielding a total response rate of 35% (n = 40).

According to Kerlinger (1986, p. 380), survey mail response rates are often about 30%. The critical question when such response rates are realized is whether the respondents are still representative of the population to which the researcher wishes to generalize. Mittag and

Thompson (2000) reported that "response profiles should be analyzed to provide at least some insight regarding the issue(s)" (p. 15). Although the results of this study may not be generalized to the entire population of American Vocational Education Research Association members, the results can still provide valuable information for selected career and technical / workforce education researchers.

Data Analysis

Data were analyzed using the Statistical Package for the Social Sciences (SPSS Version 9.0 for Windows). Descriptive statistics were used to organize and summarize the data.

Findings

Demographic Characteristics

Sixty seven percent of the respondents were males. A majority of the respondents (93%) had earned a doctoral degree. Sixty percent of the respondents revealed that they had over 15 years of experience in educational research. The respondents' work settings were as follows: university (82.5%), school district (7.5%), business (5.0%) and other (5.0%).

Perception Clusters

The 29 items evaluated nine clusters of perceptions. Table 1 presents responses to the first five items, which measured *general perceptions* and the ongoing significance controversy.

Insert Table 1 about here

Respondents were in general agreement ($\underline{M}$ = 4.47, $\underline{SD}$ = .60) that this controversy is likely to continue for many years in the future. The respondents also "agreed" ($\underline{M}$=4.25, $\underline{SD}$=.87) that

researchers should use the phrase "statistically significant," rather than "significant," to describe their results. There was general disagreement (M=1.70, SD=.88) among respondents concerning the proposition that statistical significance tests should be banned.

Table 2 shows means and standard deviations of respondents' perceptions of the *General Linear Model* (GLM). Respondents "slightly disagreed" that regression could be used to test hypotheses about means. As reported in Table 2, respondents also "slightly disagreed" that all statistical analyses are correlational.

Insert Table 2 about here

Participants were asked whether *stepwise methods* identify the best variable set, and whether the results can be used to infer variable importance. As reported in Table 3, these two views were perceived by respondents as "neutral" to "slightly agreeable" (M=3.47 to 3.55).

Insert Table 3 about here

Table 4 shows respondents' perceptions of *score reliability*. Respondents were "slightly in agreement" with item 23 (M=3.62, SD=1.12). Item 23 addressed the influence of poor reliability of data on "effect sizes".

Insert Table 4 about here

Views regarding *Type I and Type II errors* are reported in Table 5. Respondents reported a mean rating score of 2.27 for item 9 (a Type II error is impossible if the results are statistically significant).

Insert Table 5 about here

Perceptions regarding the *influence of sample sizes on statistical tests* are reported in Table 6. Respondents "disagreed" (M=2.37, SD=1.31) that "statistically significant results are more noteworthy when sample sizes are small."

Insert Table 6 about here

Table 7 shows respondents' perceptions of whether *statistical probabilities are exclusively measures of effect size*. A mean rating of 3.82 was reported for item 4 (failure to obtain statistical significance means that results were not noteworthy or important).

Insert Table 7 about here

Perceptions of *p values* are summarized in Table 8. Respondents "agreed" that "studies with non-significant results can still be very important" (M=1.45, SD=1.19).

Insert Table 8 about here

Finally, participants were asked about whether *p values evaluate population parameters and result replicability*. As revealed in Table 9, respondents' perceptions were "slightly agreeable" to "neutral" ($\underline{M}$=2.22 to 3.05).

Insert Table 9 about here

Discussion and Conclusions

It appears that AVERA members who were most comfortable with and interested in statistical issues (quantitative methods) may have been likely to respond to the survey. It should be noted that some AVERA members use only qualitative methods. These individuals may have been less likely to respond to the survey.

AVERA members' general views regarding statistical testing is consistent with previous research (Carver, 1993; Thompson, 1996; Mittag & Thompson, 2000).
These findings suggest that the controversy has raised some consciousness among AVERA researchers' perceptions on the general views of statistical testing.

Respondents were more likely to "slightly disagreed" with the two views pertaining to the General Linear Model (GLM). These findings contradict a previous study reported by Mittag and Thompson (2000). In their study, respondents were basically "neutral" on: (a) the point of whether all statistical analyses (e.g., $\underline{t}$-tests, ANOVA, $\underline{r}$, $\underline{R}$) are correlational, and (b) respondents "agreed" that regression could be used to test hypothesis about means. Statisticians have argued that parametric methods are part of a single family, and that all are correlational (Cohen, 1968; Knapp, 1978; Thompson, 1991; Mittag & Thompson, 2000). One important implication of the

GLM is that r^2 analogs can be reported as effect sizes in all analyses (Mittag & Thompson, 2000).

The two views pertaining to stepwise methods were more likely to be perceived as acceptable for identifying the best variable set and importance. These findings suggest that some AVERA researchers are not aware that stepwise methods do not identify the best predictor set of a given size (Cliff, 1987; Huberty, 1989; Thompson, 1995). In a recent study by Thomas (2000), over 70% of AVERA members indicated a need for adequate workshops on emerging statistical techniques and research methods.

Stepwise methods are especially problematic when statistical significance tests are invoked to determine stopping positions, because the methods have all the problems, in spades, associated with conventional statistical significance applications (Carver, 1987; Cohen, 1994; Thompson, 1993, 1994a, 1994b, 1994c). As a general proposition, there are readily available software programs to assist with appropriate variable selection efforts. Thus, stepwise analyses should be eschewed in favor of programs such as those offered by McCabe (1995), the Morris program distributed within Huberty's (1994) book, or SAS procedure RSQR. Regarding interpretations involving the origins of explained variance (i.e., variable ordering), a useful alternative is simple to consult standardized weights (beta weights) and structure coefficients (Thompson & Borello, 1985).

Overall, views regarding score reliability appeared to be "neutral". These findings are consistent with a similar study reported by Mittag and Thompson (2000) for the American Educational Research Association. It is important to remember that a test is not reliable or unreliable. Reliability is a property of the scores on a test for a particular population of examines... Thus, authors should provide reliability coefficients of the data being analyzed.

Interpreting the size of the observed effects requires an assessment of the reliability of the scores (Wilkinson & The APA Task Force on Statistical Inference, 1999, p.596).

Views pertaining to Type I and Type II errors appeared to be "neutral". Examination of these findings revealed a "mixed perception" of the definition of a Type I error. By definition, a Type one error can only occur if results are statistically significant (Oliver, 1981).

Respondents were more likely to have a "neutral perception" regarding: (a) whether "significance tests are partly a test of whether the researcher had a large sample," and (b) "every null hypothesis will eventually be reflected at some sample size." Mittag and Thompson (2000) reported similar findings. Several factors can influence the size of the sample used in a research study, but with the exception of cost, information about such factors is often incomplete and it becomes difficult to set an exact size (Wiersma, 2000). Hinkle and Oliver (1983) discuss estimating necessary sample size based on certain characteristics.

Studies with non-significant results can still be very important. Tyler (1931) pointed out that "differences which are statistically significant are not always socially important. The corollary is also true: differences which are not shown to be statistically significant may nevertheless be socially significant" (pp. 116-117). Meehl (1997) characterized the use of the term "significant" as being "cancerous" and "misleading" (p. 421) and advocated that researchers interpret their results in the terms of confidence intervals rather than p values. Moore (1992) noted that:

> We as vocational educators should be proud of our improving process as "research technicians". I am not advocating we do away with statistical testing. However, I am cautioning that we must not get caught up in the misguided belief that having statistically significant things makes our research significant. (p. 5)

Recommendations

1. Future AVERA researchers should be encouraged to (a) correctly interpret statistical tests, (b) always interpret effect sizes, and (c) the replicability of the results should be empirically investigated, either through actual replication of the study, or by using methods such as cross-validation, the jackknife, or the bootstrap (see Thompson, 1994).
2. Future researchers in the field may consider additional preparation in statistics so as to comprehend some of the advanced techniques which are used in the current research literature in career and technical education/workforce education.
3. Joint efforts between career education/and other fields of education should be considered in offering statistics courses at all levels due to the similarity in the use of statistics techniques across the fields.
4. From a practical standpoint of view, graduate programs of career and technical/workforce education should ensure inclusion of statistical techniques at the basic, intermediate and advanced levels so that graduate students can understand the statistical aspect of most research literature in the field.
5. "Progress has no greater enemy than habit" (McCracken, 1991, p.303). As a profession we must break out of the habit of simply describing relationships and differences between and among groups. The explanation of the phenomena must be our goal.
6. For further study, it is recommended that research be conducted to determine the nature of interpretation in qualitative research in career and technical/workforce education. The interplay of subject and object, self and problem, is usually taken for granted or ignored in both qualitative and quantitative research.

References

American Psychological Association (1992). Publication manual of the American Psychological Association (4th ed.). Washington, DC: Author.

Ary, D., Jacobs, L., & Razavieh, A. (1996). Introduction to research in education (5th ed.). Ft. worth: Holt, Rinehart, and Winston, Inc.

Biskin, B. H. (1998). Comment on significance testing. Measurement and Evaluation in Counseling and Development, 31, 58-62.

Bracey, G.W. (1988). Tips for readers of research. Phi Delta Kappan, 70(3), 257-258.

Carver, R. (1993). The case against statistical significance testing revisited. Journal of Experimental Education, 61(4), 287-292.

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

Cheek, J.G. (1998). Maintaining momentum in vocational education research. Journal of Vocational Education Research, 13(1), 1-17.

Cliff, N. (1987). Analyzing multivanate data. San Diego: Harcovert Brace Jovanovich.

Cohen, J. (1968). Multiple regression as a general data-analytic system. Psychological Bulletin, 70, 426-443.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences. New York: Academic Press.

Cohen, J. (1994). The earth is round ($p<.05$). American Psychologist, 49, 997-1003.

Daniel, L.G. (1998). Statistical significance testing: A historical overview of misuse and misinterpretations with implications for the editorial polices of educational journals. Research in the Schools, 5(2), 23-32.

Dillman, D.A. (2000). Mail and internet surveys: The tailored design method (2nd. ed.). New York, NY: John Wiley & Sons, Inc.

Gay, L. R. (1996). Educational research: Competencies for analysis and application (5th.ed.). Upper Saddle River, NJ: Prentice-Hall, Inc.

Gay, L.R., & Airasian, P. (2000). Educational research: Competencies for analysis and application (6th ed.) Upper Saddle River, NJ: Prentice-Hall, Inc.

Henkle, D.E., & Oliver, J.D. (1983). How large should the sample be? A question with no simple answer? Or... Educational and Psychological Measurement, 43, 1050-1051.

Hillison, J. (1989). Using all tools available to vocational education researchers. Journal of Vocational Education Research, 15(1), 1-8.

Huberty, C.J. (1994). Applied discriminant analysis. New York: Wiley.

Huberty, C.J. (1998). On statistical testing. Educational Researcher, 6(8), 4-9.

Huberty, C.J. (1989). Problems with stepwise methods-better alternatives In B. Thompson (Ed.). Advances in social science methodology (Vol.1, pp.43-70). Greenwich, CT: JAI Press.

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

Kerlinger, F.N. (1986). Foundations of Behavioral Research (3rd. ed.). New York: Holt, Rinehart and Winston.

Knapp, T.R. (1978). Canonical correlation analysis: A general parametric significance testing system. Psychological Bulletin, 85, 410-416.

Kotrlik, J.W. (2000). Guidelines for authors. Journal of Agricultural Educating, 41(1), inside cover.

Krejcie, R.V., & Morgan, D.W. (1970). Determining sample size of research activities. Educational and Psychological Measurement, 30, 607-608.

Kupfersmid, J.(1998). Improving what is published. American Psychologist, 43(8), 635-642.

Lynch, K.B. (1983). Qualitative and quantitative evaluation: Two terms in search of meaning. Educational Evaluation and Policy Analysis, 5(4), 461-464.

McCabe, G.P. (1975). Computations for variable selection in discriminant analysis. Techno metrics, 17, 103-109.

McCracken, J.D. (1991, December). The use and misuse of correctional and regression analysis in agricultural education research. Paper presented as the invited address at the National Agricultural Education Research Meeting, Los Angeles, CA.

Meehl, P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronal and slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.

Meehl, P. (1997). The problem is epistemdogy, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow,

S.A. Mulaik, & J.H. Steiger (Eds.), What if there were no significance tests? (pp. 393-426). Mahwah, NJ: Erlbaum.

Miller, L.E., & Smith, K.L. (1983). Handling non-response issues. Journal of Extension, 21, 45-50.

Mittag, K.C. (1999). The psychometrics group instrument: Attitudes about contemporary statistical controversies. Unpublished instrument. The University of Texas at San Antonio.

Mittag, K.C., & Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. Educational Researcher, 29(4), 14-20.

Moore, G. E. (1992). The significance of researcher vocational education: The 1992 AVERA presidential address. Journal of Vocational Education Research, 17(4), 1-4.

Nelson, N., Rosenthal, R., & Rosnow, R.L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. American Psychologist, 41, 1299-1301.

Oakes, M. (1986). Statistical inference: A commentary for the social and behavioral sciences. New York: Wiley.

Oliver, J.D. (1981). Improving agricultural education research. The Journal of American Association of Teacher Educations in Agriculture, 22(1), 9-15.

Rosenthal, R., & Gaito, J. (1963). The interpretation of level of significance by psychological researchers. Journal of Psychology, 55, 33-38.

Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researcher. Psychological Methods, 1(2), 115-129.

Shaver, J. (1992, April). What significance testing is, and what it isn't. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.

Shulman. L.S. (1970). Reconstruction of educational research. Review of educational Research, 40, 371-393.

Smith, B.B. (1984). Empirical-analytic research paradigm research in vocational education. Journal of Vocational Education Research, 9(04), 20-35.

Soltis, J.F. (1984). On the nature of educational research. Educational Researcher, 13(10), 5-10.

Thomas, H. (2000). Keeping on track to the future: The 1999 AVERA presidential address. Journal of Vocational Education Research, 25(1), 4-20.

Thompson, B., & Borello, G.M. (1985). The importance of structure coefficients in regression research. Educational and Psychological Measurement, 45, 203-209.

Thompson, B. (1991). A primer on the logic and use of canonical correlation analysis. Measurement and Evaluation and Development, 24(2), 80-95.

Thompson, B. (1994a). Guidelines for Authors. Educational and Psychological Measurement, 54, 837-847.

Thompson, B. (1994b). The concept of statistical significance testing (An ERIC/AE Clearinghouse Digest EDO-TM-94-1). Measurement Update, 4(1), 5-6, (ERIC Document Reproduction Service No. ED366 654)

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

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here. A guidelines editorial. Educational and Psychological Measurement, 55, 525-534.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25(2), 26-30.

Tyler, R.W. (1931). What is statistical significance? Educational Research Bulletin, 10, 115-118, 142.

Tyron, W.W. (1998). The inscrutable null hypothesis. American Psychologist, 53, 796.

Vacha-Haase, T., Nilsson, J.E., Reetz, O.R., Lance, T.S. & Thompson, B. (2000). Reporting practices and ATA editorial policies regarding statistical significance and effect size. Theory & Psychology, 10, 413-425.

Vacha-Haase, T., & Thompson, B. (1998, August). APA editorial polices regarding statistical significance and effect size: Glacial fields more inexorably (but glacially). Paper presented at the annual meeting of the American Psychological Association, San Francisco, CA.

Walker, H.M. (1956). Methods of research. Review of Educational Research, 26(3), 323-344.

Warmbrod, J.R. (1986). Priorities for continuing progress in research in agricultural education. Paper presented at he 35th Annual Southern Region Research Conference in Agricultural Education, Little Rock, AR.

Wiersma, W. (2000). Research methods in education: An introduction (7th. ed.). Needham Heights: Allyn and Bacon.

Wilkerson, L., & The APA Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54, 594-604.

West, C.K., & Robinson, D.G. (1980). Prestigious psycho-educational research published form 1910 to 1974: Types of explanations, focus, authorship, and other concerns. Journal of Educational Research, 73(5), 271-275.

Zhang, C. (1993, April). The determination of statistical sophistication of research in vocational education. Paper presented at the meeting of the American Educational Research Association Atlanta, GA.

Zuckerman, M., Hodgins, H.S., Zuckerman, A., & Rosenthal, R. (1993). Contemporary issues in the analysis of data: A survey of 531 psychologists. Psychological Science, 4, 49-53.

Table 1

Means and Standard Deviations of AVERA Members' General Views Regarding Statistical Testing (n = 40).

No.	Perception Statement/Item	M[a]	SD
1.	Controversies regarding the use of significance tests have existed for many years in the past, and will doubtless continue for many years in the future.	4.47	.60
2.	It would be better if everyone used the phrase "statistically significant" rather than "significant," to describe the results when the null hypothesis is rejected.	4.25	.87
3.	Most studies are conducted with insufficient statistical power against Type II error.	3.41	.85
5.	All that significance means is that the researcher rejected the null hypothesis.	3.02	1.44
4.	Science would progress more rapidly if tests of significance were banned from journal articles.	1.70	.88

[a]Note. Response scale: 1 = disagree, 5 = agree.

Table 2

Means and Standard Deviations of AVERA Members' Perceptions of the General Linear Model (n = 40).

No.	Perception Statement/Item	M	SD
26.	It is not possible to use regression to statistically test the null that means of different groups are equal.	3.70	.88
12.	All statistical analyses (e.g., *t*-tests, ANOVA, r, R) are correlational.	2.37	1.17

Note. For item 26, after recoding, 1 = agree, 5 = disagree.

For item 12, 1 = disagree, 5 = agree.

Table 3

Means and Standard Deviations of AVERA Members' Perceptions of Stepwise Methods (n = 40).

No.	Perception Statement/Item	M[a]	SD
13.	In regression and other analyses, stepwise methods can reasonably be used to identify the best subset of predictors of a given subset size.	3.55	.95
20.	When researchers do stepwise analyses, the order of the entry of the variables (1st, 2nd, etc.) provides one useful indication of the importance of the variables.	3.47	.98

[a]Note. Response scale: 1 = disagree, 5 = agree.

Table 4

Means and Standard Deviations of AVERA Members' Perceptions of Score Reliability (n = 40).

No.	Perception Statement/Item	M	SD
23.	Poor reliability of data in a given study will tend to lower or attenuate the effect sizes that are detected.	3.62	1.12
28.	Reliability does not directly affect the likelihood of obtaining significance in a given study.	3.45	1.21
7.	On its face the statement, "the reliability of the test," asserts an untruth, since reliability is not a characteristic of a given test.	2.85	1.18
19.	Testing the significance of a reliability of validity coefficient with null hypothesis that $r^2 = 0$ is not useful or productive.	2.80	.99

Note. For items 7, 19, and 23, 1 = disagree, 5 = agree.

For item 28, after recoding, 1 = agree, 5 = disagree.

Table 5

Means and Standard Deviations of AVERA Members' Perceptions of Type I and II Errors (n = 40).

No.	Perception Statement/Item	M	SD
22.	It is possible to make both Type I and Type II error in a given study.	3.37	1.21
17.	Type I errors may be a concern when the null hypothesis is not rejected.	2.72	1.19
29.	Type II errors are probably fairly common within published research.	2.52	1.01
9.	A Type II error is impossible if the results are statistically significant.	2.27	1.10

Note. For items 17, 22, 29, after recoding, 1 = agree, 5 = disagree.
For item 9, 1 = disagree, 5 = agree.

Table 6

Means and Standard Deviations of AVERA Members' Perceptions of Sample Size Influences (n = 40).

No.	Perception Statement/Item	M[a]	SD
16.	Every null hypothesis will eventually be rejected at some sample size.	3.15	1.29
25.	Significance tests are partly a test of whether the researcher had a large sample.	2.87	1.18
10.	Statistically significant results are more noteworthy when sample sizes are small.	2.37	1.31

[a]Note. Response scale: 1 = disagree, 5 = agree.

Table 7

Means and Standard Deviations of AVERA Members' Perceptions of Effect Sizes (n = 40).

No.	Perception Statement/Item	M	SD
14.	If a dozen different researchers investigated the same phenomenon using the same null hypothesis, and none of the studies yielded statistically significant results, this means that the effects being investigated were not noteworthy or important.	3.82	1.19
11.	Smaller p values provide direct evidence that study effects were larger.	3.27	1.17
24.	The p values reported in different studies cannot be readily compared, because these values are confounded with different sample sizes across studies.	3.15	1.23

Note. For items 11 and 14, after recoding, 1 = agree, 5 = disagree.
For item 24, 1 = disagree, 5 = agree.

Table 8

Means and Standard Deviations of AVERA Members' Perceptions of p Values (n = 40).

No.	Perception Statement/Item	M	SD
27.	Unlikely results are generally more important or noteworthy.	3.50	1.06
6.	Finding that $p < .05$ is one indication that the results are important.	2.80	1.41
18.	Studies with non-significant results can still be very important.	1.45	1.19

Note. After recoding 1 = agree, 5 = disagree.

Table 9

Means and Standard Deviations of AVERA Members' Perceptions of p as Replicability Evidence (n = 40).

No.	Perception Statement/Item	M	SD
8.	Smaller and smaller values for the calculated p indicate that the results are more likely to be replicated in future research.	3.05	1.21
15.	The p values that are calculated in a given study test the probability of the results occurring in the sample, and not the probability of results occurring in the population.	2.82	1.33
21.	Significance tests evaluate the probability that the results for the sample are the same in the population.	2.22	1.09

Note. For items 8 and 21, after recoding 1 = agree, 5 = disagree.
For item 15, 1 = disagree, 5 = agree.

CE061918 A

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title:	American Vocational Education Research Association Members' Perceptions of Statistical Significance Tests and Other Statistical Controversies
Author(s):	Howard R. D. Gordon
Corporate Source:	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following two options and sign at the bottom of the page.

	The sample sticker shown below will be affixed to all **Level 1** documents	The sample sticker shown below will be affixed to all **Level 2** documents	
☑ Check here **For Level 1 Release:** Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical) *and* paper copy.	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	☐ Check here **For Level 2 Release:** Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical), but *not* in paper copy.
	Level 1	**Level 2**	

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at **Level 1**.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Sign here→ please	Signature: H. Gordon	Printed Name/Position/Title: Howard R.D. Gordon Prof. of Occupational Leadership	
	Organization/Address: 434 Harris Hall. Dept. ATE, Marshall University Huntington, WV 25755	Telephone: 304 696 3079	FAX: 304 696 3077
		E-Mail Address: gordonh@marshall.edu	Date: 6/27/2001

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

Associate Director for Database Development
ERIC Clearinghouse on Adult, Career, and Vocational Education
Center on Education and Training for Employment
1900 Kenny Road
Columbus, OH 43210-1090

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to: