Research on Humanities and Social Sciences
ISSN (Paper)2224-5766 ISSN (Online)2225-0484 (Online)
Vol.5, No.20, 2015

Statistical Power and Effect Size in Educational and Psychological Research Published in Journal of AL-MANARAH for Research and Studies

Dr. Iyad Mohammed Hamadneh
Associate Proffessor, Measurement & Evaluation, Faculty of Educational Sciences – Al al-Bayt University
Director of Faculty Development Center, Mafraq- Jordan

Abstract
This study aimes to assess statistical power that was used in testing the null hypotheses and effect size in the educational and psychological research published in journal of AL-MANARAH for research and studies at Al al-Bayt University from 2010- 2014; and to detect the distribution of the used statistical power according to the different levels of the effect size. The current study included all the educational and psychological research that used (T- test & F- test) in all editions of this period. The number of these research articles was (87) and included (452) statistical tests, where (231) used the statistical (T- test) and (221) used the statistical (F- test). The necessary data was collected to achieve the study aims.Results of the study showed that about (26%) of the hypotheses contained a small effect size, and the average for (T- test) was (0.37), while the average for (F- test) was (0.13). Also, results revealed that around (63%) of the selected hypotheses had a weak statistical power test, around (11%) of the hypotheses had a medium statistical power test, and a round (26%) of the hypotheses had a large statistical power test. Finally, the study recommends that the researchers should provide the Editorial Board in AL-MANARAH journal with the amount of effect size and statistical power test besides statistical significance when they submit their research for publication in the journal.
Keywords: Effect Size, Statistical Significance, Statistical Power, Journal of AL-MANARAH for research and studies

Introduction

There have been different misconceptions about what significance testing is and what it is not (Grice & Barrett, 2014). one should have a good understanding about the basic purpose of statistical significance testing in quantitative research and about what information statistical significance testing provides for education researchers (Hedges & Rhoads, 2009; Jaradat & Judeh, 2005)

Statistical significance testing has caused much confusion among research, unfortunately, the meaning of statistical significance testing has sometimes been lost, and the importance of statistical significance tends to be grossly exaggerated in education research practice chers (Cohen, 1994; Falk & Greenbaum, 1995; Hagan, 1997; Thompson, 1993; Capraro, 2004), leading to frequent distortions or misinterpretations of results (Vacha-Haase & Nilsson, 1998), This seems to have been occurring since the origins of statistical significance in the test of the null hypothesis as proposed by Sir Ronald Fisher more than 70 years ago.

In his seminal article, Kotrlik, Williams & Jabor (2011) stated that few researchers truly understand the meaning of statistical significance. Thompson (1993) echoed this, suggesting that researchers tend to confuse the issues of statistical significance, result importance, and result generalizability.

Indeed, this problem has continued to grow and is increasingly receiving attention at arabian level (Barqi,2012; Osairy, 2012). Larry (2010) confirmed that, "A quick perusal of research journals, educational and psychological statistics textbooks, and doctoral dissertations will confirm that tests of statistical significance continue to dominate the interpretation of quantitative data in educational research".

Statistical significance is a function of effect size which is an index on a common metric that Indicates the magnitude of a relationship or effect, and must be interpreted in the context of the particular study conducted; in order to provid a satisfactory interpretation of results (Cohen, 1988; Durlak, 1995; Weinfurt, 1995).

Also, statistical significance is a function of power , the probability of rejecting a false null hypotheses, is a function of several features such as the magnitude of population parameters, the research design, and the alpha level (Schafer, 1991; Abu-Allam, 2006).

Power is expressed in a number from 0 to 1, where 1 signifies perfect power. A power analysis is recommended before the collection of data. A power of .80 is often recommended. This level of power together with expected effect size and desired alpha level can guide the researcher regarding the appropriate sample size. The researcher must keep in mind that a power of .80 still allows for a 20% Type II error rate (failing to reject a false null). A power of .80 also tends to require a large sample size, often larger than possible for many researchers (Schmidt, 1996). However, low power guides the researcher to conclude that little can be learned from the study (Larry, 2010; Daniel, 1993).

Statistical power is the probability that a test of the null hypothesis of no treatment effect will successfully reject the null hypothesis when a nonzero average treatment effect exists. In simple research designs that use simple random samples, statistical power depends on three things: the significance level of the test; the expected size of the intervention effect (the effect size); and the sample size (Cohen, 1988).

Of course, once the null hypothesis has been rejected, the issue of power becomes moot. Obviously, one simply cannot make a Type II error (i.e., fail to reject a false null hypothesis) once one has already rejected the null, This is why so few published studies report power. Because there is a prejudice against publishing statistically nonsignificant results, and power is not relevant in reports in which a Type II error is already demonstrably impossible, few published studies present power analyses. statistical power is the probability of making the correct decision when a treatment effect actually exists, high statistical power is desirable (Cadeyrn & Brenda, 2014; Schochet, 2008).

Gentry & Scott (2009) mention that too often reports of data analyses in gifted education rely on statistical significance without reporting effect size indices to help interpret quantitative findings. Without a supporting effect size index, erroneous interpretation of results can occur.

Cohen (1988) recommends, one must consider sample size, anticipated effect size and the significance criterion size together with power in designing an experiment. Since these four parameters are interrelated. However, two of these parameters, effect size and the significance criterion, are not freely alterable by the experimenter. Effect size is determeed by nature and the scientific community has established the 5% or 1% confidence limits as conventions that the analyst dare not change; in seeking an acceptable experimental design. Also, he referred to the criteria for judging the value of the effect size computed by Cohen's Standard (d), Where he considered it a small at (0.2); medium at (0.5) and large at (0.8).

Different measures for effect size have been developed over the decades (Kirk, 1996; Snyder & Lawson,1993; Grice & Barrett, 2014) provided useful and practical summaries of those measures, and categorized the variety of effect-size measures into two broad categories: measures of effect size (according to group mean differences) and measures of association strength (according to proportion of variance accounted for), it can be represented by R2, d, eta2, omega2, epsilon2, and others.

The researcher review studies related to statistical significance; Statistical Power and Effect Size; to benefit from their procedures and arranged from oldest to newest.

Daniel (1993) study examines statistical power in music education by taking an in-depth look at quantitative articles published in the "Journal of Research in Music Education" between 1987 and 1991, inclusive. Of the 109 articles of the period, 78 were quantitative, with both parametric and nonparametric procedures considered. Sample sizes were those reported by the authors. Effect sizes were estimated according to the guidelines developed by J. Cohen (1988), and his power analysis tables were used. The overall median power for the articles was 0.13 for detecting small effects, 0.64 for detecting medium effects, and 0.97 for detecting large effects. findings of the study suggest that attention should be placed on a priori power analyses of research designs. Adequate sample sizes should be chosen and a greater understanding and application of the concept of effect size is needed in music education research.

Schochet (2005) study examines issues related to the statistical power of impact estimates for experimental evaluations of education programs. The focus is on "group-based" experimental designs, because many studies of education programs involve random assignment at the group level, at the school or classroom level, rather than at the student level. The clustering of students within groups generates design effects that considerably reduce the precision of the impact estimates, because the outcomes of students within the same schools or classrooms tend to be correlated. This, statistical power is a concern for these evaluations. The report is organized into five sections. First, it discusses general issues for a statistical power analysis, including procedures for assessing appropriate precision levels. Second, it discusses reasons that a clustered design reduces the statistical power of impact estimates and provides a simple mathematical formulation of the problem. Third, it presents procedures that can be used to reduce design effects. Fourth, it provides power calculations for impact estimates under various design options and parameter assumptions.

Parker & Hagan-Burke (2007) study An obstacle to broader acceptability of effect sizes in single case research is their lack of intuitive and useful interpretations. Interpreting Cohen's d as "standard deviation units difference" and R[superscript 2] as "percent of variance accounted for" do not resound with most visual analysts. In fact, the only comparative analysis widely supported in single case research (SCR) is "percent of nonoverlapping data." This article explores five alternative interpretations of Cohen's d and R[superscript 2] effect sizes that may be more acceptable to the SCR field. They are: 1- Cohen's Statistical power analysis for the behavioral sciences, 2- Percent of All Nonoverlapping Data, 3- Binomial Effect Size Display, 4- Percentile Rank in Control Group, and 5- A common language effect-size statistic. Each of the five interpretation schemes are applied to a published data set and are evaluated according to (a) intuitive appeal, (b) relevance to visual analysis, (c) ease of calculation, and (d) technical adequacy. Three of the five appear to be improvements over prevailing practice.

Similarly, Schochet (2008) study examines theoretical and empirical issues related to the statistical power of impact estimates under clustered regression discontinuity (RD) designs. The theory is grounded in the causal inference and HLM modeling literature, and the empirical work focuses on commonly-used designs in education research to test intervention effects on student test scores. The main conclusion is that three to four times larger samples are typically required under RD than experimental clustered designs to produce impacts with the same level of statistical precision. Thus, the viability of using RD designs for new impact evaluations of educational interventions may be limited, and will depend on the point of treatment assignment, the availability of pretests, and key research questions.

Larry (2010) study provide a guide to calculating statistical power for the complex multilevel designs that are used in most field studies in education research. For multilevel evaluation studies in the field of education, it is important to account for the impact of clustering on the standard errors of estimates of treatment effects. Using ideas from survey research, the study explains how sample design induces random variation in the quantities observed in a randomized experiment, and how this random variation relates to statistical power, and the sample sizes at the various levels, the effect size, the multiple correlation between covariates and the outcome at different levels, and the heterogeneity of treatment effects across sampling units is illustrated. Both hierarchical and randomized block designs are considered. The study demonstrates that statistical power in complex designs involving clustered sampling can be computed simply from standard power tables using the idea of operational effect sizes: effect sizes multiplied by a design effect that depends on features of the complex experimental design. These concepts are applied to provide methods for computing power for each of the research designs most frequently used in education research.

Kotrlik, Williams & Jabor (2011) study about the effect size in quantitative Agricultural Education Research Journal (JAE). The purposes of this study are to describe the research foundation supporting the reporting of effect size in quantitative research and to provide examples of how to calculate effect size for some of the most common statistical analyses utilized in agricultural education research. The study revealed requires authors to follow the guidelines stated in the publication manual of the American Psychological Association [APA] in preparing research manuscripts, and to utilize accepted research and statistical methods in conducting quantitative research studies, and JAE now requires the reporting of effect size when reporting statistical significance in quantitative manuscripts.

Barqi (2012) study assess reality of the statistical and practical significance in the published papers in Umm Al-Qura University journal for educational, social and humanitarian sciences. The sample of the study is the papers that were published during the period from 1425/1430. Results of the study showed three levels of statistical significance have been used in the process of investigating the assumptions of the study (0.05, 0.01, 0.001), the rate of statistical significant examinations (71.7%), whereas the rate of non- statistical significant examinations (28.3%); There are poor rate of (0.76%) in which the statistical significance is calculated beside the practical significance from the total statistical examinations of the study. The most important recommendations of the study were to put new standards for accepting the scientific studies in the published papers in Umm Al-Qura University journal, through adopting the practical significance beside the statistical one. Also, we should pay attention to link between the concept of statistical and practical significance.

Finally, Cadeyrn & Brenda (2014) study to assess statistical power, control of experiment-wise Type I error, reporting of a priori power analyses, reporting and interpretation of effect sizes, and reporting of confidence intervals. The analyses were based on 333 papers, from which 10,337 inferential statistics were identified. The use, reporting, and interpretation of inferential statistics in nursing research need substantial improvement. Most importantly, researchers should abandon the misleading practice of interpreting the results from inferential tests based solely on whether they are statistically significant (or not) and, instead, focus on reporting and interpreting effect sizes, confidence intervals, and significance levels. Nursing researchers also need to conduct and report a priori power analyses, and to address the issue of Type I experiment-wise error inflation in their studies.

The benefit drawn by the researcher from the previous researches and studies that was because of the importance statistical power test as a factor influence in each of sample size and effect size, It must be signed to power test besides statistical significance in order to get a suitable results; high decision value and more reliable results.

Problem and Questions of the Study

In response to the findings and recommendations of the previous researches and studies that both statistical power and effect size are needed to make sound research decisions; Because the two items serve different purposes, they supplement each other, but do not substitute for one another. The current study therefore sought to assess the statistical power and effect size used in educational and psychological research published in journal of AL-MANARAH for research and studies at Al al-Bayt University from 2010- 2014.

Specifically**,** The current study seeks to answer the following questions:

1. How does the effect size - associated with null hypotheses that have been tested- distributed, according to the levels proposed by Cohen's (d)?
2. How does the statistical power distributed, according to the different levels of the power?
3. How does the statistical power distributed, according to the different levels of the effect size?

Aims of the study

This study aimed to assess statistical power that was used in testing the null hypotheses and effect size in the educational and psychological research published in journal of AL-MANARAH for research and studies at Al al-Bayt University from 2010- 2014; and to detect the distribution of the used statistical power according to the different levels of the effect size.

Importance of the study

The importance of the current study represents in raising the researchers attentions about the usefull of using statistical power when they test the null hypotheses in their research instead of relying on statistical significance only; The importance of statistical significance levels and it's relationship of practical significance from one hand, and the statistical power from other hand, and the effect of this on the decision accuracy in reject or accept the null hypothesis; Will Provide workers and Editorial Board in AL-MANARAH journal with amount of quality and quantity information about the statistical power tests used in researches published in this journal; Finally, the results of the current study may be provide recommendations and suggestions to strengthen the hypotheses testing results and the confidence in the researches results.

Operational definitions

- **Al-MANARAH for research and studies**: Is a blind academic research joirnal issued by Al- albayt University, Mafraq, Jordan, and is published by the deanship for acadimic research at Al- albayt University. The journal publishes genuine research articles and welcoms original research on current topics based on recent theoritical developments and latest international scholarship in the arts, humanities, social & educational sciences, law, religion and theology, business and finance. Manuscripts should be submitted in English or Arabic. Decisions are made by the Editorial Board based on the referees' reports.
- **Statistical Power**: Is the probability that a test of the null hypothesis of no treatment effect will successfully reject the null hypothesis when a nonzero average treatment effect exists. In the current study, it keeps in mind that a power of .80 still allows for a 20% Type II error rate (failing to reject a false null).
- **Effect Size**: Is the degree of association between and effect (e.g., a main effect, an interaction, a linear contrast) and the dependent variable; in the current study effect sizes defined as "small, d = .2," "medium, d = .5," and "large, d = .8" in (T-test); and Eta squared in (ANOVA analysis).

Limitations of the Study

The study was limited to educational and psychological research published in journal of AL-MANARAH for research and studies at Al al-Bayt University from 2010- 2014, which has been used (T- test for two independent) or (F- test in analysis of variance), or both. Also, generalizing the results of the study depend on the distribution of effect size according to Cohen's Standard (small at (0.2); medium at (0.5) and large at (0.8)); and according to the five statistical power levels.

Methodology of the Study

Participants

The population of this study consisted of all educational and psychological research published in journal of AL-MANARAH for research and studies at Al al-Bayt University from 2010- 2014, which has been used (T- test for two independent) or (F- test in analysis of variance), or both. The number of these research have been reached (87) studies.

The study sample represented (87) studies within the sample available, which Which contained a statistically tests, of which (231) were used statistical (T) and (221) used the statistical (F).

The study procedures:

The researcher made a review of studies that have been published in all volumes and issuses of AL-MANARAH Journal for research and studies at Al al-Bayt University from the period 2010- 2014, were necessary data distribution in private tables prepared for that purpose. These data included the statistical test user type; the level of statistical significance adopted by the researcher in the study; the result of statistical test; and data calculates the effect size, and then extract the statistical power of the test from Cohen's tables.

Cohen's Standard (d) values have been calculated as indicators for effect size of impact depending on the statistical test used to examine the null hypothesis. Where d: is the difference between the means, M1 - M2, divided by standard deviation, as a scale for effect size in the case of (T- test) for two independent groups.

Measures of effect size in ANOVA (F- test) are measures of the degree of association between and effect (e.g., a main effect, an interaction, a linear contrast) and the dependent variable. They can be thought of as the correlation between an effect and the dependent variable. If the value of the measure of association is squared it can be interpreted as the proportion of variance in the dependent variable that is attributable to each effect. The used measure of effect size in ANOVA is: Eta squared.

Results and Discussion

To answer the first question, " How does the effect size - associated with null hypotheses that have been tested- distributed, according to the levels proposed by Cohen's (d)" ?. Effect size were measured according to formulas that depend on the type of statistical tests used to examine the null hypothesis in each study.

Table (1) shows the number of hypotheses (significance and non- significance) distributed according to statistical test and effect size.

Table (1): Distribution of the hypotheses (significance and Non- significance) according to statistical test and effect size.

Effect Size / Statistical test	Less than 0.2		0.2- 0.49		0.50- 0.79		More than or equal 0.8		Total
	Sig.	Non.	Sig.	Non.	Sig.	Non.	Sig.	Non.	
T- test	12	99	13	47	17	19	13	11	231
F- test	35	138	12	8	7	2	10	9	221

Data in the table (1) shows clearly that the effect size was four levels, while Cohen (1988) suggested three levels (small at (0.2); medium at (0.5) and large at (0.8)); but when effect size where measured according to the three Cohen's levels, researcher found that there is an effect size less than (0.2), So it has been added a fourth level to Cohen's levels. And it is striking that the number of hypotheses that contained effect size less than 0.2 reached (111) hypotheses out of (231) hypotheses for (T-test), the equivalent of (48%) from the ratio of total hypotheses. Number of hypotheses that contained "Small" effect size reached (60) hypotheses out of (231) hypotheses for (T-test), the equivalent of (26%) from the ratio of total hypotheses. Number of hypotheses that contained "Medium" effect size reached (36) hypotheses with ratio of (16%). As for the hypotheses that contained "Large" effect size reached (24) hypotheses with ratio of (10%).

As for the statistical test (F), it is striking that the number of hypotheses were distributed into four levels. Number of hypotheses that contained effect size less than 0.2 reached (173) hypotheses out of (221) hypotheses for (F-test), the equivalent of (78%) from the ratio of total hypotheses. Number of hypotheses that contained "Small" effect size reached (20) hypotheses with ratio of (9%), and (9) hypotheses with ratio of (4%) contained "Medium" effect size, whereas (19) with ratio of (9%) contained "Large" effect size.

It should be noted that the reason why the effect size was less than the Small level of effect size (0.2) is the sample size; where the increase in the participants (sample size) would reject the null hypothesis even if it is a small effect size (Daniel, 1993; Gentry & Scott, 2009; Grice & Barrett, 2014).

As shown by the results of the current study that the average of effect size for (T-test) reached (37%). Whereas, the average of effect size for (F-test) reached (13%). This means that the practical significance for (T-test) was Medium, and for (F- test) was weak (Cadeyrn & Brenda, 2014). Also, it referred that all researchers only take statistical significance into account, to infer the differences between independent variables levels; the interpretation of their results, and make their own decisions, without reference to the practical significance. The result of the current study are consistent with Barqi (2012) study, and Kotrlik, Williams & Jabor (2011) study.

To answer the second question, "How does the statistical power distributed, according to the different levels of the power " ?. The statistical power was divided into four intervals (levels).

Table (2) shows the number of hypotheses (significance and non- significance) distributed according to statistical test (T & F) and the four intervals of statistical power.

Table (2): Distribution of the hypotheses (significance and Non- significance) according to statistical test and statistical power intervals.

Power Intervals	Sig.	Non.	Sum	Total
0- 0.50	11 (39)	101 (135)	112 (174)	286
0.501- 0.75	6 (9)	19 (7)	25 (16)	41
0.751- 0.80	4 (4)	3 (1)	7 (5)	12
0.801- 1	41 (13)	48 (11)	90 (23)	113
Total	62 (65)	171 (154)	234 (218)	452

- **The number outside the brackets indicates T-test and inside it indicates (F- test): T(F).**

Table (2) shows clearly that nearly (63%) of the examined hypotheses had statistical power (0.5) or less, i.e the probability that a test of the null hypothesis of no treatment effect will successfully reject the null hypothesis did not exceed the half. As if the researchers depend on their decisions to accept or reject the null hypothesis to the possibility of the appearance of the image and writing when throwing the coin instead of relying on statistical test, which raises the suspicion and lack of confidence in their findings (Daniel, 1993).

Also, it appears from Table (2) that only (12%) from the examined hypotheses were significance and it's test power was large, which means the safety and strength of the decisions, and this average is very modest. And (9%) from the examined hypotheses had medium statistical power test, and around (26%) from the hypotheses had large statistical power test.

Statistical tests which were non- significance and had small power reached (52%) from the study sample; This may be due to the absence of differences, or differences were exist, but the weakness of the test power prevented disclosure. Whereas, statistical tests which were significance and had small power reached (11%) from the study sample; This means that the scientific value of the decisions based on them, is questionable.

As can be seen from Table (2) that the examined hypotheses which were non-sig. and had large statistical power reached (13%), which means that there are differences or effect for the independent variable, but the statistical that was used failure in detecting these differences; This may be due to the weakness of statistical design that had been used in examining the null hypotheses (Schochet, 2008), or due the researcher violation for some the assumptions regarding the use of appropriate statistical, where the assumptions violation leads to amplify type I error, it drives the researcher to reject the null hypothesis which is correct. Conversely, the researcher can accept the null hypothesis which is wrong, as a result of amplification type II error (Hedges & Rhoads, 2009).

To find out the results of the third question, " How does the statistical power distributed, according to the different levels of the effect size"?. The statistical power test (T & F) was divided into four intervals (levels), whereas, effect size was divided into three levels, according to Cohen's Standard (d), with note that the table not include the number of hypotheses wich had effect size less than (0.2), as shown in Table (3).

Table (3): Distribution of hypotheses (significance and non- significance) according to effect size levels and statistical power levels.

Effect Size	Small		Medium		Large		Total
Power Intervals	Sig.	Non.	Sig.	Non.	Sig.	Non.	
0- 0.50	0 (5)	14 (3)	0 (4)	2 (0)	1 (1)	0 (0)	17 (13)
0.501- 0.75	1 (3)	10 (4)	0 (0)	0 (1)	1 (2)	0 (0)	12 (10)
0.751- 0.80	2 (1)	4 (1)	0 (1)	1 (0)	0 (1)	0 (0)	7 (4)
0.801- 1	10 (3)	18 (0)	16 (2)	17 (1)	12 (6)	11 (9)	84 (21)
Total	13 (12)	46 (8)	16 (7)	20 (2)	14 (10)	11 (9)	120 (48)

- **The number outside the brackets indicates T-test and inside it indicates (F- test): T(F).**

Data in Table (3) shows that the hypotheses which had Small effect size reached (49.2%) from all the examined hypotheses by (T- test); also only (50.8%) of the hypotheses achieved cohen's measure. As for the hypotheses which had medium effect size reached (92%); and the hypotheses which had large effect size reached also (0.92%), this leads to the importance of effect size in determining the statistical power test. This result consistent with Barqi (2012) study, and Cadeyrn & Brenda (2014) study.

From Table (3), we also note that the hypotheses which had Small effect size reached (42%) from all the examined hypotheses by (F- test); also only (15%) of the hypotheses achieved cohen's measure. As for the hypotheses which had medium effect size or large effect size reached (61%). Thus, this result confirmed the

importance of effect size in determining the statistical power test. This result consistent with Jaradat & Judeh (2005) study.

To study the relationship between effect size and statistical power test, Chi Square value was computed for each test (T & F) and reached (32.6 & 24.53) respectively. And it is sig. values at sig. level (0.05). This means there is a strong relationship between effect size and statistical power test.

From previous results, it can be said that that all researchers only take statistical significance into account, to infer the differences between independent variables levels; the interpretation of their results, and make their own decisions, without reference to the practical significance. Also, The role of statistical power test is complementary to the role of effect size, and there is there is a strong relationship between them. Thus, the results of the current study highlighted the usefull of using statistical power when the researchers test the null hypotheses in their research instead of relying on statistical significance only. Also, it highlighted the importance of statistical significance levels and it's relationship of practical significance from one hand, and the statistical power from other hand, and the effect of this on the researchers decision accuracy in reject or accept the null hypothesis, in order to understand their research results in a better way.

Recommendations and Suggestions

In light of the current study findings, the study recommends the researchers to refer the amount of effect size besides statistical significance in order to understand their research results in a better way. Also, it recommends the Editorial Board in journal of AL-MANARAH for research and studies that the researchers should provide them with the amount of effect size and statistical power test besides statistical significance when they submit their research for publication in the journal. The researcher suggests to make a similar study on the same subject to other Educational Journals and Theses at different Universities of Jordan, and make a comparison between their results.

References

Abu Allam, R. (2006). Effect size of the experimental treatment and significance of the statistical significance. Educational Journal, ISNN393599, 5- 150.

Barqi, T. (2012). Reality the Statistical and Practical Significance in the Published papers in Umm Al-Qura University journal for Educational, Social and Humanitarian Sciences, during the period of 1425-1430 H. Unpublished Master Thesis, Um Al-qoura University, Saudi Arabia.

Cadeyrn, G., & Brenda, H. (2014). Power, effects, confidence, and significance: An investigation of statistical practices in nursing research. International Journal of Nursing Studies, 51(5), 795–806.

Capraro, R. (2004). Statistical significance, effect size reporting, and confidence intervals: Best reporting strategies. Journal for Research in Mathematics Education, 35(1), 57-62.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd Ed.). New York: Erlbaum.

Cohen, J. (1994). The earth in round ($p < 0.05$). American Psychologist, 49, 997- 1003.

Daniel, T. (1993). A statistical power analysis of the quantitative techniques used in the "Journal of Research in Music Education," 1987 through 1991. Paper presented at the Annual Meeting of the Mid-South Educational Research Association (New Orleans, LA, November 10-12, 1993).

Durlak, J. (1995). Understanding meta-analysis. In L. G Grimm & P. R. Yarnold (Eds.), Reading and understanding multivariate statistics (pp. 319-352). Washington, DC: American Psychological Association.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. Theory & Psychology, 5, 75-98.

Gentry, M., & Scott, P. (2009). Effect sizes in gifted education research. Gifted Child Quarterly, 53(3), 219- 222.

Grice, J., & Barrett, P. (2014). A note on Cohen's overlapping proportions of normal distributions. Psychological Report, 115(3), 741- 747.

Hagan. R. L. (1997). In praise of the null hypothesis statistical test. American Psychologist, 52, 15-23.

Hedges, L.V., and Rhoads, C. (2009). Statistical power analysis in education research. U.S. Department of Education. Washington, DC: National Center for Special Education Research, Institute of Education Sciences.

Jaradat, D., and Judeh, M. (2005). The power, effect size and sample size of published research in Abhath Al-Yarmouk Humanities and Social Science Series. Jordan Journal of Science in Education, 1(1), 21- 29.

Kirk, R. (1996). Practical significance: A concept whose time has come. Educationaland Psychological Measurement, 56, 746-759.

Kotrlik, J., Williams, H., & Jabor, M. (2011). Reporting and interpreting effect size in quantitative Agricultural Education Research. Journal of Agricultural Education, 52(1), 132-142.

Larry, H. (2010). Statistical power analysis in education research. Christopher Rhoads, Department of Education, Northwestern University, U.S.

Osairy, A. (2012). (Methodological / Statistical) difficulties of scientific research for post graduate students in Faculty of Education in Um Al-qoura University. Master Thesis published, Um Al-qoura University, Saudi Arabia.

Parker, R., & Hagan-Burke, S. (2007). Useful effect size interpretations for single case research. Behavior Therapy, 38(1), 95-105.

Schafer, W. D. (1993). Power analysis in interpreting statistical non significance. Measurement and Evaluation in Counseling and Development, 23, 146-148.

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

Schochet, P. (2005). Statistical power for random assignment evaluations of education programs. U.S. Department of Education. Washington, DC: Institute of Education Sciences.

Schochet, P. (2008). Statistical power for regression discontinuity designs in education evaluations [Online]. Available from: ED Pubs: http://ies.ed.gov/ncee/.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. Journal of Experimental Education, 61, 334-349.

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

Vacha-Haase, T., & Nilsson, J. (1998). Statistical significance reporting: Current trends and uses in MECD. Measurement & Evaluation in Counseling & Development (American Counseling Association), 31(1), 12- 46.

Weinfurt, K. P. (1995). Multivariate analysis of variance. In L. G. Grimm & P. R.

Yarnold (Eds.), Reading and understanding multivariate statistics (pp. 245-276). Washington, DC: American Psychological Association.