IOSR Journal of Research & Method in Education (IOSR-JRME)
e-ISSN: 2320–7388,p-ISSN: 2320–737X Volume 1, Issue 5 (May. –Jun. 2013), PP 78-83
www.iosrjournals.org

The Roles of Statistical Significance Testing In Research

Oni, N. O., Osuntoki N. B., Fasasi, S. K. & Rahaman Aliyu
Department of Mathematics and Statistics, The Polytechnic Ibadan, Saki Campus, Oyo State, Nigeria.

Abstract: *The research methodology literature in recent years has included a full frontal assault on statistical significance testing. The purpose of this paper is to promote the position that, while significance testing as the sole basis for result interpretation is a fundamentally flawed practice, significance tests can be useful as one of several elements in a comprehensive interpretation of data. Specifically, statistical significance is but one of the three criteria that must be demonstrated to establish a position empirically. Statistical significance merely provides evidence that an event did not happen by chance. However, it provides no information about the meaningfulness (practical significance) of an event or if the result is replicable. Thus, we support other researchers who recommend that statistical significance testing must be accompanied by judgments of the events practical significance and replicability.*
Keywords:*Statistical SignificanceTesting, Educational Research, Assult, Replicability, Jacknife.*

I. Introduction

The research methodology literature in recent years has included a full frontal assault on statistical significance testing (Thompson,1993b). There are some who recommend the total abandonment of statistical significance testing as a research methodology option, while others choose to ignore the controversy and use significance testing following traditional practice. The purpose of this paper is to promote the position that while significance testing by itself may be flawed, it has not outlived its usefulness. However, it must be considered in the total context of the statistical significance as one of several criteria that must be demonstrated to establish a position empirically. Statistical significance merely provides evidence that an event did not happen by chance, however, it provides no information about the meaningfulness (practical significance) of an event or if the result is replicable.

This paper addresses the controversy by first providing a critical review of the literature. Following the review are our summary and recommendations. While the recommendations by themselves are not entirely new, they provide a broad perspective on the controversy and provide practical guidance for researchers employing statistical significance testing in their work.

II. Review of the Literature

Scholars have been using statistical testing for research purposes since the early 1700s (Huberty, 1993). In the past 300 years, applications of statistical testing have advanced considerably, most noticeably with the advent of the computer and recent technological advances. However, much of today's statistical testing is based on the same logic used in the first statistical tests and advanced in the early twentieth century through the work of Fisher, Nyman, and the Pearson family (See the appendix to Mulaik, Raju and Harshman (1997) for further information. Specifically, significance testing and hypothesis testing have remaind at cornerstone of research papers and the teaching of introductory statistics courses. (It should be noted that while the authors recognize the importance of Bayesian testing for statistical significance, it will not be discussed, as it falls outside the context of this paper). Both methods of testing hold at their core basic premises concerning probability. In what may be termed Fisher's *p value approach*, after stating a null hypothesis and then obtaining sample results (i.e. "statistics"), the probability of the sample results (or sample results more extreme in their deviation from the null) is computed, assuming that the null is true in the population from which the sample was derived (see Cohen, 1994 or Thompson, 1996 for further explanation). The Neyman Pearson or ***fixed-alpha approach*** specifies a level at which the test statistic should be rejected and it sets apriori to conducting the test of data. A null hypothesis (H_0) and an alternative hypothesis ($H_{a)}$ are stated, and if the value of the test statistic falls in the rejection region the null hypothesis is rejected in favour of the alternate hypothesis. Otherwise the full hypothesis is retained on the basis that there is insufficient evidence to reject it.

Distinguishing between the two methods of statistical testing is important in terms of how methods of statistical analysis have developed in the recent past. Fisher's legacy of statistical analysis approaches (including ANOVA methods) relies on subjective judgements concerning differences between and within groups, using probability levels to determine which results are statistically significant from each other. Karl Pearson's legacy involves the development of correlational analyses and providing indexes of association. It is because of different approaches to analyses and different philosophical beliefs that the issue of testing for statistical significance has risen. In Huberty's (1993) historical review of the importance of statistical significance testing

literature, the research community has shifted from one perspective to another, often within the same article. Currently we are in an era where the value of statistical significance testing is being challenged by many researchers. Both positions (arguing for and against the use of statistical significance test in research) are presented in this literature review, followed by a justification for our position on the use of statistical significance testing as part of a comprehensive approach.

As previously noted, the research methodology literature in recent years has included a full frontal assault on statistical significance testing. The assault is based on whether or not statistical significance testing has value in answering a research question posed by the investigators.

In fact, null hypothesis testing still dominates the social sciences (Loftus & Masson, 1994) and still draws derogatory statements concerning the researcher's methodological competence. As Falk and Greenbaum (1995) and Weitzman (1984) noted, the researchers' use of the null may be attributed to the experimenters' ignorance, misunderstanding, laziness, or adherence to tradition. Carver (1993) agreed with the tenets of the previous statement and concluded that "the best research articles are those that include no tests of statistical significance'' (p. 28). One may even concur with Cronbach's (1975) statement concerning periodic efforts to "exorcize the null hypothesis" (p. 124) because of its harmful nature. It has also been suggested by Thompson (1998) in his paper on the aetiology of researcher resistance to changing practices that researchers are slow to adopt approaches in which they were not trained originally.

In response to the often voracious attacks on significance testing, the American Psychological Association, as one of the leading research forces in the social sciences, has reacted with a cautionary tone: "*An APA task force won't recommend a ban on significance testing, but is urging psychologists to take a closer look at their data" (*Azar, 1997) In reviewing the many publications that offer advice on the use or misuse of statistical significance testing or plea for abstinence from statistical significance testing, we found the following main arguments for and against its use: (a) what statistical significance testing does and does not tell us, (b) empha-sizing effect-size interpretations, (c) result replicability, (d), importance of the statistic as it relates to sample size, (e)the use of language in describing results, and (f) the recognition of the importance of other types of information such as Type II errors, power analysis, and confidence intervals

2.1 What Statistical Significance Testing Does and Does Not Tell Us

Carver (1978) provides a critique against statistical significance tests of statistical significance, there appeared to be little change in research practices. Fifteen years later, Carver (1993) focused on the negative aspects of significance testing. His article indicted the research community for reporting significant differences when the results may trivial, and called for the use of effect size estimates and study replicability, Carvers argument focused on what statistical significance testing does not do, and proceeded to highlight ways to provide indices of practical significance and result replicability. Carver (1993) recognized that 15years of trying to extinguish the use of statistical significance testing has resulted in little change in the use and frequency of statistical significance testing. Therefore, the tone of the 1993 article differed from the 1978 article in shifting from a dogmatic anti-statistically significant approach to more of a bipartisan approach where the limits of significance testing were noted and ways to decrease their influence provided. Specifically, Carver (1993) offered four ways to minimize the importance of statistical significance testing:

a. Insist on the word *statistically* being placed in front of significance testing.
b. Insist that the results always be interpreted with respect to the data first and statistical significance second.
c. Insist on considering effect sizes (whether significant or not), and
d. Require journal editors to publicize testing prior to their selection as editors.

Shaver (1993), provides a description of what significance testing is and a list of the assumptions involved in statistical significance testing. He also methodically stressed the importance of the assumptions of random selection of subjects and their random assignment to groups. Levin (1993) agreed with the importance of meeting basic statistical assumptions, but pointed out a fundamental distinction between statistical significance testing and statistics that provide estimates of practical significance. Levin observed that a statistically significant difference gives information about *whether* a difference exists. As Levin noted, if the null hypothesis is rejected, the *p* level provides a "posteriori indication of the probability of obtaining the outcomes as extreme as or more extreme than the one obtained, given the null hypothesis is true". The effect size gives an estimate of the noteworthiness of the results. Levin made the distinction that the effect size may be necessary to obtain the size of the effect; however, it is statistical significance that provides information which alludes to whether results may have occurred by chance. In essence, Levin's argument was for the two types of significance being complementary and not competing concepts. Frick (in press) agreed with Levin: "When the goal is to make a claim about how scores were produced, statistical testing is still needed, to address the possibility of an observed pattern in the data being caused just by chance fluctuation" (in press).

One of the most important emphases in criticisms of contemporary practice is that researchers must evaluate the practical importance of result, and not only statistical significance. Thus, Kirk (1996) agreed that statistical significance testing was a necessary part of a statistical analysis. However, he asserted that the time had come to include practical significance as necessary, but insufficient for interpreting research. Suen (1992), used an 'overbearing guest' analogy to describe the current state of statistical significance testing. In Suen's analogy, statistical significance is the overbearing guest at a dinner party who inappropriately dominates the activities and conversation to the point that we forget who the host was. We cannot disinvite this guest, instead, we need to put this guest in the proper place; namely, as one of the many guests and by no means the host. (p. 78).

Suen's reference to a "proper place" is a call for researchers to observe statistical significance testing as a means to "filter out the sampling fluctuations hypothesis so that the observed information (difference, correlation) becomes slightly more clear and defined" (p. 79). The other "guests" that researchers should elevate to a higher level include ensuring the quality of the research design, measurement reliability, treatment fidelity, and using sound clinical judgment of effect size.

For Frick (in press), Kirk (1996), Levin (1993), and Suen (1992), the rationale for statistical significance testing is independent of and complementary to tests of practical significance. Each of the tests provides distinct pieces of information, and all three authors recommend the use of statistical significance testing; however, it must be considered in combination with other criteria. Specifically, statistical significance is but one of the three criteria that must be demonstrated to establish a position empirically (the other two being practical significance and replicability).

In the review of the literature, the authors were unable to find an article that argued against the value of including some form of effect size or practical significance estimate in a research report. Huberty (1993) notes that "of course, empirical researchers should not rely exclusively on statistical significance to assess results of statistical tests. Some type of measurement of magnitude or importance of the effects should also be made" (p. 329). Carver's third recommendation (mentioned previously)was the inclusion of terms that denote an effect size measure. Shaver (1993) believes that "studies should be published without tests of statistical significance, but not without effect sizes" (p. 311); and Snyder and Lawson (1990) contributed a paper to the journal of Experimental Education special edition on statistical significance testing titled "Evaluating Results Using Corrected and Uncorrected Effect Size Estimates." Thompson(1987, 1989, 1993a, 1996, 1997) argues for effect sizes as one of his three recommendations (the language use of statistical significance and the inclusion of result explicability results were the other two); Levin (1993) reminds us that "statistical significance (alpha and p values) and practical significance (effect sizes) are not competing concepts-they are complementary ones" Cortina and Dunlap (1997), Frick (1995,in press), and Robinson and Levin (1997) agree that a measure of the size of an effect is indeed important in providing results to a reader.

We agree that it is important to provide an index of not only the statistical significance, but also a measure of its magnitude. Robinson and Levin (1997) took the issue one step further and advocated for the use of adjectives such as strong /large, moderate/medium, etc, to refer to the effect size and to supply information concerning p values. However, some authors lead us to believe that they feel it may be necessary only to provide an index of practical significance and that it is unnecessary to provide statistical significance information. For example, it could be con-cluded from the writings of Carver (1978, 1993) and Shaver (1993) that they would like to abandon the use of statistical significance testing results. He did assert that you can attach a p-value to an effect size, but "it is far more in-formative to provide a confidence interval" (Cohen, 1990, p 1310). Levin, in his 1993 article and in an article co-authored with Robinson (1997), argued against the idea of a single indicator of significance.

III. Result Replicability

Carver (1978) was quick to identify that neither significance testing nor effect sizes typically inform the researcher regarding the likelihood that results will be replicated in future research. Schafer (1993) felt that much of the criticism of significance testing was misfocused. Schafer concluded that readers of research should not mistakenly assume that statistical significance is an indication that the results may be replicated in future.

According to Thompson (1996), "If science is the business of discovering replicable effects, because statistical significance tests do not evaluate result replicability, then researcher should use and report some strategies that do evaluate the replicability of their results" Robinson and Levin (1997) were in total agreement with Thompson's recommendations of external result replicability. However, Robinson and Levin (1997) disagreed with Thompson when they concluded that internal replication analysis constitutes "an acceptable substitute for the genuine 'article'' (p. 26). Thompson (1997), in his rejoinder, recognized that external replication studies would be ideal in all situations, but concludes that many researchers do not have the stamina for external replication, and internal replicability analysis helps to determine where noteworthy results originate.

In terms of statistical significance testing, all of the arguments offered in the literature concerning replicability report that misconceptions about what statistical significance tells us are harmful to research. The authors of this paper agree, but once again note that misconceptions are a function of the researcher and not the test statistic. Replicability information concerning noteworthy results.

Importance of the Statistic as it Relates to Sample Size

According to Shaver (1993), a test of statistical significance "addresses only the simple question of whether a result is a likely occurrence under the null hypothesis with randomization and a sample of size n" indicates the importance of sample size in the H_O decision-making process. As reported by Meehl (1967) and many authors since, with a large enough sample and reliable assess-ment, practically every association will be statistically significant. Two salient points applicable to this discussion were highlighted by Thompson first is the relationship of n to statistical significance, providing a simulation that shows how, by varying n to create a large enough sample, a difference between two values can change a non-significant result into a statistically significant result. The second property of significance testing Thompson alluded to was an indication that "superficial understanding of significance testing has led to serious distortions, such as researchers interpreting significant results involving large effect sizes" (p.2). following this line of researchers, Thompson (1993a) humorously noted that "tired researchers, having collected data from hundreds of subjects, then conducts a statistical test to evaluate whether there were a lot of subjects, which the researchers already know, because they collected the data and they are tired" (p. 363). Thus, as the sample size increases, the importance of significance testing is reduced. However, in small sample studies, significance testing can be useful, as it provides a level of protection from reporting random results by providing information about the chance of obtaining the sample statistics, given the sample size n, when the null hypothesis is exactly true in the population.

IV. The Use of Language in Describing Results

Carver (1978, 1993), Cronbach (1975), Morrison and Henkel (1970), Robinson and Levin (1975), and Thompson (1987, 1989, 1993a, 1996, 1997) all stressed the need for the use of better language to describe significant results. As Schneider and Darcy (1984) and Thompson (1989) noted, significance is a function of at least seven interrelated features of a study where the size of the sample is the most influential characteristic. Thompson (1989) used an example of varying sample sizes with a fixed affect size to indicate how a small change in sample size affects the decision to reject, or fail to reject H_O. The example helped to emphasize the cautionary nature that should be practiced in making judgements about the null hypothesis and raised the important issue of clarity in writing. These issues were stressed by Thompson (19996) when he called for the use of the term "statistically significant" when referring to the process of rejecting H_O based on an alpha level. It was argued that through the use of specific terminology, the phrase "statistically significant" would not be confused with the common semantic meaning of significant.

While applauding Thompson for his "insightful analysis of the problem and the general spirit of his policy recommendations" (p.21), Robinson and Levin were quick to counter with quips about "language police" and letting editors focus on content and substance and not on dotting the i's and crossing the t's. However, and interestingly, Robinson and Levin (1997) proceeded to concur with Thompson on the importance of language and call for researchers to use words that are more specific in nature. It is Robinson and Levin's (1997) recommendation that, instead of using the word statistically significant, researchers use statistically nonchance or statistically real, reflecting the test's intended meaning. The authors' rationale for changing the terminology reflects their wish to provide clear and precise information.

In respect to the concerns raised concerning the use of language, it is not the practice of significance testing that has created the statistical significance debate. Rather, the underlying problem lies with careless use of language and the incorrect assumptions made by less knowledge able readers and practitioners of research. As before, Cohen (1994), referred to the misinterpretations that result from statistical testing (e.g, the belief that p-values are the probability that the null hypothesis is false). Cohen again suggested exploratory data analysis, graphical methods, and placing an emphasis on estimating effect sizes using confidence intervals. Once more, the basis for the argument against statistical significance testing falls on basic misconceptions of what the p-value statistic represents.

One of the strongest rationales for not using statistical significance values relies on misconceptions about the meaning of the p-value and the language used to describe its purpose. As Cortina and Dunlap (1997) noted, there are many cases where drawing conclusions based on p value are perfectly reasonable. In fact, as Cortina and Dunlap (1997), Frick (1995), Levin (1993), and Robinson and Levin (1997) pointed out, many of the criticisms of the p value are built on faulty premises, misleading examples and incorrect assumptions concerning population parameters, null hypotheses, and their relationship to samples.

V. The Recognition of the Importance of Other Types of Information

Other types of information are important when one considers statistical significance testing. The researcher should not ignore other information such as type II errors, power analysis, and confidence intervals. While all of these statistical concepts are related, they provide different types of information that assist researchers in making decisions. There is an intricate relationship between power, sample size, effect size, and alpha (Cohen, 1988). Cohen recommended a power level of 80 for no other reason than that for which Fisher set an alpha level of 0.5 – it seemed a reasonable number to use. Cohen believed that the effect size should be set using theory, and the alpha level should be set using what degree of Type I error the researcher is willing to accept based on the type of experiment being conducted. In this scenario, n is the only value that may vary, and through the use of mathematical tables, is set at a particular value to be able to reach acceptable power, effect size, and alpha levels. Of course, in issues related to real-world examples, money is an issue and therefore sample sizes may be limited.

It is possible that researcher have to use small n because of the time, money and accuracy aimed at. Cohen (1990) addresses the problems motioned above by asking researchers to plan their research using the level of alpha risk they want to take, the size of the effect they wish to find, a calculation sample size, and the power they want . if one is unable to use a simple size of sufficient magnitude, one must compromise power, effect size, or increasing your alpha level '(p.1310).

The sentiment was shared by Schafer (1993) who also believed that researchers should set alpha levels, conduct power analysis, decide on the size of the sample, and design research studies that would increase effect size (e.g. through the careful addition of covariates in regression analysis or extending treatment interventions). It is necessary to balance sample size against power, and this automatically means that we do not fix one of them. It is also necessary to balance size and power against cost, which means that we do not arbitrarily fix sample size. All of the recommendation may be conducted prior to the data collection and therefore before the data analysis. The recommendation, in effect, provides evidence that methodological prowess may overcome some of the posterior problem researchers find.

VI. Summary and Recommendations

We support other researchers who state that statistical significance testing must be accompanied by judgment of the event's practical significance and replicability. However, the likelihood of a chance occurrence of an event must not be ignored. We acknowledge the fact that the importance of significance testing is reduced as sample size increases. In large sample experiments, particularly those involving multiple variables, the role of significance testing diminishes because event small, non-meaningful differences are often statistically significant. In small-sample studies where assumption such as random samp-ling are practical, significance testing provides meaningful protection from random results. It is important to remember that statistical significance is only one criterion useful to inferential researchers. In addition to statistical significance, practical significance, and replicability, researchers must also consider Type II Errors and sample size. Furthermore, researchers should not ignore other techniques such as confidence interval. While all of these statistical concepts are related, they provide different types of information that assist researchers in making decision.

Our recommendations reflect a moderation mainstream approach. That is, we recommend that in situations where the assumptions are tenable, statistical significance testing can still be applied. However, we recommend that the analyses always be accompanied by at least one measure of practical significance, such as effect size. The use of confidence intervals can be quite helpful in the interpretation of statistically significance or statistically nonsignificant result. Further do not consider a hypothesis or theory 'proven' even when both the statistical and practical significance have been established; the results have to be replicability. Even if it is not possible to establish external replicability for a specific study, internal approaches such as jackknife or bootstrap procedures are often feasible. Finally, please note that as sample sizes increase, the role of statistical significance becomes less important and the role of practical significance increases. This is because statistical significance can provide false comfort with results when sample sizes are large. This is especially true when the problems is multivariate and the large sample is representative of the target population. In these situations. Effect size should weight heavily in the interpretations.

References

[1]. American Psychological association (1994). Publication manual of the American Pscychological association (4th ed.) Washington, DC Author.

[2]. Azar, B. (1997), APA task force urges a harder look at data (on-line). Available:http://www.apa.org/monitor/mar97/stats. html.

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

[4]. Carver, R. P. (1993). The case against statistical significance testing revisited. The journal of experimental Education, 61(4), 287-292.

[5]. Cohen J. (1988). Statistical power analysis for the behavioural sciences (2nd ed). Things I have learned so far. American Psychologist, 45 (12), 1304-1312.

[6]. Cohen J. (1994). The Earth Round (p less than. 05). American Psychologist, 49(12), 997-1003.

[7]. Cortina J. M., & Dunlap. W.P. (1997). On the logic and purpose of significance testing. *Psychologist Method*, 2(2), 161-172
[8]. Crobach L.J. (1975). Beyond the two discipline of Psychology *Psychologist, 30, 116-127*
[9]. Falk, R., & Greenbaum, C.W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. Theory & *Psychology, 5, 75-98*
[10]. Frick, R. W. (1995). Accepting the null hypothesis. Memory and cognition, 33,132-138
[11]. Frick, R. W. (in press). Interpreting statistical testing: process, not populations and random sampling. *Behavior research method, instruments. & Computers.*
[12]. Harris, M. J. (1991). Significance tests are not enough: The role of effect- size estimation in theory corroboration. *Theory & Psychology, l, 375-382.*
[13]. Heldref Foundation. (1997). Guidelines for contributors. *Journal of experimental education,65,95-96.*
[14]. Huberty, C.J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. The Journal of Experimental Education, 61 (4), 317-333.
[15]. Hunter, J. E. (1997). Needed: A ban on the significance test Psychological Science, 8(1), 3-7
[16]. Kirk, R.E. (1996). Practical significance: A concept whose time has come. Educational and Psychological measurement, 56(5), 746-59.
[17]. Levin J.R. (1993), Statistical significance testing from three perspectives. The Journal of Experimental Education, 61(4), 378-382.
[18]. Loftus, G. R., & Masson, M. J. (1994). Using confidence intervals in within-subject designs. Psycholomic Bulletin & Review, 1, 476-490
[19]. Meehl P.E. (1967), Theory-testing in Psychology and physics: A methodological paradox. Philosophy of Science, 34, 103-115
[20]. Morrison, D. E., & Henkel, R. E (Eld). (1970). The significance test controversy. Chicago: Aldine.
[21]. Mulaik, S. A., Raju, N. S. & Harshman, R. A. (1997). There is a time and place for significance testing. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds), What if there were no significance tests? (pp. 65-115). Mahwah, NJ: Eribaum.
[22]. Murphy, K. R. (1997). Editorial, Journal of Applied Psychology, 92, 3-5
[23]. Robinson D.H., & Levin, J. R. (1997). Reflections on Statistical and substantive significance, with a slice of replication. Educational Researcher, 26,(5), 21-26
[24]. Seheneider, A.L. & Darcy,R. E. (1984). Policy implications of using significance tests in evaluation research. Evaluation review. 8, 573- 582
[25]. Shaver, J. P. (1993). What statistical significance testing is, and what it is not. The Journal of Experimental Education, 6(4), 293-316.
[26]. Shea, C. (1996). Psychologists debate accuracy of "significance test." Chronicle of Higher Education, 42(49), A12, A16.
[27]. Snyder, P., & Lawson, S. (1993). Evaluation the results using corrected and uncorrected effect size estimates. The Journal of Experimental Education, 61(4), 334-349.
[28]. Snyder, P. A., & Thompson, B. (in press). Use of tests of statistical significance and other analytic choice in a school psychology journal: Review of practices and suggested alternatives. School Psychology Quarterly.
[29]. Sue, H. K. (1992). Significance testing: Necessary but insufficient topics in Early Childhood Special Education, 12(1), 66-81
[30]. Task Force on Statistical Inference Initial Draft Report (1996). Report to the Board of Scientific Affairs. American Psychological Association (On-line). Available: http//www.apa.org/science/tfsi.html.
[31]. Thompson, B. (1987, April). The use (and misuse) of statistical significance testing. Some recommendations for improved editorial policy and practices. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC. (ERIC Document Reproduction Service No. ED 287 868).
[32]. Thompson, B. (1989). Statistical significance, result important, and result generalizability: Three note-worthy but somewhat different issues. Measurement and Evaluation in Counselling and Development 22, 2-5
[33]. Thompson, B. (1993a). The use of statistical significance tests in research: Bootstrap and other alternatives the Journal of Experimental Education, 61(4). 361-377.
[34]. Thompson, B. (Guest Ed) (1993a). Statistical significance testing in contemporary practice (Special issue). The Journal of Experimental Education, 61(4).
[35]. Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25(2), 26-30.
[36]. Thompson, B. (1997). Editorial policies regarding statistical significance tests: Further comments. Educational Researcher, 26(5), 29-32
[37]. Thompson, B. (1998, January), Why "encouraging" effect size reporting isn't working: The aetiology of researcher resistance to changing practices. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, TX. (ERIC Document ED Number forthcoming)
[38]. Thompson, B. (in press). Canonical correlation analysis. In L. Grimm & P. Yarnold (Eds.) Reading and under-standing multivariate statistics (Vol. 2) Washington, DC: American Psychological Association.
[39]. Thompson, B., & Snyder, P.A. (1997). Statistical significance testing practices in the journal of Experimental Education. Journal Experimental Education, 66, 75-83
[40]. Thompson, B., & Snyder, P. A. (in press). Statistical significance testing and reliability analyses in recent JCD research articles. Journal of Counselling and Development.
[41]. Vacha-Haase, T., & Nilsson, J. E. (in press). Statistical significance reporting: current trends and usages within MECD. Measurement and Evaluation in Counselling and Development.
[42]. Weitzman, R. A. (1984). Seven treacherous pitfalls of statistics, illustrated, Psychological Reports, 54, 355-363