

قوة الاختبارات الإحصائية وحجم الأثر في البحوث
التربوية المنشورة في مجلة جامعة القدس
المنشورة للأبحاث والدراسات

د. حمدي يونس أبو جراد
قسم علم النفس
كلية التربية - جامعة القدس المفتوحة

قوة الاختبارات الإحصائية وحجم الأثر في البحوث التربوية المنشورة في مجلة جامعة القدس المفتوحة للأبحاث والدراسات

د. حمدي يونس أبو جراد
قسم علم النفس
كلية التربية - جامعة القدس المفتوحة

الملخص

هدفت الدراسة الحالية إلى التعرف على قوة الاختبارات الإحصائية في البحوث التربوية المنشورة في مجلة جامعة القدس المفتوحة للأبحاث والدراسات واشتملت الدراسة على جميع البحوث التربوية التي استخدمت الاختبارات الإحصائية (F)، (t) في جميع أعداد مجلة جامعة القدس المفتوحة للفترة بين (٢٠٠٢ - ٢٠١٠) وقد بلغ عدد هذه الدراسات (٧٤) دراسة احتوت على (٤٤٥) اختباراً إحصائياً منها (٢٢٦) استخدم الإحصائي (t) و(٢١٩) استخدم الإحصائي (F). جمعت البيانات اللازمة لحساب حجم الأثر وقوة الاختبار الإحصائي وأشارت النتائج إلى أن حوالي (٢٧٪) من الفرضيات احتوت على حجم أثر صغير وقد بلغ متوسط حجم الأثر لاختبار (t) (٠,٣٨)، بينما بلغ متوسط حجم الأثر لاختبار (F) (٠,١٢). وأشارت النتائج أيضاً أن حوالي (٧١٪) من الفرضيات التي اختبرت كانت قوة الاختبار الإحصائي فيها ضعيفة وحوالي (٦٪) من الفرضيات كانت قوة الاختبار الإحصائي فيها متوسطة وحوالي (٢٣٪) من الفرضيات كانت قوة الاختبار الإحصائي فيها عالية.

الكلمات المفتاحية: قوة الاختبار، حجم الأثر.

The Power of the Statistical Tests and the size of Effect of the Educational Researches Published in Al Quds Open University Journal for Researches and Studies

Dr. Hamdy Y. Abu Jarad
Faculty of Education
Al Quds Open University-Gaza

Abstract

The current study aimed at identifying the power of the statistical tests in the educational researches published in Al-Quds Open University Journal for researches and studies. The study included all the educational researches that used the statistical (t) & (F) tests in all editions of Al-Quds Open University Journal for the period between (2002-2010). The number of these studies was (74) and included (445) statistical tests, where (226) used the statistical (t) and (219) used the statistical (F).

The necessary data was collected to calculate both the size of the effect and the power of the statistical test. The results pointed out that about (27%) of the hypotheses contained a small size and the average for (t) test was (0.38) while the average for the (F) test was (0.12).

The results indicated that around (71%) of the selected hypotheses had a poor statistical power test, around (6%) of the hypotheses had a medium statistical test power, and (23%) of the hypotheses had a high statistical power test.

Key words: power test, effect size.

قوة الاختبارات الإحصائية وحجم الأثر في البحوث التربوية المنشورة في مجلة جامعة القدس المفتوحة للأبحاث والدراسات

د. حمدي يونس أبو جراد
قسم علم النفس
كلية التربية - جامعة القدس المفتوحة

المقدمة

يبنى الباحثون في العلوم الإنسانية والاجتماعية نوعين من الفرضيات تعرف الأولى بالفرضية الصفرية Null Hypothesis وهي التي تشير إلى أن الاختلاف في قيمة احصائي العينة المأخوذة من مجتمع التجريب عن قيمته في المجتمع الأصلي هو اختلاف يرجع إلى الصدفة أو أن الفرق بين احصائي عينة ونفس الاحصائي في عينة أخرى هو فرق صدفة. أي أنه فرق ليس له قيمة وأنهما بالفعل لا ينسبان إلى مجتمعين مختلفين. وتسمى الثانية بالفرضية البديلة Alternative Hypothesis وهي الفرضية التي يحاول الباحث إثباتها مقابل رفضه للفرضية الصفرية وتسمى الفرضية البديلة بفرضية البحث حيث يفترض الباحث أن نتائج البحث تدعم ما كان قد توقعه من خلال الإطار النظري الذي كونه عن المشكلة مدار البحث والذي يشير عادة إلى وجود الفرق أو إلى عدم التساوي المشار إليه في الفرضية الصفرية. (عودة، وملكاوي، 1992).

ومن المعروف أن عملية اختبار الفرضيات الإحصائية تجيب عن سؤال يتعلق بكون البيانات التي جمعها الباحث كافية لرفض الفرضية الصفرية (دالة إحصائية) أو غير كافية لرفض تلك الفرضية الصفرية (غير دالة إحصائية) بمعنى أن الجواب عن الأسئلة الإحصائية يكون على الصورة: نعم توجد دلالة إحصائية أو لا توجد دلالة إحصائية (Thompson, 1988). ويقصد بالدلالة الإحصائية احتمال رفض الفرضية الصفرية باستخدام الأساليب الإحصائية. أما الدلالة العملية فيقصد بها أن تكون الفروق الإحصائية أو العلاقات بين المتغيرات كبيرة إلى درجة تبرر عملية الأخذ بنتائجها. ويشير الصياد (1988) إلى أن الدلالة الإحصائية شرط ضروري لصناعة قرار تربوي أو نفسي. ولكنه ليس كافياً؛ فالكفاية تتحقق بحساب قوة العلاقة بين المتغير المستقل، والمتغير التابع، وقوة العلاقة هي ما يسمى بالدلالة العملية (Hays, 1973).

ومن الملاحظ في البحوث التربوية افتقارها إلى دراسة الدلالة العملية للبحث ومناقشتها واقتصارها على المقاييس الإحصائية التقليدية مما يقلل من أهمية النتائج فالأصل هو تقدير الدلالة العملية ، خاصة إذا علمنا أن قوة الاختبار الإحصائية تكمن في علاقتها بالدلالة الإحصائية وبالدلالة العملية وأن تحديد الدلالة العملية للفرق الإحصائي يعتبر مقياساً إضافياً يعكس أهمية هذا الفرق.

وبناء عليه فإن الاعتماد على الدلالة الإحصائية في اتخاذ القرار بقبول الفرضية الصفريّة أو برفضها يوقع الباحث في نوعين من الخطأ:

الخطأ من النوع الأول (Type I error) ويحدث هذا الخطأ إذا رفض الباحث الفرضية الصفريّة وهي صحيحة ويطلق عليه مستوى الدلالة الإحصائية (α). وتعتبر القيمة المرتفعة لمستوى الدلالة عن نقصان احتمال قبول الفرضية الصفريّة ، بينما تعبر القيمة المنخفضة لمستوى الدلالة الاحصائية عن زيادة احتمال قبول الفرضية الصفريّة.

الخطأ من النوع الثاني (Type II error) ويحدث الخطأ من النوع الثاني عند قبول الباحث للفرضية الصفريّة وهي خطأ ويرمز له بالرمز (β) . ويعتمد الخطأ من النوع الثاني جزئياً على مستوى الدلالة، أما في حالة قبول الباحث للفرضية الصفريّة وهي صحيحة فإن القرار المتخذ يكون سليماً وهذا ما يعرف بفترة الثقة ويرمز له بالرمز ($1 - \alpha$). وأما إذا قام الباحث برفض الفرضية الصفريّة وهي خاطئة فإن القرار المتخذ يكون سليماً أيضاً وهذا ما يعرف بقوة الاختبار الإحصائي ويرمز له بالرمز (1 - β).

ويحدد الباحث مقدار الخطأ من النوع الأول الذي يمكن ارتكابه في فحص الفرضية الصفريّة ويكون احتمال الوقوع في مثل هذا الخطأ مساوياً لقيمة مستوى الدلالة الإحصائية (α). وغالباً ما يستخدم الباحثون مستويات دلالة إحصائية مساوية (0.05) و(0.01) ، وهو أمر متفق عليه وليس له دليل علمي أو منطقي (Winer, Brown & Michels, 1991).

ويفترض أن يحدد الباحث مستوى الدلالة الإحصائية التي يستخدمها في فحص فرضياته قبل القيام بجمع البيانات وإجراء التحليلات الإحصائية. (Winer, 1971).

وقد ناقش هوبكينز (Hopkins, 2011) مفهوم مستويات الدلالة الإحصائية وتطورها بقوله: ما هو الخاص في القيمة 0.05؟ في الحقيقة لا شيء ، شخص ما قرر أن هذه القيمة معقولة. ولذا فإننا الآن جامدون عندها. كما أصبحت ($p > 0.01$) أيضاً تقليداً لتحديد الدلالة الإحصائية فكلما القيمتين من الآثار القديمة للأيام التي سبقت استخدام الحاسب الآلي، عندما كان من الصعب حساب القيمة الدقيقة لـ (p) لإحصاء الاختبار، وبدلاً من

ذلك استخدم الناس الجداول لقيم إحصاء الاختبار المقابلة لقيم قليلة وعشوائية لـ (p) وهي $(0,05)$ ، $(0,01)$ وفي بعض الأوقات $(0,001)$.

لقد أصبحت هذه القيم مقدسة كعتبة حدية للقيم التي تحدد الدلالة الإحصائية، وعادة ما يكون على الباحث أن ينص على أي واحدة يستخدم. فعلى سبيل المثال، إذا نصّ الباحث على أنّ مستوى الدلالة هو $0,05$ ، إذاً يسمح له باعتبار أي نتيجة لها قيمة (p) تساوي أو أقل من $0,05$ بأنها دالة إحصائياً.

كما اعتبر ماكلين وايرنيسست (Mclean & Ernest, 1997) أنّ مستوى الدلالة $(0,05)$ أصبح حارس بوابة عشوائي للمعرفة العلمية، بحيث أنّ النتائج التي لا تقابل المعيار ينظر على أنها غير مهمة، وأنّ هذه الآلية في تحديد مستوى الدلالة الإحصائية ثم اختبار الدلالة الإحصائية للوصول إلى القرار الإحصائي، تبدو محددة لصنع القرار مما يعطي شعوراً بعدم الموضوعية العلمية في اتخاذ القرار.

ويرى بعض الباحثين المهتمين بمشكلات اختبارات الدلالة الإحصائية في البحث التربوي ضرورة إعادة النظر في التقليد العلمي المتبع حالياً في البحث التربوي الذي يركز على اختبار الفرضيات من خلال استخدام اختبارات الدلالة الإحصائية، وقد عبر عن مثل هذا الرأي كيبستر (Keaster, 1988) حيث ذكر أن علماء الاجتماع يجب أن يعملوا بشكل أفضل لإعادة فحص أغراضهم من استخدام اختبارات الدلالة الإحصائية، وأن يحاولوا أن يستنبطوا طرقاً أفضل للوصول إلى تلك الأغراض بدلاً من الاستمرار في اللجوء إلى تكتيكات هي في أحسن أحوالها تضلل البحث الامبريقي.

وقد ذهب كارفر (Carver, 1993) إلى حد القول بأن عدم استعمال اختبارات الدلالة الإحصائية سيكون أفضل للبحث التربوي حيث ذكر بأنه إذا استطاع أحد ألا يستخدم الدلالة الإحصائية مرة أخرى لتحديد دلالة الفرق، فإن الباحثين سيضطرون لاستخدام التصميمات التي تظهر بوضوح الأهمية العلمية للفرق، ويرى بأنه بدون استخدام الدلالة الإحصائية، يضطر الباحثون إلى تلمس مشكلات الاستدلال العلمي بدلاً من الالتزام باختبار الدلالة الإحصائية، وأن البحث التربوي سوف يكون أفضل إذا أوقف اختبار الدلالة الإحصائية لنتائجه.

وقد ذهب كرونباخ إلى القول بأن الوقت قد حان للتخلي عن الفرضية الصفرية، بينما وصف شلمان حال الباحثين التربويين في اعتمادهم على اختبارات الدلالة الإحصائية بقوله إن الباحثين التربويين أسلموا أنفسهم لعبودية اختبار الفرضية الإحصائية (Keaster, 1988).

لقد دعت تلك النظرة لاختبارات الدلالة الإحصائية العديد من الباحثين إلى القول بخطر اختبارات الدلالة الإحصائية والدعوة لحظرها ومنعها من النشر فقد اعتبر سميث: "اختبار الدلالة الإحصائية معيقاً لنمو المعرفة العلمية، ولا يقوم أبداً بأي إسهام إيجابي" (Keaster, 1988).

ولقد ذهب إلى هذا الاتجاه نيكس وبارنت (Nix & Barnette, 1998) حيث اعتبروا فقدان اختبارات الفرضيات الصفوية الإحصائية لن يعيق نمو المعرفة العلمية، بل سيتقدم بالمعرفة ولقد أشارا إلى ضرورة التخلي عن تقاليد اختبارات الدلالة الإحصائية .

كما واجهت اختبارات الدلالة الإحصائية انتقاداً آخرًا بسبب اعتمادها المفرط على حجم العينة وعدم القدرة على توضيح معنى النتائج والباحثون الذين لديهم خبرة العمل مع عينات كبيرة يدركون بأن كل الفرضيات الصفوية سوف يتم رفضها عند مستوى معين من حجم العينة . ولقد حذر (Kirk, 1996) من التفسيرات الخاطئة لمعنى الدلالة الإحصائية والاستعاضة عنها بالدلالة العملية لأثر المتغير المستقل على المتغير التابع أو حجم الأثر.

ويتم حساب حجم الأثر من خلال حساب الفروق في الأوساط الحسابية باستخدام الانحراف المعياري كوحدة قياس لمقدار الفرق بين تلك المتوسطات كما هو الحال في اختبار (t) للعينات المترابطة والمستقلة (Cohen, 1988). أما في حالة المجموعتين الضابطة والتجريبية فتعتبر نقطة الإسناد هي وسط المجموعة الضابطة وحجم الأثر هو العلامة المعيارية للوسط الحسابي لأداء المجموعة التجريبية مقارنة بأداء المجموعة الضابطة (Glass & Hopkins, 1984).

وفي حال تحليل التباين فإن حجم الأثر يعبر عن العلاقة بين المتغير المستقل أو المتغيرات المستقلة من جهة . والمتغير التابع أو المتغيرات التابعة من جهة أخرى . ويتم ذلك من خلال حساب مقدار التباين في المتغير التابع أو المتغيرات التابعة الذي يمكن تفسيره من خلال المتغير المستقل . أو المتغيرات المستقلة (Mcclain, 1995).

أي أن قيمة حجم الأثر تعبر عن الدرجة التي يمكن التنبؤ بها بالمتغير التابع من خلال المتغير المستقل (Mahadevan, 2000)

ومن المؤشرات المستخدمة لقياس حجم الأثر Effect size مؤشر كوهن (d) ومربع ابتا (η²) لفردمان Friedman ومربع أوميغا (ω²) لهيز Hays ومربع ايبلسون (ξ²) لكوهن. (Fagley & Mckinney, 1983).

ويرى كيلو (Kellow, 1998) أن أول ما يميز حجومات التأثير هو إمكانية تفسيرها بشكل

مستقل عن حجم العينة، وهي مفيدة في تحديد المقدار الفعلي للفروق بين متوسطات المجموعات أو درجة التلازم بين المجموعات. ولذا فإنها تزود الباحثين بمعلومات تتجاوز حدود اختبارات الدلالة الإحصائية، وتمنحهم فهماً واضحاً للبيانات التي بين أيديهم. ولقد أوجز هوستون (Huston, 1993) فوائد مقاييس حجم التأثير على النحو التالي:

١. حجم التأثير يشير إلى درجة وجود الظاهرة في المجتمع بمقياس متصل، بحيث يعني الصفر عدم وجود الظاهرة.

٢. تزود الباحثين بمؤشرات للدلالة العملية بخلاف اختبارات الدلالة الإحصائية.
٣. يمكن استخدامها في المقارنة الكمية بين نتائج دراستين أو أكثر كما هو مستخدم في التحليل البعدي *Meta Analysis*.
٤. يمكن استخدامها في تحليل القوة الإحصائية لتحديد عدد العناصر المطلوبة في دراسة معينة.

ولقد أشار كوهين (Cohen, 1988) إلى معايير للحكم على قيمة حجم الأثر المستخرجة بواسطة مؤشر كوهن (d) حيث اعتبره صغيراً عند القيمة (٠,٢) ومتوسطاً عند القيمة (٠,٥) واعتبره كبيراً عند القيمة (٠,٨).

ويعتبر تحليل القوة الإحصائية أحد الاختبارات المكمل لاختبارات الدلالة الإحصائية حيث تعبر عن احتمالية رفض الفرضية الصفرية بطريقة صحيحة، ويفيد تحليل القوة الإحصائية في تفسير النتائج شبه المعدومة *null results* كما أن معرفة احتمالية الحصول على نتائج دالة إحصائياً تفيد الباحث في تحديد مدى الجدوى من تنفيذ الدراسة وذلك لتوفير الوقت والجهد والتكلفة.

وتتأثر قوة الاختبار الإحصائي بالعديد من العوامل أبرزها حجم العينة فكلما زاد حجم العينة زادت قوة الاختبار الإحصائي، وفي هذا الصدد يرى سنايدر ولوسون (Snyder & Lawson, 1993) أن حجم العينة هو العامل الحاسم والأساس في تقرير ما إذا كانت نتائج أي دراسة دالة أو غير دالة إحصائياً، وليس معنى هذا أن حجم العينة هو السبب في زيادة قوة الاختبار الإحصائي وإنما قيمتها مستوى الدلالة (α) وخطأ النوع الثاني (β) وكذلك تباين المجتمع لهما أثر كبير على قوة الاختبار بجانب حجم العينة وحجم الأثر (Weinberger & Goldberg, 1979).

لقد حظيت موضوعات الدلالة الإحصائية والدلالة العملية وحجم الأثر وقوة الاختبارات الإحصائية باهتمام بالغ من قبل الباحثين فقد أجرى نصار (٢٠٠٥) دراسة هدفت إلى تقديم

مفهوم حجم الأثر كأسلوب إحصائي مكملاً لفحص الفرضيات الإحصائية وبرهن الباحث أن حجم الأثر أقل تأثيراً بحجم العينة من الأساليب الإحصائية التي تستخدم لفحص الفرضيات وقدمت الدراسة بعض المؤشرات الإحصائية التي تستخدم للدلالة على قيمة حجم الأثر في حالة استخدام اختبار (t) للعينات المستقلة والمتراطة واختبار (F) في حالة تحليل التباين الأحادي ومعامل ارتباط بيرسون والانحدار البسيط.

وهدفت دراسة جرادات وجوده (٢٠٠٥) إلى التعرف على قوة الاختبار الإحصائي وحجم الأثر وحجم العينة للدراسات المنشورة في مجلة أبحاث اليرموك وتألقت عينة الدراسة من الدراسات التي استخدمت الاختبارات الإحصائية (t) (F) و (tr) المنشورة في العدد الأول من كل مجلد من مجلدات مجلة أبحاث اليرموك سلسلة العلوم الإنسانية والاجتماعية للفترة (١٩٨٥ - ٢٠٠١) وقد بلغ عدد هذه الدراسات (٥٤) دراسة احتوت على (٧٨٥) اختباراً إحصائياً وبينت نتائج الدراسة أنَّ ما يزيد على (٧٠٪) من الفرضيات التي تم اختبارها في الدراسات المنشورة قد تم رفضها وحوالي (٥٢٪) من الاختبارات الإحصائية حققت الحد الأدنى للقوة (٠,٨) حسب معيار كوهن ولم يزد عدد الفرضيات التي اقتربت بحجم أثر كبير على (٢٦,٥)٪ من مجموع الفرضيات التي تم اختبارها وأكثر من (٤٦)٪ من العينات المستخدمة لا يزيد عدد أفرادها على (١٠٠).

وهدفت دراسة حجمات، وعليان (١٩٩٧) إلى التعرف على واقع الدلالات الإحصائية والعملية وقوة الاختبارات الإحصائية المستخدمة في رسائل ماجستير الإرشاد النفسي والتربوي بالجامعة الأردنية وتكونت عينة الدراسة من (١٨٨) فرضية صفرية استخدم في فحصها الإحصائي (F) وذلك في (٢٨) رسالة من رسائل الماجستير التي استخدمت تصاميم تجريبية وشبه تجريبية في مجال الإرشاد النفسي وأشارت نتائج الدراسة إلى أن حوالي (٨٥)٪ من الفرضيات الصفرية استخدم في فحصها مستوى دلالة (٠,٠١) وأن حوالي (٦٣)٪ من الفرضيات المفحوصة التي كانت دالة إحصائياً لم تكن ذات دلالة عملية أما بالنسبة لقوة الاختبار فقد اتضح أن حوالي نصف الاختبارات الإحصائية المستخدمة في فحص الفرضيات الصفرية والدالة إحصائياً تمتعت بقوة اختبار من مستوى متوسط وكبير وأن حوالي (٢٥)٪ من الاختبارات الإحصائية لم تكن دالة إحصائياً وكانت قوتها ضعيفة وأشارت النتائج أيضاً إلى أن (٣,٢)٪ من الفرضيات الصفرية والمفحوصة بالإحصائي (F) كانت دالة عملياً ولكن لم تكن دالة إحصائياً.

وهدفت دراسة صائغ (١٩٩٦) إلى التعرف على واقع الدلالة العملية ومدى تأثيره بحجم

العينة وقوة الاختبار ومستوى الدلالة ونوع الاختبار في رسائل الماجستير التي قدمت في كلية التربية بجامعة أم القرى بمكة المكرمة حتى عام ١٩٩٤ وتم أخذ عينة قصديه من قيم اختباري (t) . (F) الدالة إحصائياً حيث بلغ عددها (1٤٨) قيمة من مجتمع البحث البالغ عدده (١٠٣) رسالة ماجستير وبينت نتائج الدراسة أن (٤٧٪) من الفروق الدالة إحصائياً كانت قيم مربع ايتا ومربع أوميجا فيها منخفضة وتركزت قيم حجم التأثير بمستوياته المختلفة عند قوة اختبار مرتفعة بنسبة (٧٠,٥٪) وأن حجم العينة كان مناسباً عند حجم تأثير صغير وأنه لا يوجد أثر لمستوى الدلالة الإحصائية على الدلالة العملية، وأنه لا توجد علاقة بين الدلالة العملية ونوع البحث تربوي كان أم نفسي.

وهدفت دراسة الصياد (١٩٨٨) إلى التعرف على واقع الدلالة العملية وحجم العينة المصاحبين للدلالة الإحصائية لاختبار (t) في البحث التربوي والنفسي وتكونت عينة الدراسة من (١٥) دراسة في مجال علم النفس والصحة النفسية و(١١) دراسة في مجالات التربية المختلفة بالإضافة إلى (٤) رسائل ماجستير في مجال التربية و(٩) رسائل في مجال علم النفس التعليمي. ولقد بلغ تكرار الاختبار الإحصائي (t) (٤٦٣) مرة . وقد أشارت النتائج إلى أنّ الدلالة العملية (حجم التأثير) لاختبار (t) هي من النوع الضعيف . وأنّ الباحثين يميلون إلى التخفي وراء حجم العينة في سبيل الحصول على دلالة إحصائية . كما أنهم يميلون أيضاً لاستخدام مستويي الدلالة الإحصائية (٠,٠١) و (٠,٠٥) حيث يتم استخدامها بعد جمع البيانات و تحليلها.

مشكلة الدراسة

يلاحظ في كثير من الأبحاث النفسية والتربوية المنشورة في المجالات المتخصصة اعتمادها على الدلالة الإحصائية في قبول أو رفض فرضيات البحث إلا أن النتيجة ذات الطبيعة الثنائية للدلالة الإحصائية (قبول أو رفض) لم تعد ترضى كثيراً من العلماء الذين يستخدمون الفرضية الصفرية لمعرفة مقدار التأثير أكثر من مجرد التعرف على وجوده من عدمه. كما تزايد اعتراف الباحثين بقصور ومحدودية اختبارات الدلالة الإحصائية في تشكيل النتائج العلمية (Palomares, 1990). وقد سبب عدم إدراك الباحثين لمفهوم الدلالة الإحصائية ظهور عدة استخدامات خاطئة لاختبارات الدلالة الإحصائية . ولما كان معيار الدلالة الإحصائية لقبول أو رفض الفرضية الصفرية لا يكفي لاتخاذ قرارات دقيقة وموثوق بها حيث تتأثر الدلالة الإحصائية بالعديد من العوامل أبرزها حجم

العينة حيث يقل الخطأ المعياري كلما كبر حجم العينة وبالتالي تكون الدلالة الإحصائية دالة لحجم العينة أو لمستوى الدلالة بدلاً من أن تكون دالة للفروق العملية بين المتوسطات. فقد اشترطت العديد من المجلات البارزة في مجال علم النفس والعلوم التربوية على الباحثين تضمين قيمة حجم الأثر وقوة الاختبار في أبحاثهم (Tapia & Marsh, 2002).

ويؤكد أرنيمان (Arneman, 2003) على أن هناك (٢٣) مجلة متخصصة في علم النفس التربوي والعلوم ذات العلاقة به كالعلوم التربوية وعلم النفس بما في ذلك مجلتان تعدان من أهم المجلات المتخصصة في هذا المجال تطلب من الباحثين الراغبين في نشر الأبحاث الكمية من خلالها بتضمين أبحاثهم قيمة حجم الأثر وقوة الاختبار، وعلى الرغم من ذلك يلاحظ أن الكثير من الأبحاث التربوية والنفسية المنشورة في المجلات المتخصصة تخلو في كثير من الأحيان من الإشارة إلى الدلالة العملية (حجم الأثر).

ويبني الباحث دراسته على أساس مراجعة دقيقة للأدب يمكنه من افتراض وجود فروق بين المتوسطات أو وجود علاقة بين المتغيرات المستقلة والمتغيرات التابعة مما يعني أنه افترض مسبقاً بوجود حجم أثر مقبول من المحتمل أن يمكنه من رفض الفرضية الصفرية الخاطئة وهذا ما يسمى بقوة الاختبار، ونظراً لأهمية قوة الاختبار كعامل مؤثر في كل من حجم العينة وحجم التأثير فإنه يجب الإشارة إلى قوة الاختبار بجانب الدلالة الإحصائية وذلك للحصول على نتيجة قيمة وقرار ذي قيمة عالية ونتائج أكثر موثوقية.

أهداف الدراسة

تهدف الدراسة الحالية إلى:

- ١- تقييم قوة الاختبارات الإحصائية المستخدمة في فحص الفرضيات الصفرية للبحوث المنشورة في مجلة جامعة القدس المفتوحة للأبحاث والدراسات
- ٢- تقييم حجم الأثر في البحوث المنشورة في مجلة جامعة القدس المفتوحة للأبحاث والدراسات
- ٣- معرفة توزيع قوة الاختبارات الإحصائية المستخدمة حسب المستويات المختلفة لحجم الأثر

أسئلة الدراسة

- ١- كيف يتوزع حجم الأثر المقترن بالفرضيات الصفرية التي تم اختبارها حسب المستويات التي اقترحها كوهن لحجم الأثر؟

- ٢- كيف تتوزع قوة الاختبارات الإحصائية المستخدمة حسب المستويات المختلفة للقوة؟
 ٣- كيف تتوزع قوة الاختبارات الإحصائية المستخدمة حسب المستويات المختلفة لحجم الأثر؟

أهمية الدراسة

- ١- تأتي أهمية الدراسة الحالية من أهمية مستويات الدلالة الإحصائية وافتراضها بالدلالة العملية من جهة وقوة الاختبارات الإحصائية المستخدمة من جهة أخرى وتأثير ذلك على دقة القرار في رفض أو قبول الفرضية الصفرية.
 ٢- تساعد الدراسة الحالية على إثارة اهتمام الباحثين والوعي بمدى أهمية استخدام قوة الاختبارات الإحصائية عند اختبار فرضياتهم الصفرية بدلاً من الاعتماد على الدلالة الإحصائية فقط.
 ٣- رقد الباحثين والقائمين على هيئة تحرير المجلة بمعلومات كمية ونوعية عن قوة الاختبارات الإحصائية للاختبارات الإحصائية المستخدمة في البحوث المنشورة في المجلة .
 ٤- تقديم مقترحات جديدة لتدعيم نتائج فحص الفرضيات مما ينعكس بشكل مباشر على الثقة في النتائج.

محددات الدراسة

اقتصرت الدراسة الحالية على البحوث التربوية المنشورة في مجلة جامعة القدس المفتوحة للأبحاث والدراسات التي استخدمت الاختبارات الإحصائية (t) و(F).

منهجية الدراسة وإجراءاتها:

مجتمع الدراسة

تكون مجتمع الدراسة من جميع الدراسات التي استخدمت الاختبارات الإحصائية (t) المستخدم عادة في اختبار الدلالة الإحصائية للفرق بين متوسطى عينتين مستقلتين و(F) المستخدم عادة في تحليل التباين وقد بلغ عدد هذه الدراسات (٧٤) دراسة وذلك في الفترة (٢٠٠٢-٢٠١٠).

عينة الدراسة

تكونت عينة الدراسة من جميع الدراسات المنشورة في مجلة جامعة القدس المفتوحة

في الفترة (٢٠٠٢-٢٠١٠) والتي احتوت على الاختبار الاحصائي (t) أو (F) أو كليهما وقد بلغ عدد هذه الدراسات (٧٤) دراسة احتوت على (٤٤٥) اختبارا إحصائيا منها (٢٢٦) استخدم الإحصائي (t) و (٢١٩) استخدم الإحصائي (F).

إجراءات الدراسة

تم مراجعة الدراسات المنشورة في جميع أعداد مجلة جامعة القدس المفتوحة للأبحاث والدراسات للفترة بين (٢٠٠٢-٢٠١٠) وتم توزيع البيانات الضرورية في جداول خاصة أعدت لذلك الغرض واشتملت هذه البيانات على نوع الاختبار الإحصائي المستخدم ومستوى الدلالة الإحصائية الذي اعتمده الباحث ونتيجة الاختبار الإحصائي والبيانات الخاصة بحساب حجم الأثر ومن ثم استخراج قوة الاختبار الإحصائي من جداول كوهن (Cohen, 1977) وقد حسبت قيم d لكوهن كمؤثرات لحجم الأثر باختلاف الاختبار الإحصائي المستخدم لفحص الفرضية الصفرية فقد اعتبر d مساوية: لقيمة الفرق بين متوسطي عينتين مستقلتين مقسوماً على الانحراف المعياري كمقياس لحجم الأثر في حالة الاختبار (t) للفرق بين متوسطي عينتين مستقلتين. وقيمة ابنا (خارج قسمة مجموع مربعات الانحرافات بين المجموعات على مجموع مربعات الانحرافات الكلية) كمقياس لحجم الأثر في حالة الاختبار (F).

عرض النتائج ومناقشتها

عرض نتائج السؤال الأول

نص السؤال الأول في الدراسة الحالية على: " كيف يتوزع حجم الأثر المقترن بالفرضيات الصفرية التي تم اختبارها حسب المستويات التي اقترحها كوهن لحجم الأثر؟". وللإجابة عن هذا السؤال تم حساب حجم الأثر حسب العلاقة الخاصة التي تختلف باختلاف الاختبار الإحصائي المستخدم لفحص الفرضية والجداول رقم (١) يبين عدد الفرضيات الإحصائية الدالة وغير الدالة موزعة حسب نوع الاختبار وحجم الأثر.

الجدول رقم (١)

عدد الفرضيات الإحصائية الدالة وغير الدالة موزعة حسب نوع الاختبار وحجم الأثر

المجموع	حجم الأثر				الاختبار
	أقل من ٠,٢ دال - غير دال	٠,٢ - ٠,٤٩ دال - غير دال	٠,٥ - ٠,٧٩ دال - غير دال	٠,٨ - فما فوق دال - غير دال	
٢٢٦	٩٧ ١٠	٤٩ ١١	٢٠ ١٨	١٠ ١١	T
٢١٩	١٤٥ ٣٤	٦ ١٠	١ ٦	٧ ١٠	F

يتضح من الجدول رقم (1) أنَّ مستويات حجم الأثر أربعة بينما اقترح كوهن ثلاثة مستويات هي: الصغير (0,2 - 0,49) والمتوسط (0,5 - 0,79) والكبير (0,8 - فما فوق) وعند حساب حجم الأثر حسب المستويات المقترحة تبين للباحث أنَّ هناك حجم أثر أقل من (0,2) لذا أضاف الباحث مستوى رابعاً لمستويات كوهن الثلاثة.

والملفت للنظر أن عدد الفرضيات التي احتوت على حجم أثر أقل من الصغير بلغت (107) فرضيات من أصل (226) فرضية للاختبار (t) أي ما يعادل (47٪) من نسبة الفرضيات الكلي. كما بلغ عدد الفرضيات الإحصائية الذي احتوى على حجم أثر صغير (10) فرضية أي ما يعادل (27٪) من العدد الكلي للفرضيات الإحصائية للاختبار (t).

وبلغ عدد الفرضيات الإحصائية التي احتوت على حجم أثر متوسط (38) فرضية أي ما يعادل (17٪) أما بالنسبة للفرضيات التي احتوت حجم أثر كبير فقد بلغ عددها (21) فرضية أي ما يعادل (9٪).

أما بالنسبة للاختبار الإحصائي (F) فقد توزع عدد الفرضيات على المستويات الأربعة بشكل لافت للنظر حيث بلغ عدد الفرضيات التي اقترنت بحجم أثر أقل من صغير (179) من أصل (219) فرضية أي حوالي (82٪) من الفرضيات . و(16) فرضية اقترنت بحجم أثر صغير أي ما يعادل (7٪) و(7) فرضيات اقترنت بحجم أثر متوسط أي ما يعادل (3٪) و(17) فرضية اقترنت بحجم أثر كبير أي ما يعادل (8٪) . وتجدر الإشارة إلى أنَّ السبب الذي جعل من حجم الأثر أقل من الحد الأدنى لحجم الأثر الصغير ذا الدلالة الإحصائية بمعنى أنَّ الفرق بين متوسطي عينتين مستقلتين يقل عن عشري الانحراف المعياري للمجتمع ذا دلالة إحصائية هو حجم العينة حيث أنَّ زيادة عدد أفراد العينة يساعد على رفض الفرضية الصفرية حتى لو كان حجم الأثر صغيراً.

وتتفق نتيجة الدراسة الحالية مع نتائج دراسة جردات وجودة (2005). ودراسة حجات وعليان (1997) كما تبين من نتائج الدراسة أنَّ متوسط حجم الأثر للاختبار (t) بلغ (0,38) بينما بلغ متوسط حجم الأثر للاختبار F (0,12) وهذا يعني أنَّ الدلالة العملية للاختبار (t) كانت متوسطة ولاختبار F كانت ضعيفة وما يشار إليه أنَّ جميع الباحثين اكتفوا بالدلالة الإحصائية لاستنتاج الفروق بين مستويات المتغيرات المستقلة . وتفسير نتائجهم واتخاذ قراراتهم ولم يشيروا إلى الدلالة العملية.

عرض نتائج السؤال الثاني

نص السؤال الثاني على: "كيف تتوزع قوة الاختبارات الإحصائية المستخدمة حسب المستويات المختلفة للقوة؟" وللإجابة عن هذا السؤال تم تقسيم فئات القوة إلى أربعة فئات والجداول رقم (٢) يبين ذلك.

الجدول رقم (٢)

عدد الفرضيات الإحصائية الدالة وغير الدالة موزعة حسب فئات قوة الاختبار ونوع الاختبار الإحصائي

فئات القوة	دال	غير دال	المجموع	المجموع الكلي
صفر - ٠,٥	١١ (٤٥) ×	١١٠ (١٤٩)	١٢١ (١٩٤)	٢١٥
٠,٥٠١ - ٠,٧٥	١ (٦)	١٦ (٢)	١٧ (٨)	٢٥
٠,٧٥١ - ٠,٨	١ (١)	١ (٠)	٢ (١)	٣
٠,٨٠١ - ١	٢٧ (٨)	٤٩ (٨)	٨٦ (١٦)	١٠٢
المجموع	٥٠ (٦٠)	١٧٦ (١٥٩)	٢٢٦ (٢١٩)	٤٤٥

* يشير الرقم الموجود بين قوسين إلى البيانات الخاصة بالاختبار (F)، والرقم خارج القوس إلى البيانات الخاصة بالاختبار (t)

يتضح من الجدول رقم (٢) أنّ حوالي (٧١٪) من الفرضيات التي اختبرت كانت قوة الاختبار الإحصائي فيها تساوي ٠,٥ أو أقل أي أن احتمال الرفض الصادق للفرضية الصفرية الخاصة باستخدام هذه الاختبارات لا يتعدى النصف وكان الباحث يعتمد قراره لقبول أو رفض الفرضية الصفرية إلى احتمال ظهور الصورة أو الكتابة عند رمي قطعة نقود غير متحيزة بدلا من الاعتماد على الاختبار الإحصائي ما يثير الشك وعدم الثقة في النتائج.

وأشارت النتائج إلى أن ما نسبته (١٠٪) فقط من الفرضيات المفحوصة دالة احصائيا وقوة الاختبار لها عالية ما يعني سلامة وقوة القرار المتخذ وهذه نسبة متواضعة جدا .

كما بينت النتائج أن ما نسبته (٦٪) من الفرضيات كانت قوة الاختبار الإحصائي فيها متوسطة وحوالي (٢٣٪) من الفرضيات كانت قوة الاختبار الإحصائي فيها عالية.

وقد شكلت الاختبارات الإحصائية غير الدالة إحصائياً والضعيفة في قوتها (٥٨٪) من الاختبارات في عينة الدراسة وقد يعود ذلك إلى عدم وجود فروق أو أنّ الفروق موجودة ولكن ضعف قوة الاختبار حال دون الكشف عنها.

أما الاختبارات الإحصائية الدالة إحصائياً والضعيفة في قوتها فقد شكلت حوالي (١٣٪) وهذا يعني أنّ القيمة العلمية للقرارات المبينة عليها مشكوك فيها.

وأشارت النتائج أيضاً أن ما نسبته (١٣٪) من الفرضيات المفحوصة غير دالة احصائيا وقوة الاختبار لها عالية ما يعني أنّ هناك فروق أو أثر للمتغير المستقل ولكن الإحصائي المستخدم

فمثل في الكشف عن هذه الفروق وقد يكون سبب ذلك ضعف التصميم الإحصائي المستخدم في فحص الفرضيات الصفرية أو انتهاك الباحث لبعض الافتراضات المتعلقة باستخدام الإحصائي المناسب حيث يؤدي انتهاك الافتراضات إلى تضخيم الخطأ من النوع الأول Type I error مما يدفع الباحث إلى رفض الفرضية الصفرية وهي صحيحة بالمقابل يمكن أن يقبل الباحث الفرضية الصفرية وهي خطأ نتيجة لتضخيم الخطأ من النوع الثاني Type II error كما أن استخدام الباحث أسلوباً إحصائياً لا يتفق نظرياً مع الافتراضات المتوفرة في تلك البيانات يؤدي إلى انخفاض مستوى الخطأ المعياري. وبالتالي يحصل الباحث على دلالة إحصائية غير حقيقية. (Bormond & Robert, 2005).

عرض نتائج السؤال الثالث

نص السؤال الثالث في الدراسة الحالية على: "كيف تتوزع قوة الاختبارات الإحصائية المستخدمة حسب المستويات المختلفة لحجم الأثر؟" وللإجابة عن هذا السؤال تم تقسيم فئات القوة إلى أربعة فئات وحجم الأثر إلى ثلاثة فئات مع ملاحظة عدم تضمين الجدول عدد الفرضيات التي لها حجم أثر أقل من الحد الأدنى الذي حدده كوهين والجدول رقم (3) يبين ذلك.

الجدول رقم (3)

عدد الفرضيات الدالة وغير الدالة موزعة حسب فئة القوة وحجم الأثر لكل من الإحصائي (t) و(F)

المجموع	كبير		متوسط		صغير		حجم الأثر فئات القوة
	غير دال	دال	غير دال	دال	غير دال	دال	
٢٢ (١٨)	٠ (٠)	٢ (٢)	٣ (٠)	٠ (٥)	١٧ (٤)	٠ (٧)	٠,٥-٠
١٣ (٥)	٠ (٠)	٠ (١)	٠ (٠)	٠ (٠)	١٢ (٢)	١ (٢)	٠,٧٥-٠,٥٠
١ (١)	٠ (٠)	٠ (١)	٠ (٠)	٠ (٠)	٠ (٠)	١ (٠)	٠,٨-٠,٧٥
٨٣ (١٦)	١٠ (٧)	١٠ (٦)	١٨ (١)	١٧ (١)	١٩ (٠)	٩ (١)	١-٠,٨٠
١١٩ (٤٠)	١٠ (٧)	١٢ (١٠)	٢١ (١)	١٧ (٦)	٤٨ (٦)	١١ (١٠)	المجموع

الرقم بين الأقواس يدل على قيمة الاختبار (F)

يتضح من الجدول رقم (3) أن ما نسبته (٤٩,٦٪) من الفرضيات التي اختبرت بواسطة الإحصائي (t) قد اقترنت بحجم أثر صغير. وأن (٥١٪) فقط من الفرضيات حققت معيار كوهين. أما بالنسبة للفرضيات التي اقترنت بحجم أثر متوسط أو كبير فنجد ما نسبته

(٩٧٪) و(٩٥٪) من الفرضيات التي اختبرت حققت معيار كوهن مما يدل على أهمية حجم الأثر في تحديد قوة الاختبار الاحصائي.

أما فيما يتعلق بالاحصائي (F) فيتبين من خلال الجدول أن ما نسبته (٤٠٪) من الفرضيات التي اختبرت قد افترنت بحجم أثر صغير، وأن (٦٪) منها حققت معيار كوهن، أما بالنسبة للفرضيات التي افترنت بحجم أثر متوسط أو كبير فنجد أن ما نسبته (٦٣٪) من الفرضيات التي اختبرت حققت معيار كوهن مما يدل على أهمية حجم الأثر في تحديد قوة الاختبار الاحصائي.

ولدراسة العلاقة بين قوة الاختبار وحجم الأثر تم استخدام اختبار مربع كاي للاستقلال لكل اختبار على حدة، وبلغت قيم مربع كاي المحسوبة (٣١,٧)، (٢٢,٣٩) لكل من اختبار (t) و(F) على التوالي، وهي قيم دالة احصائيا عند مستوى دلالة (٠,٠٥).

وهذا يعنى وجود علاقة قوية بين حجم الأثر، وقوة الاختبار الاحصائي حيث بلغت قيم معامل الاقتران (٠,٤٦) و(٠,٥٩) لكل من اختباري (t) و(F) على التوالي لأن القيمة العظمى لمعامل الارتباط في جدول التوافق في هذه الحالة (٠,٨٢).

خلاصة القول أنه يجب الانتباه إلى حجم العينة المناسب، والتصميم الاحصائي للدراسة لضمان قوة اختبار لا تقل عن (٠,٨)، كما أن كتابة النتائج بأشكال أخرى كإضافة فترات الثقة التي تسمح للقارئ بتحديد قوة الاختبار من خلال تحديد حجم فترة الثقة (الأصغر تعبر عن الفترة، والأكبر تعبر عن القوة)، وحجم الأثر وتحليل بواسون وكتابة القيمة الاحتمالية لجميع الاختبارات وليس فقط للاختبارات الدالة إحصائيا جميعها تدعم وتعزز نتائج البحث وتزود القارئ بفهم أعمق للنتائج.

وبما لا شك فيه أن تحليل قوة الاختبار الاحصائي لا يمكن أن ينفي الدور الأساسي للدلالة الإحصائية ولكن دوره يعتبر مكملاً لدور الدلالة الإحصائية، فاختبارات الدلالة الإحصائية لا تزود القارئ بالمعلومات الكافية لرفض الفرضية الصفرية، فضلاً عن تأثرها بحجم العينة فجميع الفرضيات الصفرية سوف يتم رفضها عند مستوى معين من حجم العينة، وعليه فإن الأثر العملي لحجم الأثر سيكون قليلاً.

التوصيات

١. الباحثون بالإشارة إلى مقدار حجم الأثر وقوة الاختبار وذلك لفهم نتائج دراساتهم بصورة أفضل والاعتماد على هذه النتائج لبناء أبحاث وفق أسس نظرية سليمة.

٢. نشر الثقافة الإحصائية في أوساط الباحثين من خلال إعطاء دورات إلزامية للأساتذة في مجال التحليلات الإحصائية.
٣. عرض البحوث المقبولة للنشر بالمجلة على متخصص في الإحصاء أو القياس والتقويم ولا يكتفي بأراء المحكمين خصوصا إذا لم يكونوا متخصصين في الإحصاء أو القياس والتقويم.
٤. إضافة أحد الأساتذة المتخصصين في الإحصاء أو القياس والتقويم لهيئة تحرير المجلة.

المراجع

- حجمات، تحسين وعلبان، خليل (١٩٩٧). واقع الدلالات الإحصائية والعملية وقوة الاختبارات الإحصائية المستخدمة في رسائل ماجستير الإرشاد النفسي والتربوي في الجامعة الأردنية. مجلة دراسات العلوم التربوية. ٢٤(٢). ٣٩٨-٤٠٨.
- جرادات، ضرار، وجودة، ماجد (٢٠٠٥). قوة الاختبار الاحصائي وحجم الأثر وحجم العينة للدراسات المنشورة في مجلة أبحاث اليرموك. المجلة الأردنية في العلوم التربوية. (١) ٢١-٢٩.
- الصائغ، ابتسام (١٩٩٦). الدلالة الإحصائية والدلالة العملية لاختبار "ت" و"ف". دراسة تحليلية تقويمية من خلال رسائل الماجستير التي قدمت في كلية التربية بجامعة أم القرى بمكة المكرمة حتى عام ١٤١٥هـ. رسالة ماجستير غير منشورة، كلية التربية، جامعة أم القرى.
- الصيد، عبد المعطى أحمد (١٩٨٨). الدلالة العملية وحجم العينة المصاحبان للدلالة الإحصائية لاختبار(ت) في البحث التربوي والنفسي العربي، ورقة عمل مقدمة في مؤتمر البحث التربوي: الواقع والمستقبل، المجلد الثاني، رابطة التربية الحديثة، المركز القومي للبحوث التربوية، القاهرة.
- عودة، أحمد سليمان، وملكاوي، فتحي حسن (١٩٩٢). أساسيات البحث العلمي في التربية والعلوم الإنسانية، أريد- الأردن: مكتبة الكتاني.
- نصار، يحيى (٢٠٠٥). استخدام حجم الأثر لفحص الدلالة العملية للنتائج في الدراسات الكمية. مجلة العلوم التربوية والنفسية. ٢٧(٢). ٣٦-٥٩.

Arnemmn, k. (2003). Areview of the panoply of the effect size choices. *ERIC, Document Reproduction (Service No.ED473811)*.

Bormond, G. & Robert E. (2005). Success for all: First year results from the National randomized field trail, *Education Evaluation and policy Analysis, 27(1), 1-22*.

- Carver, P.(1993). The case against statistical significance testing, revisited. *Journal of Experimental Education*, 61(4), 287-92, Sum 1993.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*(rev. ed). New York: Academic Press
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- Fagley, N. & Mckinney, I. J. (1983). Reviewer bias for statistical significance results: reexamination. *Journal of Counseling Psychology*, 30(2), 298-300.
- Glass , G.& Hopkins, K.(1984). *Statistical Method in Education and Research*, 2nd ed. Prentice - Hall, INC. Englewood Cliffs. New Jersey.
- Gravetter. J. (1988). *Statistics for the behavioral sciences*, 2nd ed. McGraw-Hill Book Company: New York.
- Hays, W. L.(1973). *Statistical for the social sciences*, 2nd edition. Holt: Rinehart and Winston Inc, New York.
- Hopkins, G. (2011). *A New View of Statistics. (Online Book)*. Retrieved on September 23 2011, from: <http://sportsci.org/resource/stats/newview.html>
- Huston, L.(1993). *Meaningfulness, statistical general significance, effect size, and power analysis: a general discussion with implications for manova*. Paper presented at annual meeting of the mid – south educational research association (2nd , New Orleans, LA, November 9-12, 1993).
- Keaster, D. (1988). *Statistical significance testing: from routine to ritual*. Paper presented at annual meeting of the mid – south educational research association (Louisville, KY, November 9-11, 1988).
- Kellow, J.(1998). Beyond statistical significant tests: the importance of using other estimates of treatment effects to interpret evaluation results. *American Journal of Evaluation*, 19(1), 123-34, win 1998.
- Kirk, R. (1996). Practical significance: A concept whose time has com. *Educational and psychological measurement*, 56(4).746-759.
- Mahadevan, L.(2000).*The effect size statistic :Overview of various choices*.(Eric Document Reproduction Service No.ED438308).
- Mcclain, A.(1995).*Effect size as an alternative to statistical significance testing*. Paper presented at the annual meeting for the American Educational Research Assosiation. San Francisco CA(Eric Document Reproduction Service No.ED382639).

- Mclean , J. & Ernest, J. (1997). *Has testing for statistical significance outlived its usefulness?*. Paper presented at annual meeting of the mid – south educational research association, (26th, Memphis, TN, November 12-14, 1997)
- Nix, T. & Barnette, J. (1998). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. *Research in the Schools*, 5(2), 3-14.
- Palomares , R. (1990). *Alternatives to statistical significance testing*. paper presented at annual meeting of the Mid-South educational research Association (19th New Orleans, LA, November14-16 1990)
- Snyder, P. & Lawson, S.(1993).Evaluating results using correct and uncorrected effect size estimates. *Journal of experimental Education*, 61, 334-349.
- Tapia, M.& Marsh, G.(2002). *Interpreting and reporting effect sizes in research investigations*. (ERIC Document Reproduction Service No.ED471293).
- Thompson, B. (1988). *Encouraging effect size reporting is not working: The etiology of research resistance to changing practices*. Paper presented at the annual meeting of the southwest Educational Research Association (Houston, Tx, January1988). Eric Document Reproduction service No. ED 416214.
- Weinberg, S. & Goldberg, K. (1979). *Basic Statistics for education and behavioral science*. Boston: Hughton Mifflin.
- Winer, B.(1971). *Statistical Principles in experimental design*, 2nd ed. N.Y. : Mc Graw Hill.
- Winer, B., Brown, D.& Michels, K.(1991). *Statistical Principles in experimental design*, 3rd ed . N.Y. :Mc Graw Hill.