قوة الاختبار الإحصائي وحجم الأثر وحجم العينة للدراسات المنشورة في مجلة أبحاث اليرموك- سلسلة العلوم الإنسانية والاجتماعية

ضرار جرادات* و ماجد جودة**

تاريخ تسلم البحث 2004/8/8 تاريخ قبوله 2004/11/29

ملخص: حاولت هذه الدراسة أن تقيّم قوة الاختبارالاحصائي وحجم الأثر وحجم العينة في الدراسات المنشورة في 16 مجلداً من مجلدات مجلة أبحاث اليرموك- سلسلة العلوم الإنسانية والاجتماعية. واحتوت عينة الدراسة على 785 اختباراً إحصائياً من نوع F المستخدم في تحليل التباين و t للعينات المستقلة و t لمعامل الارتباط. وأمكن من البيانات التي احتوتها الدراسات المنشورة حساب قيمة d لكوهن كمقياس لحجم الأثر، ومن ثم استخراج قوة الاختبار الإحصائي من جداول كوهن. وقد خلصت هذه الدراسة إلى الاستنتاج أن العامل الرئيس وراء النسبة العالية (72.5%) من الفرضيات الصفرية المرفوضة والنسبة العالية (52%) من الاختبارات الإحصائية التي حققت الحد الأدنى للقوة (0.8) الذي يقترحه كوهن، قد يكون الحجم الكبير لعيّنات الدراسة وليس حجم الأثر المقترن بالفرضيات الصفرية التي جرى اختبارها، إذ لم تزد نسبة الفرضيات الصفرية التي اختبرت واقترنت بحجم أثر كبير عن 26.5%.

(الكلمات المفتاحية: قوة الاختبار، حجم الأثر، حجم العينة، الدلالة الاحصائية، الدلالة العملية)

Test Power, Effect Size and Sample Size of Published Research in *Abhath Al-Yarmouk Hummanities and Social Science Series*

Derar Jaradat, *Faculty of Education. Yarmouk University, Irbid, Jordan.*
Majed Judeh, *Yarmouk Model School, Yarmouk University.*

Abstract: *This study tried to evaluate the test power, effect size and sample size of the research published in 16 volumes of Abhath Al-Yarmouk- Hummanities and Social Sciences Series. The sample consisted of 785 statistical test, where 3 conventional tests, were used, namely the F test for ANOVA, the t test for means of independent samples and the t test for the correlation coefficient. From data reported by the studies, Cohen d as a measure of effect size was calculated for each tested hypothesis and the power for the used statistical test was obtained from tables prepared by Cohen.The study concluded that the major factor behind the high percentage (72.5%) of rejected hypotheses and the high percentage (52%) of statistical tests reaching Cohen criterion for minimal power (0.8) might have been the large sample size and not the effect size since hypotheses associated with a large effect size were only 26.5%.*

***(Keywords**: Test power, effect size, sample size, statistical significance, practical significance).*

مقدمة: يبني الباحثون في العلوم الإنسانية والاجتماعية عادة نوعين من الفرضيات تعرف الأولى بالفرضية الصفرية (Null Hypothesis) وتسمى الثانية بالفرضية البديلة (Alternative Hypothesis) . ويمكن وصف الفرضية الصفرية بأنها ادعاء أو زعم للباحث حول مقدار مؤشر أو مؤشرات لمجتمع أو مجتمعات، أو حول مقدار تأثير أو تأثيرات لمتغير أو متغيرات مستقلّة على متغير أو متغيرات تابعة. وهذا الادعاء الذي يبنيه الباحث يحاول التحقق من صحته بإخضاع البيانات التي يجمعها من عينة أو عينات مأخوذة من مجتمع أو مجتمعات مستهدفة بمجهوده البحثي، للاختبار الإحصائي المناسب.

والاختبار الإحصائي (Statistical Test) ليس إلاّ مجموعة من المعالجات الإحصائية لبيانات عينة أو عينات تهدف إلى مساعدة الباحث في الوصول إلى قرار حول فرضيته أو فرضياته الصفرية وذلك برفضها أو الفشل في رفضها. ويصل الباحث إلى قراره هذا بعد مقارنة المؤشر الإحصائي المحسوب من بيانات العينة مع قيمة ذلك الإحصائي الحرجة، عندما تكون خصائص توزيع مؤشرات العينات العشوائية المشابهة لعينة الباحث معروفة.

ولا يعني بالضرورة أن اتّباع الباحث لإجراءات إحصائية محددة ومعروفة كفيل بصحة قراره حول الفرضية الصفرية التي لا يعرف شيئاً عن صحتها أو عدم صحتها، الأمر الذي يجعل الفرصة مواتية لوقوع الباحث في نوعين من الأخطاء يُعرف أحدهما بالخطأ من النوع الأول، عندما يرفض الباحث فرضية صفرية صحيحة، ويعرف الآخر بالخطأ من النوع الثاني عندما يفشل الباحث في رفض فرضية صفرية خاطئة.

ويساوي أعلى احتمال للخطأ من النوع الأول ألفا (α)، وهي ما تعرف بمستوى الدلالة الإحصائية. وتعني كلمة دلالة في أبسط صورها أن الفرق بين القيمة المفترضة لمعلم في المجتمع وقيمته المحسوبة من بيانات العينة كبير إلى القدر الذي يصعب عزوه إلى أخطاء العينات. ويحدد الباحث عادة قيمة ألفا قبل أن ينخرط في جمع بيانات دراسته، ويجعل من هذه القيمة معياراً يحتكم إليه في اتخاذ القرار حول فرضيته الصفرية. أما احتمال الخطأ من النوع الثاني والذي يُعرف بـ بيتا (β) فهو وإن كان قابلاً للحساب قبل أن يجمع الباحث بياناته، إلاّ أن العادة جرت أن يتم احتساب مقداره بعد أن ينتهي الباحث من تحليلاته الإحصائية.

* كلية التربية، جامعة اليرموك.
** المدرسة النموذجية لجامعة اليرموك.
© حقوق الطبع محفوظة لجامعة اليرموك، اربد، الأردن.

نظـرة علـى الأدب الـسابق: مـع أن اسـتخدام اختبـارات الدلالـة الإحـصائية مـن قبـل البـاحثين الكميّـين فـي العلـوم الإنـسانية والاجتماعية قد يعود إلى سبع عقود ونيّف، إلاّ أنّ هذا الاستخدام كان ولا يزال موضع جدل ونقاش (Kirk, 1996; Shaver, 1993). فقد وجهت على مدى العقود الماضية انتقادات كثيرة لاختبارات الدلالة الإحصائية. ففي حين حذّر بعض الباحثين من التفسيرات الخاطئة لمعنى الدلالة الإحصائية (Kirk, 1996) شكك آخرون في قيمتهـا وذهبـوا إلـى الاسـتنتاج بأنهـا إفـساد لطريقـة العلـم (Carver, 1978, 1993)، وعليه فلا بد من التخلص منها، واقترحوا بدلاً من ذلك أن يقوم الباحثون بإعطاء معلومات عن حجم الأثـر أو التـأثير أو مـا يعـرف بالدلالـة العمليـة (Practical Significance) لأثـر المتغيـر المـستقل علـى المتغيـر التـابع فـي مجهوداتهم البحثية.

ولعلّ أكثر ما يُعاب على اختبارات الدلالة الإحصائية اعتمادها القوي على عدد أفراد عينة الدراسة، إذ كلما زاد ذلك العدد كلما زاد احتمال رفض الفرضية الصفرية (Xiato, 2001)، إلاّ أن حجم الأثر يتـأثر هـو الآخر بمـا يمكن أن يُسمى حجم الدراسـة، الـذي يتضمن عدد أفرادها وعدد معالجاتها (مجموعاتها) معاً. وقد دلّل موري ودوسر (Murry & Doser, 1987) على العلاقة الخطية المباشرة بين قيمة مربع ايتا ($\eta 2$) (كمقياس لحجم الأثر) وقيمة الإحصائي F (كمقياس للدلالة الإحصائية)، واستنتجا أن كلاً من المؤشرين دالّة للآخر، فإذا رفضت الفرضية الصفرية بناء على قيمة F المحسوبة ستُرفض تلك الفرضية بناء على قيمة مربع ايتا ($\eta 2$) المحسوبة والعكس صحيح أيضاً.

وإذا مـا تـم تجـاوز الجـدل حـول اختبـارات الدلالـة الإحـصائية والمواقف المتباينة منها يجد البعض (Brewer, 1972; Wampold Furlong & Atkinson 1983) مـا يـشبه الإجمـاع بـين الباحثين علـى أهميـة المعلومـات وجوهريتهـا التاتجـة عـن قـوة الاختبـارات الإحـصائية التـي يـستخدمونها، هـذه المعلومـات التـي يطلـب كـوهن (Cohen, 1977) مـن البـاحثين تحديـدها قبـل إتمـام دراسـاتهم وإعطاء معلومات عنها عند نشر تلك الدراسات لاحقاً. وتأتي قيمة المعلومات عن قوة الاختبار الإحصائي من تدليلها على دقة قرار الباحث وصدقه برفض الفرضية الصفرية الخاطئة من جهة، وتدليلها على احتمال الخطأ من النوع الثاني من جهة أخرى. وهذه العلاقة التكاملية بين احتمال رفض الفرضية الصفرية الخاطئة واحتمال الفشل في رفضها تشير بوضوح إلى أهمية أن تكون قوة الاختبار الإحـصائي مرتفعـة، ويقتـرح كـوهن (Cohen, 1977) كقاعـدة اجتهاديـة بـأن لا تقـل قـوة الاختبـار الإحـصائي المـستخدم لاختبـار الفرضـية الـصفرية عـن 0.8 وللتـدليل علـى أهميـة أن تكـون قـوة الاختبـار الإحـصائي مرتفعـة، افتـرض وعلـى سـبيل المثـال أن قـوة اختبـار إحـصائي مـا تـساوي 0.5 وهـذا يعنـي أن احتمـال رفـض الفرضـية الـصفرية عنـدما تكـون خاطئـة هـو 0.5 أي أن احتمـال أن يكون قرار الباحث صائباً هو 0.5، وهو الاحتمال نفسه الذي يتوفر للباحث لو احتكم إلى نتيجة رمي قطعة نقود غير متحيزة بدلاً من إجراء الاختبار الإحصائي!.

وتتأثر قوة الاختبار الإحصائي بعوامل عديدة لعل من أهمها عدد أفراد عينة الدراسة، ومستوى الدلالة الإحصائية الذي يعتمده الباحث لاختبار فرضياته الصفرية، وحجم الأثر أو التأثير، الذي سمي اصطلاحاً بالدلالة العملية، إذ تزيد قوة الاختبار الإحصائي بزيادة مقادير هذه العوامل وتنقص بنقصانها. وإذا كان المقصود بعدد أفراد العينة ومستوى الدلالة الإحصائية غنياً عن التعريف إلاّ أن ما يُقصد بالدلالة العملية قد يختلف باختلاف أسئلة الباحثين الجوهرية، ولعلّ هذا يُفسّر تعدد المؤشرات الإحصائية التي تدلل على مقدار تلك الدلالة، فقد أورد كيرك(Kirk, 1996) 40 مقياساً لها، وصنفها في فئات ثلاث هي: مقاييس قوة العلاقة، ومقاييس حجم الأثر، ومقاييس أخرى. ومع أن هذه المقاييس تتفاوت في خصائصها الإحصائية كتقديرات لحجم الأثر، إلاّ أن هذا العدد الكبير منها قد يكون شاهداً على أن الباحثين قد لا يجدون في نتائج اختبارات الدلالة الإحصائية وحدها معلومات كافية للخروج بإجابات دقيقة وصحيحة لأسئلتهم البحثية الجوهرية، أو لمعرفة ما إذا كانت بياناتهم الامبريقية تؤيد أو تدحض فرضياتهم العلمية وليس فرضياتهم الإحصائية.

وقد اقترح كوهن (Cohen, 1977) تصنيفاً لمستويات حجم الأثر عندما يتم حسابه باستخدام مؤشر d لكوهن في ثلاث فئات: صغير ومتوسط وكبير، ويحتوي الجدول 1 مقادير حجم الأثر الصغيرة والمتوسطة والكبيرة في حال استخدام ثلاثة اختبارات إحصائية هي: اختبار F واختبار t للفرق بين متوسطي عينتين مستقلتين واختبار t (من الآن ولاحقاً t_r)؛ لاختبار فرضية حول ما إذا كان معامل بيرسون للارتباط بين متغيرين في المجتمع يختلف عن الصفر. وتجدر الملاحظة أن الحدود الدنيا لمستويات حجم

جـدول (1): تـصنيف كـوهن لحجـم الأثـر حـسب نـوع الاختبـار الإحصائي المستخدم

حجم الأثر	d في حالة استخدام اختبار t	d في حالة استخدام اختبار F	d في حالة استخدام اختبار t_r*
صغير	0.49 – 0.2	0.24 - 0.1	0.29 - 0.1
متوسط	0.79 - 0.5	0.39 – 0.25	0.49 - 0.3
كبير	0.8 - فما فوق	0.40 - فما فوق	0.5 - فما فوق

* يساوي حجم الأثر قيمة معامل الارتباط r .

الأثر تبعد مسافات متساوية عن بعضها بعضاً، في حين أن الحدود العليا لهذه المستويات مفتوحة.

وقد حظيت موضوعات قوة الاختبار الإحصائي والدلالة العملية والدلالة الإحصائية باهتمام كبير من الباحثين، فقد وجد كيرك (Kirk, 1996) في مراجعة للأبحاث المنشورة في أربع دوريات علمية أن نسبة الأبحاث التي أوردت مؤشرات للدلالة العملية قد تراوحت بين 12% و 77% للدوريات الأربع. كما وجد هاس وواكتر وسولومون (Hasse, Waccter & Solomon, 1982) أن

متوسط الدلالة العملية للأبحاث المنشورة في مجلة علم النفس الإرشادي يساوي 0.58، وحسب بروّر (Brewer, 1972) قوة الاختبارات الإحصائية للأبحاث المنشورة في مجلة البحث التربوي وجد أن متوسط القوة هو 0.23 و 0.56 و 0.79 لحجم الأثر الصغير والمتوسط والكبير على التوالي. وبالمثل وجد تشيس وتشيس (Chase & Chase, 1976) أن قوة الاختبارات الإحصائية للبحوث المنشورة في مجلة علم النفس التطبيقي كانت عموماً عالية خاصة عندما كانت الدلالة العملية متوسطة وعالية.

واستنتج الصياد (1988) من مراجعة للأبحاث المنشورة في مجلة البحث التربوي أن الباحثين يلجأون إلى زيادة عدد أفراد العينة من أجل رفض فرضياتهم الصفرية، في حين أن الدلالة العملية للأبحاث المنشورة في تلك المجلة كانت في المتوسط ضعيفة.

وفي الأردن وجد ابراهيم (2000) أن الأبحاث المنشورة في مجلة دراسات للفترة 1997- 1998 ذات دلالة عملية متوسطة وضعيفة، في حين كانت قوة الاختبارات الإحصائية المستخدمة على الأغلب عالية، كما وجد حجمات وعليان (1997) أن 63% من الفرضيات الدالة إحصائياً في رسائل ماجستير الإرشاد النفسي والتربوي في الجامعة الأردنية كانت ذات دلالة علمية ضعيفة، في حين كانت قوة اختبار تلك الفرضيات عالية.

وحديثاً خلص الدرابيع (2003) من مراجعة للدراسات المنشورة في مجلة مؤتة للبحوث والدراسات سلسلة العلوم الإنسانية والاجتماعية للفترة بين 1996- 2000 إلى الاستنتاج أن هذه الدراسات "تتمتع بقوة اختبار عالية ودلالة عملية مقبولة ... ويمكن أن تؤخذ نتائجها على المستوى العملي" (ص174)، بشرط أن تكون نتائج هذه الدراسات ذات دلالة إحصائية.

مشكلة الدراسة وأهميتها: حاولت هذه الدراسة تقييم قوة الاختبارات الإحصائية المستخدمة في فحص الفرضيات الصفرية لعينة من الأبحاث المنشورة في مجلة أبحاث اليرموك- سلسلة العلوم الإنسانية والاجتماعية للفترة من 1985- 2001، كما حاولت هذه الدراسة، من خلال معالجة البيانات التي تحتويها هذه الأبحاث المنشورة، إعطاء معلومات تفصيلية وشاملة تربط بين قوة الاختبارات الإحصائية المستخدمة وحجم الأثر، وأعداد أفراد العينات، ونوع الاختبار الإحصائي المستخدم. وبشكل أكثر تحديداً حاولت هذه الدراسة الإجابة عن الأسئلة التالية:

1- كيف يتوزع حجم الأثر المقترن بالفرضيات الصفرية التي تم اختبارها حسب المستويات التي اقترحها كوهن لحجم الأثر التي يوضحها جدول1 ؟
2- كيف تتوزع قوة الاختبارات الاحصائية المستخدمة حسب المستويات المختلفة للقوة ؟
3- ما نسبة الاختبارات الاحصائية التي تزيد قوتها عن الحد الأدنى الذي اقترحه كوهن لقوة الاختبار الاحصائي (0.8) ؟ وكيف تختلف تلك النسبة باختلاف حجم الأثر المقترن بالفرضيات الصفرية التي تم اختبارها ؟
4- هل يمكن وصف العلاقة بين قوة الاختبار الاحصائي من جهة والتغير المتصاحب لحجم الأثر وحجم العينة من جهة أخرى ؟ وهل حجم الأثر أم حجم العينة كان العامل الحاسم المؤثر في قوة الاختبارات الاحصائية للفرضيات الصفرية التي تم اختبارها ؟

ويمكن النظر إلى هذه الدراسة على أنها استكمال للمجهودات التي حاولت التعرف على المؤشرات المتعلقة بحجم الأثر، وقوة الاختبار الإحصائي للدراسات المنشورة في الدوريات الأردنية، التي بدأت بدراسة ابراهيم (2000) للدراسات المنشورة في مجلدين من مجلدات مجلة دراسات- سلسلة العلوم الإنسانية والاجتماعية التي تصدر عن الجامعة الأردنية، ثم دراسة الدرابيع (2003) للدراسات التي نُشرت في خمسة مجلدات من مجلة مؤتة للبحوث والدراسات- سلسلة العلوم الإنسانية والاجتماعية التي تصدر عن جامعة مؤتة.

ومن المؤمل أن تساعد هذه الدراسة على إثارة الاهتمام لدى الباحثين الأردنيين المهتمين بدراسة الظواهر الإنسانية والاجتماعية في المجتمع الأردني بمسائل تتعلق بقوة الاختبارات الإحصائية، التي يستخدمونها في اختبار فرضياتهم الصفرية وأحجام الأثر التي تقترن بتلك الفرضيات بدلاً من الركون إلى طقوس اختبارات الدلالة الإحصائية، التي يعتقد كارفر (Carver, 993, 1978) أنها مفسدة للطريقة العلمية، ويصفها كيرك (Kirk, 1996) بالتمرينات الساذجة التي تجيب على أسئلة تختلف عن أسئلة العلم.

الطريقة والإجراءات

عينة الدراسة: استخدمت طريقة العيّنة المنتظمة للحصول على عيّنة هذه الدراسة، التي تألفت من الدراسات التي استخدمت الاختبارات الإحصائية: F المستخدم عادة في تحليل التباين، و t المستخدم في اختبار الدلالة الإحصائية للفرق بين متوسطي عينتين مستقلتين، واختبار t_r المستخدم في اختبار اختلاف معامل ارتباط بين متغيرين في المجتمع عن الصفر. والدراسات التي تحتويها عينة هذه الدراسة منشورة في العدد الأول من كل مجلد من مجلدات مجلة أبحاث اليرموك سلسلة العلوم الإنسانية والاجتماعية للفترة 1985- 2001. وقد بلغ عدد هذه الدراسات 54 دراسة احتوت على 785 اختباراً إحصائياً وتوزّعت على أربعة ميادين معرفية هي: التربية وعلم النفس، والاقتصاد والعلوم الإدارية، وعلم الاجتماع، والعلوم السياسية.

الإجراءات: بعد مراجعة الدراسات المنشورة في العدد الأول من كل مجلد من مجلدات مجلة أبحاث اليرموك- سلسلة العلوم الإنسانية والاجتماعية للفترة بين 1985- 2001 استُبعدت الدراسات التي لم تستخدم اختبارات الدلالة الإحصائية كجزء من المجهود البحثي، وصُنّفت الدراسات المتبقية حسب الميدان المعرفي للبحث المنشور. ثم تم بعد ذلك تفريغ البيانات الضرورية في جداول خاصة أعدّت لذلك الغرض، واشتملت هذه البيانات على نوع الاختبار الإحصائي المستخدم ومستوى الدلالة الإحصائية الذي اعتمده الباحث، إن وجد، ونتيجة الاختبار الإحصائي، عدد أفراد

العينة إضافة لبعض البيانات الإحصائية الأخرى والتي كانت ضرورية لحساب حجم الأثر في الدراسة، ومن ثم استخراج قوة الاختبار الإحصائي من جداول كوهن (Cohen, 1977).

وقد حسبت قيم d لكوهن كمؤشرات لحجم الأثر حسب العلاقة الخاصة بها، التي تختلف باختلاف الاختبار الإحصائي المستخدم لفحص الفرضية الصفرية. فقد اعتبرت d مساوية: لقيمة معامل الارتباط بين المتغيرين كمقياس لحجم الأثر في حالة الاختبار الإحصائي t_r وقيمة الفرق بين متوسطي عينتين مستقلتين مقسوماً على الانحراف المعياري للمجتمع كمقياس لحجم الأثر في حالة الاختبار t للفرق بين متوسطي عينتين مستقلتين، وقيمة الجذر التربيعي لـ (ايتا2/ (1-ايتا2)) كمقياس لحجم الأثر في حالة الاختبار F، ولمزيد من التفصيل حول كيفية حساب قيم d لكوهن باختلاف الاختبار الإحصائي يمكن الرجوع إلى كتاب كوهن (Cohen, 1977).

وقد تبين من بيانات الدراسات التي تم مراجعتها في عينة الدراسة، أن أكثر من نصف هذه الدراسات لم تُحدّد مستوى للدلالة الإحصائية للاختبار الإحصائي قبل جمع البيانات أو بعد تحليل النتائج. ولأن قيمة مستوى الدلالة الإحصائية ضرورية لحساب قوة الاختبار الإحصائي، فقد اعتمد في هذه الدراسة قيمة الاحتمال P Value (التي تظهر في الجداول التي تحتوي عليها الدراسات المنشورة) لأن تكون القيمة المحسوبة للإحصائي بهذا المقدار أو أكبر بسبب أخطاء العينات العشوائية كبديل لمستوى الدلالة الإحصائية. وقد يكون من الضروري التذكير أن اتّباع مثل هذا الإجراء سيعمل أحياناً على تقليل قوة الاختبار الإحصائي المستخرجة من الجداول (عندما تكون قيمة p صغيرة)، وعلى زيادتها أحياناً أخرى (عندما تكون قيمة p كبيرة) . كما تجدر الإشارة إلى أن عدد أفراد العيّنة في المجموعة (المعالجة) عند حساب قوة الاختبار الإحصائي F والاختبار الإحصائي t قد حسب بقسمة العدد الكلي لأفراد العينة على عدد مستويات المتغير المستقل عند استخدام اختبار F لفحص الدلالة الإحصائية للأثر الرئيسي للمتغير المستقل، وعلى عدد التقاطعات عند استخدام اختبار F لفحص الدلالة الإحصائية للتفاعلات، وعلى اثنين عند استخدام اختبار t للفرق بين متوسطي عينتين مستقلتين، بمعنى أنه تم افتراض تساوي عدد الأفراد في المجموعات الفرعية لكل دراسة. وهذا هو الإجراء الذي يقترحه كوهن (Cohen, 1977) عند عدم تساوي عدد الأفراد في المجموعات الفرعية. وقد اعتبر العدد الكلي لأفراد العينة كعدد للأفراد في المجموعة عند إيجاد قوة الاختبار الإحصائي t_r.

النتائج ومناقشتها: تمت الإشارة سابقاً إلى الجدل القائم حول أهمية كل من الدلالة الإحصائية والدلالة العملية على أنهما معياران للحكم على قيمة الأدلة الامبريقية التي توفرها المجهودات البحثية حول الظواهر المدروسة. كما أشير إلى ما يشبه الإجماع على أن يحتوي التقرير البحثي المنشور على معلومات عن كلا الدلالتين حتى يستطيع قارئ التقرير أن يُقيّم بشكل مستقل عن الباحث نتائج دراسة ما والاستنتاجات التي بنيت على تلك النتائج.

وبالرجوع إلى بيانات الدراسات المنشورة في مجلة أبحاث اليرموك- سلسلة العلوم الإنسانية والاجتماعية، التي شكلّت عينة هذه الدراسة تبين أن من بين الفرضيات التي تم اختبارها باستخدام الاختبارات الإحصائية F و t و t_r ، التي بلغ عددها 785 فرضية الذي يظهر أن أغلب الفرضيات التي رُفضت كانت تتعلق بظواهر

جدول (2): عدد الفرضيات الإحصائية الدالة وغير الدالة موزعة حسب نوع الاختبار والميدان المعرفي

الاختبار	التربية وعلم النفس		الاقتصاد والعلوم الإدارية		علم الإجتماع		العلوم السياسية		المجموع			
	دال	غير دال	دال	غير دال	دال	غير دال	دال	غير دال	دال	غير دال	مجموع الفرضيات	النسبة المئوية
F	101	95	41	14	-	-	-	1	142	110	252	%32.10
T	94	30	10	1	57	14	5	21	166	66	232	%29.60
tr	145	28	85	23	-	-	16	4	246	55	301	%38.30
المجموع	340	153	136	38	57	14	21	25	554	231	785	100
النسبة المئوية	%43.30	%19.50	%17.30	%4.90	%7.20	%1.80	%2.70	%3.30	%70.70	%29.30	%100	

تربوية ونفسية، وإن حوالي 38.3% من هذه الفرضيات كانت حول علاقات ارتباطية بين متغيرين، في حين كان الباقي فرضيات حول تساوي متوسطين أو أكثر في المجتمع.

وللتعرف على مقادير حجم الأثر في الدراسات المنشورة في مجلة أبحاث اليرموك- سلسلة العلوم الإنسانية والاجتماعية، وكيف توزعت هذه المؤشرات حسب المستويات التي اقترحها كوهن (Cohen, 1977) التي احتواها جدول1 في هذه الدراسة، تم إضافة مستوى رابع لمستويات كوهن الثلاثة يحتوي على مقادير حجم الأثر التي تقل عن الحدّ الأدنى لحجم الأثر الصغير والدالة إحصائياً في الوقت نفسه. ويظهر من جدول 3 أن حوالي 8.5% من مقادير حجم الأثر كانت أقل من الحدّ الأدنى لحجم الأثر

جدول (3): عدد الفرضيات الإحصائية الدالة ومتوسط عدد أفراد العينة* (N) موزعين حسب نوع الاختبار وحجم الأثر

نوع الاختبار	F		t		tr		المجموع		النسبة التراكمية لعدد الفرضيات
حجم الأثر	عدد الفرضيات	متوسط N	عدد الفرضيات	متوسط N	عدد الفرضيات	متوسط N	عدد الفرضيات	متوسط N	
أقل من الحجم الصغير	12	310	26	266	9	280	47	280	%8.50
الصغير	19	191	59	313	79	224	167	250	%38.60
المتوسط	20	129	32	260	83	187	135	196	%63
الكبير	81	181	49	100	75	221	205	176	%100
المجموع	142	187	166	233	246	213	554	212	
النسبة المئوية	%25.60		%30		%44.40		%100		

* متوسط عدد أفراد العينة مقرب إلى أقرب واحد صحيح.

الصغير، وإذا ما أضفنا إلى ذلك أعداد الفرضيات التي تقترن بحجم الأثر الصغير نجد أن حوالي 38.6% من مقادير حجم الأثر تقع ضمن هذين المستويين.

وقد يثار التساؤل حول السبب الذي يجعل من حجم أثر أقل من الحد الأدنى لحجم الأثر الصغير ذا دلالة إحصائية، أو بعبارة أخرى كيف يمكن مثلاً أن يكون فرق بين متوسطي عينتين مستقلتين يقل عن عشري الانحراف المعياري للمجتمع ذا دلالة إحصائية عند استخدام اختبار t للفرق بين متوسطي عينتين مستقلتين. والحقيقة أن مثل هذا التساؤل قد يختفي إذا ما عرفنا بعض المعلومات عن عدد الأفراد الكبير في المجموعات الفرعية للدراسات المختلفة الذي يوضحه جدول 3، حيث تحتوي خلاياه على متوسط عدد الأفراد في المجموعة الواحدة عند استخدام اختبار F واختبار t، وفي المجموعة الكلية عند استخدام اختبار t_r، ومن الواضح أن متوسط عدد أفراد العينة كبير بشكل خاص للفرضيات التي اقترنت بحجم أثر أقل من الحدّ الأدنى لحجم الأثر الصغير أو بحجم أثر صغير بالمقارنة بالفرضيات الأخرى. ولعلّ في هذا الأمر دليلاً على أن زيادة عدد أفراد العينة يساعد إلى مدى بعيد على رفض الفرضية الصفرية حتى ولو كان حجم الأثر صغيراً.

وثمة أمر آخر يظهره جدول 3 وهو أن غالبية الفرضيات التي رفضها الباحثون الذين نشروا دراساتهم في مجلة أبحاث اليرموك- سلسلة العلوم الإنسانية والاجتماعية قد اقترنت بحجم أثر متوسط أو أقل. وإذا ما أضفنا إلى هذه النتيجة حقيقة أن حوالي 44% من هذه الفرضيات المرفوضة كانت حول معامل ارتباط بين متغيرين في مجتمع، وأن قيمة معامل الارتباط الذي يُصنف ضمن حجم الأثر المتوسط تتراوح بين 0.3- 0.5 لأدركنا ضعف العلاقة من جهة وقلّة المنفعة العملية لمثل هذه العلاقة الضعيفة إذا أريد استثمارها وتوظيفها لأغراض مستقبلية.

ولإعطاء معلومات شاملة عن الدراسات المنشورة في مجلة أبحاث اليرموك- سلسلة العلوم الإنسانية والاجتماعية للفترة 1985- 2001، تمّ بناء جدول 4، الذي يحتوي على بيانات تفصيلية عن مستوى حجم الأثر، ونوع الاختبار الإحصائي المستخدم ونتيجة اختبار الفرضية الصفرية، وعدد الفرضيات الصفرية التي تم اختبارها، ومتوسط كل من حجم الأثر، وقوة الاختبار الإحصائي، وعدد أفراد العينة، التي تقترن بكل مجموعة من الفرضيات ذات الخصائص المشتركة من حيث نوع الاختبار ومستوى حجم الأثر ونتيجة الاختبار الإحصائي للفرضية الصفرية. ويتضح من جدول 4 أن مقدار متوسط قوة الاختبار الإحصائي يتغير بتغير كل من متوسط عدد أفراد العينة ومتوسط حجم الأثر. ومع أن مثل هذه النتيجة متوقعة إلاّ أنه قد يكون من الصعب الوصف الدقيق لخصائص العلاقة التي تربط بين قوة الاختبار من جهة وعدد أفراد العينة وحجم الأثر معاً من جهة أخرى. فمن المعروف أن قوة الاختبار الإحصائي تزداد بزيادة كل من عدد أفراد العينة وحجم الأثر لكن هذه الزيادة قد تكون محكومة بعوامل ثلاثة: الأول: أن قوة الاختبار الإحصائي لا يمكن أن تزيد عن واحد صحيح، وبالتالي ليس من المتوقع أن تكون هذه العلاقة خطية سواء مع عدد أفراد العينة أم مع حجم الأثر، والثاني: أن مقدار التغير في عدد أفراد العينة وحجم الأثر قد يكون متفاوتاً من جهة وفي اتجاه واحد أو في اتجاهين متعاكسين من جهة أخرى. أما العامل الثالث؛ فيتعلق بخصائص العلاقة بين عدد أفراد العينة ومقدار حجم الأثر، التي يصعب وصفها أو التنبؤ بها. وقد أظهرت بيانات هذه الدراسة أن هذه العلاقة ضعيفة وموجبة (0.176) عند محاولة وصفها

جدول (4): عدد الفرضيات الإحصائية الدال وغير الدال ومتوسطات حجم الأثر والقوة وعدد أفراد العينة* موزعة حسب نوع الاختبار وحجم الأثر

نوع الاختبار	مستوى حجم الأثر	دال				غير دال				المجموع			
		عدد الفرضيات	متوسط حجم الأثر	متوسط القوة	متوسط عدد أفراد العينة	عدد الفرضيات	متوسط حجم الأثر	متوسط القوة	متوسط عدد أفراد العينة	عدد الفرضيات	متوسط حجم الأثر	متوسط القوة	متوسط عدد أفراد العينة
F	صغير	41	0.14	0.78	226	105	0.08	0.21	150	146	0.1	0.37	172
	وسط	20	0.29	0.89	129	3	0.29	0.41	22	23	0.29	0.83	115
	كبير	81	8.51	0.98	181	2	0.79	0.99	67	83	8.3	0.98	179
مجموع F		142	4.9	0.91	187	110	0.1	0.23	145	252	2.8	0.61	169
t	صغير	85	0.27	0.59	299	61	0.14	0.23	235	146	0.21	0.44	272
	وسط	32	0.6	0.89	260	3	0.61	0.6	215	35	0.6	0.87	256
	كبير	49	1.27	0.94	100	2	0.97	0.83	32	51	1.26	0.93	98
مجموع t		166	0.63	0.75	233	66	0.19	0.27	228	232	0.5	0.61	231
tr	صغير	88	0.2	0.63	230	55	0.08	0.26	222	143	0.15	0.49	227
	وسط	83	0.39	0.94	187	-	-	-	-	83	0.39	0.94	186
	كبير	75	0.66	0.99	221	-	-	-	-	75	0.66	0.99	221
مجموع tr		246	0.4	0.85	213	55	0.08	0.26	222	301	0.34	0.74	214
المجموع الكلي		554	1.63	0.83	212	231	0.12	0.25	187	785	1.19	0.66	205

* متوسط عدد أفراد العينة مقرب إلى أقرب واحد صحيح.

بمعامل بيرسون للارتباط. خلاصة القول إن محاولة فهم مصدر أو مصادر الفروق بين متوسطات قوة الاختبارات الإحصائية التي يحتوي عليها جدول 4 لا بد وأن تأخذ في الحسبان العوامل الثلاثة السابقة فرادى ومتصاحبة.

وبالرغم من الصعوبات المشار لها في الفقرة السابقة، فإنه من الممكن تقديم بعض الأوصاف العامة للبيانات في جدول 4، منها أن متوسط عدد أفراد العينة ومتوسط حجم الأثر لكل مجموعة فرضيات تمثّل تقاطع اختبار إحصائي معين مع أحد مستويات حجم الأثر هما أكبر للفرضيات التي رفضها الباحثون منها للفرضيات المناظرة، التي فشلوا في رفضها. ولعل هذا قد يعزز من جهة الاستنتاج بأن هذين العاملين (متوسط عدد أفراد العينة ومتوسط حجم الأثر) قد أثّرا بقوة على نتيجة الاختبار الإحصائي ويقوّي من جهة أخرى الاعتقاد بأن الباحثين قد يلجأون عادة إلى زيادة عدد أفراد العينة لرفض فرضياتهم الصفرية. وعلى سبيل المثال، وكما أشير سابقاً، يكون معامل بيرسون للارتباط بين 100 زوج من المشاهدات لمتغيرين دال إحصائياً عند مستوى الدلالة 0.05 إذا كان مقداره 0.195، ولا يخفى بطبيعة الحال ضعف مثل هذه العلاقة.

وللوقوف على حجم العينات التي استُخدمت بياناتها في المعالجات الإحصائية الوصفية والاستنتاجية للدراسات المنشورة في مجلة أبحاث اليرموك- سلسلة العلوم الإنسانية والاجتماعية بُني جدول 5 لهذا الغرض. ولم يتم الالتزام في جدول 5 بطول ثابت لفئة عدد أفراد العينة على أمل أن يساعد ذلك على فهم أفضل لخصائص توزيع أفراد العينات في الدراسات المنشورة. وقد يكون من الضروري التذكير مرة أخرى أن عدد أفراد العينة في حالتي الاختبار F، و t هو العدد الذي احتوته المجموعة الواحدة وليس عدد أفراد العيّنة الكلية للدراسة.

جدول (5): عدد الفرضيات الإحصائية الدالة وغير الدالة موزعة حسب فئة عدد أفراد العينة ونوع الاختبار

نوع الاختبار	F		t		tr		المجموع		النسبة التراكمية للمجموع الكلي
فئة العينة	دال	غير دال	دال	غير دال	دال	غير دال	كلي	غير دال	
صفر – 50	29	35	51	24	10	-	90	59	149	**%19**
51 – 100	57	30	41	6	65	14	163	50	213	%46.10
101 – 200	25	30	8	12	42	3	75	45	120	%61.40
201 – 300	2	-	2	1	79	26	83	27	110	%75.40
301 – 400	7	7	2	-	45	12	54	19	73	%84.70
أكثر من 400	24	8	62	23	5	-	89	31	120	%100
المجموع	142	110	166	66	246	55	554	231	785	

ومن تفحص جدول 5 يتبين أن أكثر من 46% من العينات لا يزيد عدد أفرادها عن 100، ومع أن 100 فرد في المجموعة الواحدة (لكل معالجة) قد يكون ملائماً عند استخدام اختبار F أو اختبار t، إلاّ أنه قد يكون قليلاً في حالة الاختبار t_r، إذ من المتوقع أن يؤدي العدد القليل لأفراد العينة في هذه الحالة إلى عدم استقرار مؤشرات العينة المحسوبة كتقديرات لمؤشرات المجتمع، بالرغم من رفض الفرضية الصفرية، وإلى عدم الدقة في تقدير مؤشرات المجتمع عند بناء فترات ثقة حول معالم المجتمع. وعلى سبيل المثال لا الحصر تكون قيمة معامل بيرسون للارتباط 0.195 دالة إحصائياً عند مستوى الدلالة 0.05 (بافتراض فرضية بديلة غير متجهة) عندما يكون عدد أفراد العينة 100، وهذا يعني أن فترة الثقة 95% التي يمكن أن يقع ضمنها معامل الارتباط في المجتمع هي 0.003 – 0.387، وهذه الفترة واسعة جداً ويمكن

وصف التقديرات المبنية عليها بعدم الدقة مقارنة بفترة لا يزيد طولها عن ربع طول هذه الفترة مثلاً.

ومن الملفت للنظر في جدول 5 أن ربع الفرضيات التي تم اختبارها إحصائياً في الدراسات المنشورة في مجلة أبحاث اليرموك زاد عدد أفراد المجموعة الواحدة فيها عن 300 فرد، بمعنى أن أقل عدد ممكن لأفراد العينة الكلية في حالتي الاختبار t، و F هو 600. كما يظهر من جدول 5 أن حوالي 83% من الفرضيات الدالة إحصائياً كان يزيد عدد أفراد المجموعة الواحدة منها عن 100 فرد. ولعل هذا يُفسّر جزئياً (إذا ما أخذ في الاعتبار حجم الأثر أيضاً) تأثير العدد الكبير لأفراد العينة على قوة الاختبار الإحصائي من جهة، وعلى نتيجة هذا الاختبار من جهة أخرى.

ولتقييم قوة الاختبارات الإحصائية التي استخدمت في الدراسات المنشورة في مجلة أبحاث اليرموك- سلسلة العلوم الإنسانية والاجتماعية أعدّ الجدول 6 لهذه الغاية. ولتسهيل عمل الأحكام حول قوة هذه الاختبارات، تفاوتت أطوال فئات القوة الأربع في جدول 6 وتم اختيار نهايات الفئات بشكل مقصود من أجل

جدول (6): عدد الفرضيات الدالة وغير الدالة موزعة حسب فئات قوة الاختبار ونوع الاختبار الإحصائي

نوع الاختبار	F		t		tr		المجموع			النسبة التراكمية
فئات القوة	دال	غير دال	دال	غير دال	دال	غير دال	دال	غير دال	الكلي	للمجموع الكلي
صفر – 0.500	4	102	37	61	26	48	67	211	278	%35.40
0.750 – 0.501	21	5	22	2	30	5	73	12	85	%46.20
0.800 – 0.751	2	-	3	-	11	-	16	-	16	%48.30
1 - 0.801	115	3	104	3	179	2	398	8	406	%100
المجموع	142	110	166	66	246	55	554	231	785	

حساب نسبة الاختبارات التي تقل قوتها عن مقادير هذه النهايات. ويتضح من جدول 6 أن أكثر من 35% من الاختبارات الإحصائية التي أجريت تساوي قوتها 0.5 أو أقل، بمعنى أن احتمال الرفض الصادق للفرضية الصفرية الخاصة باستخدام هذه الاختبارات قد لا يزيد عن النصف في أفضل الأحوال. ومثل هذه النتيجة قد تكون مثيرة للقلق حول قوة هذه الاختبارات، لأن الباحث يستطيع أن يحتكم في قراره حول الفرضية الصفرية إلى احتمال ظهور الصورة أو الكتابة عند رمي قطعة نقود غير متحيزة، ويرفض الفرضية أو يفشل في رفضها بدلاً من الاحتكام إلى الاختبار الإحصائي، وهو بذلك يعمل على توفير الجهد والتكلفة للمجهود البحثي المتمثل بتطوير أدوات جمع البيانات، وجمع هذه البيانات، وإجراء التحليلات الإحصائية لها. كما يتبين من جدول 6 أن حوالي 11% من الاختبارات تتراوح فرص الرفض الصادق للفرضيات الصفرية الخاطئة لها بين فرصة من اثنتين إلى ثلاث فرص من أربع.

وإذا ما حاولنا تبني المعيار الذي اقترحه كوهن (Cohen, 1977) بأن لا تقل قوة الاختبار الإحصائي عن 0.8 نجد من جدول 6 أن حوالي 52% من الاختبارات الإحصائية التي استخدمت في الدراسات المنشورة في مجلة أبحاث اليرموك قد حققت هذا المعيار.

وقد تكون البيانات التي احتواها جدول 6 وحدها غير كافية لتقييم قوة الاختبارات الإحصائية في الدراسات المنشورة في مجلة أبحاث اليرموك بشكل دقيق، إذ غاب عن جدول 6 المعلومات التي تربط بين قوة الاختبار ومستوى حجم الأثر التي يحتويها جدول 7. الذي يتضح من بياناته أن أكثر من 55% من الفرضيات التي اختبرت قد اقترنت بحجم أثر صغير، وأن 80% من الاختبارات الإحصائية لهذه الفرضيات لم يحقق المعيار الذي اقترحه كوهن. وإذا ما أضفنا الفرضيات المقترنة بحجم الأثر الصغير والمتوسط معاً نجد أنها تزيد عن 73% من مجموع

جدول (7): عدد الفرضيات الدالة وغير الدالة موزعة حسب فئة القوة وحجم الأثر

حجم الأثر	صغير		وسط		كبير		المجموع			النسبة التراكمية
فئة القوة	دال	غير دال	دال	غير دال	دال	غير دال	دال	غير دال	الكلي	للمجموع الكلي
صفر – 0.500	61	206	5	5	1	-	67	211	278	%35.40
0.750 – 0.501	59	11	8	-	6	1	73	12	85	%46.20
0.800 – 0.751	11	-	3	-	2	-	16	-	16	%48.30
1 - 0.801	83	4	119	1	196	3	398	8	406	%100
المجموع	214	221	135	6	205	4	554	231	785	

الفرضيات التي تم اختبارها، إلاّ أن 36% فقط من الاختبارات الإحصائية لهذه الفرضيات قد حقق معيار كوهن للحد الأدنى لقوة الاختبار الإحصائي، إلاّ أن الأمر يختلف في حالة حجم الأثر المتوسط أو الكبير، إذ إن 85% و 95% من الفرضيات التي اقترنت بحجم أثر متوسط أو كبير على التوالي قد حققت معيار كوهن. وربما يدل ذلك مرة أخرى على أهمية حجم الأثر في تحديد قوة الاختبار الإحصائي.

الاستنتاجات: بالرغم من أن ما يزيد عن 70% من الفرضيات التي تم اختبارها في الدراسات المنشورة في مجلة أبحاث اليرموك- سلسلة العلوم الإنسانية والاجتماعية قد تم رفضها، إلاّ أن هذه النسبة العالية للرفض لم تكن راجعة على ما يبدو لكبر حجم الأثر الذي يقترن بهذه الفرضيات، إذ شكلت الفرضيات التي اقترنت بحجم أثر صغير أو متوسط ما نسبته 73.5% من مجموع الفرضيات التي تم اختبارها.

وبالمثل قد لا تكون النسبة العالية (حوالي 52%) من الاختبارات الإحصائية التي حققت الحد الأدنى للقوة (0.8) كما يقترح كوهن، راجعة هي الأخرى إلى كبر حجم الأثر الذي يقترن بالفرضيات التي تم اختبارها، إذا لم تزد نسبة الفرضيات التي اقترنت بحجم أثر كبير عن 26.5% فقط من مجموع الفرضيات التي تم اختبارها، الأمر الذي يقود إلى الاستنتاج بأن العدد الكبير

لأفراد العينات التي بنى الباحثون اعتماداً على بياناتها قراراتهم في رفض الفرضيات الصفرية أو الفشل في رفضها، ربما كان العامل الحاسم وراء هاتين النسبتين العاليتين، ومثل هذا الاستنتاج قد يدعم ما خلصت إليه مجهودات سابقة (ابراهيم، 2000؛ حجمات وعليان، 1997؛ الدرابيع، 2003) راجعت الدراسات المنشورة في دوريات أردنية من أن غالبية الاختبارات الإحصائية التي أجرتها تلك الدراسات ذات قوة عالية إلا أن معظمها قد اقترن بحجم أثر متوسط أو صغير.

لقد حاولت هذه الدراسة أن تقدّم وصفاً موضوعياً ومعمقاً للمؤشرات الإحصائية المتعلقة بقوة الاختبار الإحصائي وحجم الأثر وعدد أفراد العينة للدراسات المنشورة في مجلة أبحاث اليرموك- سلسلة العلوم الإنسانية والاجتماعية للفترة بين 1985- 2001. وقد تكون الطريقة التي تناولت بها هذا الدراسة مثل تلك المؤشرات قد اختلفت كثيراً عن الطريقة التي تناولت بها مجهودات أخرى الدراسات المنشورة في دوريات أردنية أخرى. ولعل أهم ما يميز منهجية هذه الدراسة هو التركيز المتصاحب على مجمل العلاقات المتشابكة التي تربط بين قوة الاختبار الإحصائي وحجم الأثر، وعدد أفراد العينة ونتيجة الاختبار الإحصائي لكل فرضية في محاولة للوصول إلى فهم أفضل وأعمق للمعالجات الإحصائية التي أجراها مؤلفو هذه الدراسات من جهة، وإلى تقييم أكثر واقعية لنتائج هذه الدراسات وللاستنتاجات التي بنيت عليها من جهة أخرى.

وغني عن القول أن المعلومات عن مؤشرات قوة الاختبار الإحصائي وحجم الأثر وعن عدد أفراد العينة في الدراسات المنشورة في مجلة أبحاث اليرموك- سلسلة العلوم الإنسانية والاجتماعية لا تكفي وحدها للحكم على قيمة ومنفعة المجهودات البحثية التي قام بها ناشرو هذه الدراسات، إذ يحتاج التقويم الشامل والدقيق لهذه المجهودات البحثية إلى معلومات وافية عن أهمية المشكلات المدروسة ودلالاتها الإنسانية والاجتماعية، وعن مدى التزام تصميماتها البحثية بالمعايير المعروفة للتصاميم الجيدة، وعن مدى توفر الخصائص الجيدة لأداة القياس في أدوات جمع بياناتها، وعن مدى ملاءمة تحليلاتها الإحصائية للإجابة عن الأسئلة التي طرحتها هذه الدراسات، وأخيراً وليس آخراً عن مدى الانسجام المنطقي بين استنتاجات هذه الدراسات ونتائجها. ومن المؤمل أن توفر مراجعات مستقبلية للدراسات المنشورة في الدوريات الأردنية مثل هذه المعلومات للاسترشاد بها من الباحثين من جهة ومن رؤساء هيئات التحرير من جهة أخرى.

المصادر والمراجع

ابراهيم، رشاد ابراهيم (2000). *واقع الدلالة الإحصائية والدلالة العملية وقوة الاختبارات الإحصائية المستخدمة في بحوث مجلة دراسات- العلوم التربوية- في الأعوام (1997- 1998) في الجامعة الأردنية*. رسالة ماجستير غير منشورة، الجامعة الأردنية، عمان، الأردن.

حجمات، تحسين وعليان، خليل (1997). واقع الدلالات الإحصائية والعملية وقوة الاختبارات الإحصائية المستخدمة في رسائل ماجستير الإرشاد النفسي والتربوي في الجامعة الأردنية. *مجلة دراسات العلوم التربوية، 24*، 398- 408.

الدرابيع، ماهر (2003). واقع الدلالات الإحصائية والعملية وقوة الاختبارات الإحصائية المستخدمة في بحوث مجلة مؤتة للبحوث والدراسات- سلسلة العلوم الإنسانية والاجتماعية. *مؤتة للبحوث ولدراسات، 18*، 151- 176.

الصياد، عبد المعطي أحمد (1988). الدلالة العملية وحجم العينة المصاحبتين للدلالة الإحصائية لاختبار (ت)، في البحث التربوي والنفسي العربي. *مؤتمر بحوث التربوي الواقع والمستقبل، المجلد الثاني، رابطة التربية الحديثة، المركز العربي للبحوث التربوية، القاهرة.*

Brewer, J. K. (1972). On the power of statistical tests in the American Education Research Journal. *American Educational Research Journal, 9* (3), 391- 401.

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

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

Chase, L. J., Chase, R. B. (1976). A statistical power analysis of applied psychological research. *Journal of applied psychology, 61*, 234- 237.

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

Haase, R. F., Wacchter, D. M., & Solomon, G. S. (1982). How significant is a significant difference? Average effect size in counseling psychology. *Journal of counseling psychology, 29*, 58- 65.

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

Murray, L. W., &s Dosser, Jr., D. A. (1987). How significant is a significant difference? Problems with the measurement of magnitude of effect. *Journal of Counseling Psychology, 34* (1), 68- 72.

Penick, J. E., Brewer, J. K. (1972). The power of statistical tests in science teaching research. *Journal of Research in Science Teaching, 9*, 377- 381.

Shaver, J. P. (1993). What statistical significance testing is, and what is not. *The Journal of Experimental Education, 61*, 293- 316.

Wampold, B. E., Furlong, M. J., & Atkinson, D. R. (1983). Statistical significance, power, and effect size: A response to the reexamination of reviewer bais. *Journal of Counseling Psychology, 30*, 459- 463.

Xitao, F. (2001). Statistical significance and effect size in education research: Two sides of a coin. *Journal of Educational Research, 94* (5), 275-283.