International Journal for Research in Education

Volume 43 | Issue 2 Article 4

2019

An Analytical Study about the Dialectical Use of Statistical Significance in Psychological and Educational Research

عايش صباح dr
جامعة مولاي الطاهر سعيدة الجزائر, aichsabah@yahoo.fr

المجلة الدولية للأبحاث التربوية- جامعة الإمارات العربية المتحدة
المجلد (43) العدد (2) يوليو 2019

Statistical significance in psychological and educational research: methodological issues

Aiche Sabah
Faculty of Human and Social Sciences, Hassiba Benbouali University of Chlef, Algeria

Abstract.

The statistical significance is almost 300 years ago. The aim was to test validity of the results in social sciences. Statistical significance faces many comments and arguments from researchers who asked to neglect its importance not only in social sciences but also in all scientific research. Over 50 years of discussion, researchers still ask: the same question, why do educationalists and psychologists are still using statistical significance.

Keywords:statistical significance ; psychological and educational research ; methodological issues .

الدلالة الإحصائية في البحوث النفسية والتربوية: قضايا منهجية

عايش صباح
كلية العلوم الإنسانية والاجتماعية جامعة حسيبة بن بوعلي الشلف- الجزائر

مستخلص البحث :

لقد اعتمدت الدلالة الإحصائية منذ أكثر من 300 سنة، وخدمت غرضا هاما في النهوض بالبحوث في العلوم الاجتماعية، ومع ذلك، ساد الكثير من الجدل حول سوء استخدام وتفسير هذا الاختبار . إن الانتقادات المقنعة بعدم جدوى استخدام اختبار الدلالة الإحصائية التي يمكن العثور عليها في جميع المجالات العلمية تقريبا في العلوم الاجتماعية، وعلوم الحياة، والاقتصاد، وعلم الاجتماع، وعلم النفس وعلوم التربية وغيرها تحتاج أن تؤخذ على محمل الجد، لكن وبعد نصف قرن من الحجج المقنعة والدعوة إلى اعتماد ممارسات بديلة في بعض التخصصات، مثل علم النفس وعلوم التربية، يبقى السؤال المطروح هو: "لماذا لا يرغب الباحثون النفسيون والتربويون في التخلي عن هذه الممارسات الاحصائية؟"

الكلمات المفتاحية: الدلالة الإحصائية؛ البحوث النفسية والتربوية؛ قضايا منهجية.

المجلة الدولية للأبحاث التربوية- جامعة الإمارات العربية المتحدة
المجلد (43) العدد (2) يوليو 2019

مقدمة

يشار إلى الاحصاء باعتباره ذلك العلم الذي يهتم بجمع البيانات ومن ثم تنظيمها وعرضها باستخدام الأساليب العلمية لتحليلها واستخلاص النتائج منها، وللإحصاء أهمية كبيرة في مختلف العلوم، فهو يؤثر ويتأثر بها في نطاق تطورها المستمر عبر التقدم التكنولوجي المعاصر، حيث تحتل الطرق والنظريات الإحصائية مكانة مرموقة في شتى العلوم، ذلك أن افتقاد الجهد الاحصائي في أي بحث علمي يجعله مهددا بعدم الدقة، والاعتماد على نتائجه يعد ضربا من المخاطرة.

والعلوم التربوية والنفسية كغيرها من العلوم الأخرى تحتاج إلى الإحصاء، حيث يحتل هذا الأخير مكانة مهمة في البحوث التربوية والنفسية والاجتماعية، فهو يعين الباحثين على جمع البيانات الكمية، واختيار الأساليب المناسبة في معالجة تلك البيانات، وتحليل نتائج بحوثهم المختلفة التي يتم إجرائها على عينات صغيرة ممثلة لمجتمع البحث، واستخلاص بعض الاستنتاجات من دراسة عينة صغيرة لصياغة تعميمات يمكن تطبيقها على مجتمعات أكبر حجما، فاستخدام الإحصاء في الوقت الحاضر للحصول على التعميمات العلمية Scientific Generalizations من البيانات المتوفرة يعتبر من أهم أغراض الإحصاء.

وقد اهتم الباحثون في مجال علم النفس والتربية بالتحقق من دقة النتائج التي يتم التوصل إليها عند التحقق من صحة الفروض، لذلك اعتمدوا على عدة مفاهيم إحصائية من بينها الدلالة الإحصائية، حيث يقوم أغلب الباحثين التربويين بحساب ما يسمى "الدلالة الاحصائية" ، وهي قيمة الاحتمال الفاصل بين رفض أو عدم رفض الفروض الاحصائية للبحث، ويرمز لمستوى الدلالة بالرمز (α) وتقرأ الفا، وعادة ما يستخدم التربويون مستوى دلالة إما 0.05، أو 0.01 ، وطبقا لمستوى الدلالة والقيمة المحسوبة للاختبار الاحصائي يتم مقارنة القيمة الجدولية "القيمة الحرجة" للاختبار الاحصائي بالقيمة المحسوبة وعليه يتم رفض أو عدم رفض الفرض الاحصائي، والواقع أن مستوى الدلالة ما هو إلا قيمة احتمالية، وهي شرط ضروري ولكن غير كاف لرفض الفرض الاحصائي، وتركيز الباحثين أو اقتصارهم على هذ المفهوم يعد قصورا معيبا في فهم الأساليب الإحصائية الاستدلالية (سلامة،2004، ص3).

لقد اعتمدت الدلالة الإحصائية منذ أكثر من 300 سنة، وقد خدمت غرضا هاما في النهوض بالبحوث في العلوم الاجتماعية، ومع ذلك، ساد الكثير من الجدل حول سوء استخدام وتفسير اختبار الدلالة إحصائية، وقد أشار إلى ذلك Pedhazur & Schmelkin (1991) بقولهما : قليلة ربما هي القضايا المنهجية التي تولد عنها الكثير من الجدل بين العلماء الاجتماعيين والسلوكيين مثل استخدام الدلالة الإحصائية (Daniel, 1998).

وكان هذا الخلاف واضحا في أدبيات الإحصاء في العلوم الاجتماعية منذ وقت طويل، فقد تطرقت العديد من المقالات والكتب إلى المشاكل المتعلقة بالدلالة الإحصائية، حيث تعرضت منهجية البحث في السنوات الأخيرة إلى هجوم على استخدام اختبار الدلالة الإحصائية؛ وقد قدمت طبعة كاملة من مجلة Experimental Education (Carver, 1993) لاستكشاف هذا الجدل.

وقد انقسم الباحثون بين من يوصي بالتخلي الكامل عن اختبار الدلالة الإحصائية كخيار منهجي(McLEAN & Ernest, 1998)، من منطلق أنه مع رسوخ هذا الأسلوب أصبحت اختبارات الدلالة الاحصائية لاختبار الفرضيات في الأبحاث التربوية طريقة تقليدية، وهذا ما أشار إليه كسيتر "Keaster" (1998) بقوله أن اختبارات الدلالة الإحصائية قد تحولت من أساليب إلى طقوس (بابطين، 2001، ص 2)، فيما تجاهل البعض الآخر هذا الجدل، وقرر استخدام اختبار الدلالة

الاحصائية وفقا للممارسات التقليدية، التي تعتمد اختبار الدلالة الاحصائية كأسلوب احصائي علمي لتقرير ما إذا كانت الفروق أو الارتباطات أو العلاقات بين المتغيرات دالة إحصائيا.

ومنذ عام 1962، وطيلة العقود الماضية، انتقدت عدد من المقالات سوء استخدام الدلالة الاحصائية ومن هذه الأبحاث نذكر:(Bakan, 1966; Daniel, 1998; Dar, 1987; Gigerenzer & murray, 1987; Hubbard & Lindsay, 2008; Morrison & Henkel, (2006

ولعل أبلغ انتقاد وجه لاستخدام الدلالة الاحصائية في البحوث هو دعوة الجمعية الأمريكية للإحصاء"ASA" (2016) الباحثين إلى حسن استخدام هذه الأداة والوقف الفوري عن استخدامها كأداة وحيدة لاتخاذ القرار(Wasserstein & Lazar, 2016).

لذا تحاول هذه الورقة البحثية الكشف عن أهم مشكلات اختبار الدلالة الاحصائية ومدى جدواها في البحوث التربوية والنفسية ، وذلك من خلال الإجابة على التساؤلات التالية:

- ما أهمية الدلالة الاحصائية في البحوث التربوية والنفسية؟
- ما هي أهم الانتقادات الموجهة للدلالة الاحصائية؟
- ما هي الحلول البديلة للدلالة الاحصائية؟

أهمية البحث والحاجة إليه

تعتبر الدراسة الحالية مساهمة في أدبيات البحث في الاحصاء النفسي، حيث تساعد في فهم الدلالة الاحصائية ودورها في علم النفس خصوصا في ظل الانتقادات الكبيرة والجدل القائم حولها حاليا.

وإن كانت هذه المقالة وغيرها من المقالات والدراسات الأخرى لن تغير واقع استخدام الدلالة الاحصائية التي ترسخت في العلوم النفسية والتربوية إلا أنها تعبر عن وعي الباحثين بعدم جدوى هذه الأخيرة وتعتبر مؤشرا ايجابيا قد يؤدي إلى إيجاد حلول بديلة حاسمة في المستقبل.

حدود البحث

الحدود المنهجية: لقد تم الاعتماد على منهج البحث النوعي، حيث تم من خلال هذا البحث تحليل مختلف الآراء حول استخدام الدلالة الاحصائية ، وتوضيح أهمية الدلالة الاحصائية وأهم مآخذها ومن ثم الحلول البديلة كما وردت في أدبيات الدلالة الاحصائية.

الحدود البشرية: يعتمد هذا البحث على المجهود الفردي للباحثة.

الحدود المرجعية: إن المراجع التي تم اعتمادها في هذا البحث هي الكتب والابحاث التي قامت بدراسة الدلالة الاحصائية سواء العربية منها أم الانجليزية.

أولا: أهمية الدلالة الاحصائية في البحوث التربوية والنفسية:

إن زيادة البحوث العلمية الكمية ومجموعات البيانات المعقدة في السنوات الأخيرة أدى إلى توسيع نطاق تطبيقات الأساليب الإحصائية، وقد خلق هذا آفاقا جديدة للتقدم العلمي، لكنه أثار أيضا مخاوف بشأن الاستنتاجات المستخلصة من البيانات البحثية، ذلك أن صحة الاستنتاجات العلمية بما

المجلة الدولية للأبحاث التربوية - جامعة الإمارات العربية المتحدة
المجلد (43) العدد (2) يوليو 2019

في ذلك العلوم التربوية والنفسية يعتمد على أكثر من الأساليب الإحصائية نفسها؛ فتقنيات اختيار الأسلوب الاحصائي المناسب، وإجراء التحليلات بشكل صحيح، والتفسير الصحيح للنتائج الإحصائية أيضا يلعب دورا رئيسيا في ضمان سلامة النتائج.

وقد عززت العديد من الاستنتاجات العلمية المنشورة مفهوم "الدلالة إحصائية"، فما هي الدلالة الاحصائية؟ ، وما هي أهم الانتقادات التي وجهت إليها؟

تشير "الدلالة الإحصائية" "Statistical Significance" إلى النتيجة إذا ما كانت ناتجة عن الصدفة أو المتغيّرية "Variability" في العيّنة (حسن،2016، ص1) ، فهي تهدف إلى الكشف عن مدى اقتراب المقاييس الاحصائية للعينات من مقاييس الاصل (المجتمع) وتزداد الثقة في مقاييس العينة كلما اقتربت من أصلها أو كلما كان تذبذبها حول هذا الأصل ضيق (الحداد، 2006، ص 67) .

وبذلك يكون المفهوم المبسط لاختبار الدلالة الإحصائية بأنه إجراء ضمن اختبار الفرضيات لتحديد احتمالية النتيجة (أ) عند افتراض صحة الفرضية الصفرية H0 وافتراضية عشوائية البيانات(عشوائية الاختبار والتعيين وعند حجم العينة (n) (باهي، 2011، ص 1390) .

وتعد الطريقة الإحصائية أكثر الطرق شيوعا من حيث كونها طريقة من طرق صناعة القرار تحت شرط حالة عدم التأكد، أي شرط الاحتمالية، لذا يصيغ الباحث في ظل استخدام الطريقة الإحصائية نوعين من الفروض الإحصائية كترجمة لغرض البحث تحت الدراسة (البارقي، 2012، ص 14)، هذان النوعان من الفروض هما الفرض الصفري الذي ينفى أو يلغى وجود الظاهرة بشكل أو بآخر، والفرض البديل هو الذى يتحدث عن وجود الظاهرة بشكل أو بآخر، ويفترض الباحث صحة الفرض الذى يرغب في اختباره ، ثم يفحص نتائج هذا الفرض في ضوء توزيع العينة الذى يعتمد على صحة الفرض ، وإن أي حكم أو قرار يتخذه الباحث بصدد الفرض الصفري يحتمل الصحة أو الخطأ ويكون هكذا أمام نوعين من الاحتمالات:

- وجود الظاهرة في العينة وليس لها وجود فعلي في المجتمع الأصلي؛ هذا الاحتمال نرمز له بالرمز (α) وينطق ألفا Alpha (الخطأ من النوع الأول أو "خطأ ألفا"type one error).

– عدم وجود الظاهرة في العينة ولكن لها وجود في المجتمع؛ هذا الاحتمال نرمز له بالرمز (β) و ينطق بيتا Beta (الخطأ من النوع الثاني أو "خطأ بيتا" Type two error) ، ألفا وبيتا كلاهما يعبر عن الشك في النتيجة التي توصلنا إليها أو الشك في طريقة الاستدلال، ويرتبط بخطأ ألفا ما يسمى بـ "الدلالة الإحصائية "، وهى تعبر عن الثقة ، ويرتبط بخطأ بيتا ما يسمى بـ "قوة الاختبار الإحصائي" "Power of Statistical Test" (الرفاعى وآخرون،2005، ص 21، 22) .

وقد أشار (Johnson, 1999) إلى أنه توجد عدة تفسيرات للدلالة الاحصائية، فينظر تارة إلى الدلالة الاحصائية على أنها احتمال الحصول على النتائج عن طريق الصدفة، حيث تشير قيمة الدلالة الصغيرة إلى أن النتائج المتحصل عليها لا ترجع للصدفة، أما قيم الدلالة الكبيرة فتشير أن النتائج المحصل عليه كانت نتيجة للصدفة. وفي أحيان أخرى تعتبر قيمة الدلالة (P-1) يساوي ثبات النتيجة "reliability of the result"، أي احتمال الحصول على نفس النتيجة عند إعادة التجربة، وفي ظل هذا التفسير يشار إلى الفروق الدالة إحصائيا غالبا بالمصطلح "ثابت" "reliable ". أما التفسير الثالث فهو اعتبار أن قيمة (p) تمثل احتمال أن الفرض الصفري يكون صحيحا، وهذا هو التفسير المباشر الذي يدرس بطريقة مباشرة السؤال الذي يهتم به الباحث.

ولقد حظيت مجالات التربية وعلم النفس بعلاقة وثيقة مع العلوم الإحصائية لعقود من الزمن، حيث أدت التطورات الهامة في قضايا القياس النفسي والتصميمات البحثية إلى الاستفادة من علم الإحصاء ودمج مختلف هذه التطورات الإحصائية لصالح تطور علم النفس، ويعتبر الاختبار المعروف بالدلالة الإحصائية الوسيلة الأكثر استخداما لتقييم الفرضيات والنظريات.

حيث يتمتع اختبار الدلالة الإحصائية بشعبية كبيرة في علم النفس منذ ما يزيد عن 50 عاما، وفي مجال الإحصاء منذ ما يقرب من 80 عاما، إلا أنه منذ 1960، بدأ تراكم الأدب النقدي في علم النفس والعلوم السلوكية فيما يتعلق بجدوى استخدام الدلالة الاحصائية على سبيل أبحاث كل من (Harlow, Mulaik, & Steiger, 2013; Hubbard, 2015; Morrison & Henkel, 2006; Nickerson, 2000).

أثارت هذه الكتابات موجة من الانتقادات فيما يتعلق بقبول الممارسات في مجال العلوم الاجتماعية والسلوكية التي تسند قراراتها إلى الدلالة الاحصائية، حتى أصبحت تعرف باسم "اختبار الجدل" "The Significance Test Controversy" الذي ظهر في شكل مؤلف للباحثين (Morrison & Henkel, 2006)، ويرى هؤلاء النقاد أن العديد من الباحثين يسيئ استخدام الاختبارات الإحصائية، الأمر الذي يقوض صحة الاكتشافات المبلغ عنها.

وعلى الرغم من كثرة الانتقادات الموجهة للدلالة الإحصائية إلا أن الباحثين في علم النفس وعلوم التربية مازالوا يستخدمونها بشكل غير لائق، ويسيئون فهمها في أغلب الاحيان، غير مبالين بكثرة الجدل الذي تولد عن استخدام هذا الاختبار(Haig, 2017).

هذا الجدل جعل مجموعة من الباحثين يطالعوننا بالسؤال التالي في شكل مؤلف: "ماذا لو لم تكن هناك دلالة إحصائية؟" " What if there were no significance tests ? (Harlow et al., 2013) ، وقد كان هذا السؤال أول ما واجه الباحثين في مطلع القرن العشرين؛ وقد استعراض هذا المؤلف إعادة النظر في الدلالة الاحصائية على مر السنين، وكذا العديد من الحجج المؤيدة والمعارضة لاختبار الدلالة الاحصائية.

فيما تعرضت الباحثة (Kline, 2004) في كتابها "Beyond significance testing" إلى ضرورة تغيير ممارسات تحليل البيانات في علم النفس والتخصصات ذات الصلة، وهذا نتيجة للجدل الكبير والطويل حول الاختبارات الإحصائية في العلوم السلوكية، وتزايد عدد المجلات التي تطلب معلومات حول حجم التأثير؛ لهذا جاء مؤلفها "ما وراء الدلالة الاحصائية" كي يوضح عيوب الدلالة الاحصائية ويستعرض طرق بديلة لها مثل تقدير حجم التأثير وإعادة النظر في تقدير فترة الثقة.

فيما يرى (Denis, 2003) أنه على الرغم من الانتقادات الموجهة للدلالة الاحصائية عبر سنوات من الزمن، فإن الدلالة الاحصائية هي النموذج الأكثر استعمالا في الاستدلال الإحصائي في علم النفس؛ فمنذ "فيشر" (1925) استخدم علماء النفس بشكل روتيني اختبار الدلالة الاحصائية على الرغم من العيوب المنهجية والفلسفية العميقة المتعلقة بها، وهذا راجع حسبه إلى افتقار البدائل المتاحة لذلك، وقد سعى هذا الباحث إلى تقديم بدائل للدلالة الاحصائية ؛ فيما سماها (Reese, 2004) "المقدس 0.05" "The sacred 0.05"، وقال أن اختيار هذه القيمة الحرجة هو خيار تعسفي ، (Cohen, 1995)هاجم أيضا مبادئ فرضية العدم واختبار الدلالة الاحصائية في ورقة كلاسيكية بعنوان "الأرض كروية ($p < 0.05$) The earth is round " بقوله: "بعد أربعة عقود من الانتقادات اللاذعة للدلالة الاحصائية ، طقوس افتراض فرضية العدم والمستوى المقدس لدلالة الاحصائية 0.05 لا تزال قائمة".

المجلة الدولية للأبحاث التربوية- جامعة الإمارات العربية المتحدة
المجلد (43) العدد (2) يوليو 2019

ويعود استخدام للدلالة الاحصائية لعام (1710)، لكن استخدامها بالشكل الحالي يرجع إلى "كارل بيرسون الذي وضع في عام (1900) اختباره كاي مربع الذي يعتبر أول اختبار إحصائي حديث، وبعد وقت قصير نشر "Gosset" العمل الرائد لتطوير اختبار ت t test.

والنهجين الأكثر تأثيرا على اختبار الدلالة الإحصائية الحديثة، وضعا من قبل Fisher (1925، 1935) و Neyman–Pearson (1933) في أوائل ومنتصف 1900، وقد تم تطوير نهج فيشر لاختبار الفرضيات الإحصائية باعتباره النهج العام للاستدلال العلمي، في حين تم تصميم نموذج نايمان بيرسون لاتخاذ القرارات التطبيقية، في نهج فيشر، يتم صياغة فرضية معينة لكي يتم اختبارها عن طريق بيانات احصائية، وهناك نوعين من الفرضيات ، فرضية صفرية وفرضية بديلة، وبناء على بيانات العينة نقوم إما برفض Ho أو بعدم رفض Ha ، ويجب تحديد فرضية العدم والفرضية البديلة بشكل صحيح وإلا سوف تكون النتائج غير صحيحه في عملية اختبار الفرضية، وهكذا فإن استخدام التعيين العشوائي، وفرضية العدم، والاتساق والكفاءة ، والدلالة الاحصائية، ومستوى الدلالة 0,05 هي بعض من الاسهامات البارزة لـ "فيشر"(Levine, Weber, Park, & Hullett, 2008).

إلا أنه عندما أدخل الإحصائي البريطاني رونالد فيشر اختبار مؤشر "بي" في عشرينات القرن الماضي، لم يقصد به أن يكون اختبارا حاسما نهائيًا، كانت نيته أن يكون الاختبار طريقة غير رسمية للحكم على الدلالة الإحصائية للدليل بالمعنى التقليدي: أي أنه يستحق نظرة ثانية، وكانت الفكرة إجراء تجربة، ثم النظر فيما إن كانت النتائج تتسق مع ما تسفر عنه فرصة عشوائية أم لا.

وينبغي للباحثين أولًا تعيين فرضية العَدم (الصفرية) التي يريدون إبطالها، مثل عدم وجود ارتباط أو اختلاف بين مجموعتين، ثم يقومون بدور المُدافِع عن فرضية العدم، ثم يحسِبون احتمال الحصول على نتائج لا تقل تطرفًا عن النتائج المرصودة فعلًا في التجربة، هذا الاحتمال هو مؤشر "بي"، ورأى فيشر أنه كلما صغرت قيمة المؤشر، تعاظمت أرجحية بطلان فرضية العدم (نوتزو، 2014، ص 29)، ويستخدم اختبار الدلالة الإحصائية للاستدلال على خصائص المجتمع من خلال العينة المأخوذة منه، وتسمح تلك الاختبارات للفرد بحساب احتمالية الحصول على نتائج العينة بافتراض أن الفرض الصفري صحيحًا (أي افتراض أن المجتمع يتصف بما هو مفترض في الفرض الصفري) (حسن،2016 ، ص5) .

وقد رأى فيشر أن الفشل في رفض الفرض الصفري يعني في الحقيقة أن بياناتنا لا تكفي للاختيار بين هذه البدائل، والأصح عندئذ تعليق الحكم، وقد اتخذ Nyman & Pearson (1933) موقفا مختلفا وأكثر عملية إزاء هذه المسألة، فموقف تعليق الحكم يقول لنا (وخاصة لمتخذي القرارات العملية هنا) انظروا حتى يتم إجراء بحوث أخرى ومن نتائجها يمكن حسم المسألة ورفض الفرض الصفري، بينما الفرض الصفري قد يكون أصيلا بالفعل في نظرية البحث ذاتها، ناهيك أنه قد لا تتوافر للباحث الإمكانات لتكرار البحث عدة مرات، ولذلك اقترح ببيرسون وزميله على الباحث أن يختار بين قبول الفرض الصفري أو رفضه، وإنما ببساطة سوف نتصرف ولو مؤقتا حتى تتوافر لنا بيانات أكثر ملائمة كما لو كان صحيحا وفي حالتي القبول أو الرفض يجب أن يكون اهتمامنا أكثر تركيزا على احتمال القبول الزائف أو الرفض الزائف للفرض الصفري(المطيري وآخرون، د.س، 37) .

وبينما احتدم الصراع بين الغرماء، اعتبر "نيمن" بعض أعمال "فيشر" رياضيًّا «أسوأ من كونها حتى عديمة الفائدة»؛ واعتبر "فيشر" نهج "نيمن" «طفوليًّا» و«مفزعًا بالنسبة إلى حرية الفكر في الغرب»، في حين فقد باحثون آخرون صبرهم، وشرعوا في كتابة كتب إرشادية في الإحصاء للباحثين، ونظرا إلى أن كثيرا من المؤلفين لم يكونوا إحصائيين، وليس لديهم الفهم الكافي لأي من

النهجين، فقد أنشأوا نظاما هجينا حشروا فيه مؤشر "فيشر" لقوة الدليل سهل الحساب، مع نظام "نَيمن وبيرسون" الصارم القائم على القواعد، يقول "جودمان": «لم يكن أبدا القصد من حساب مؤشر "بي" أن يُستخدَم على نحو استخدامه اليوم" (نوتزو، 2014، ص 29) .

ويرى (Ziliak & McCloskey, 2008) أنه على مدى السنوات الثمانين الماضية، يبدو أن بعض العلوم أخطأت في إسناد قراراتها إلى الدلالة الإحصائية ، إذ على الرغم من أنها تبدو في البداية مسألة تفصيل إحصائي بسيط إلا أن الأمر ليس كذلك؛ فالإحصاءات والمتغيرات والمعاملات هي الأدوات العلمية الأساسية ، كما أن الاحصاء الرياضي هو الإنجاز الاجتماعي والعملي والجمالي الجيد، ولا أحد يستطيع أن يشكك بمصداقيته، إن الفهم البشري للاحتمال وعدم التيقن كان لينخفض كثيرا لولا الأساليب الإحصائية المختلفة كإحصاءات بايز Bayesian Statistics ، ووظائف غاما، ومنحنى الجرس، وبقية الإحصاءات، إلا جزء واحدا من الإحصاء الرياضي يعتبر خطأ فادحا، فالنتائج العلمية المرتبطة بالاختبار والقياس والتفسير المستندة الى "الدلالة الإحصائية" كما فعلت بعض العلوم لأكثر من ثمانين عاما، كانت فكرة سيئة للغاية.

وقد حقق (Meehl, 1967, 1978, 1997) واحدة من أقوى الانتقادات لاستخدام اختبار الدلالة الإحصائية في علم النفس حين اعتبر أن الاستخدام الواسع للدلالة الإحصائية في اختبار الفرضيات والنظريات النفسية أمر معيب، لأن دعم الفرضية أو النظرية التي يحصل عليها الباحث بناء على رفض فرضية العدم ضعيف جدا، فأحيانا ينطلق الباحثون في علم النفس وعلوم التربية من فرضية صفرية تنفى أو تلغى وجود الظاهرة أي أن الفرق بين معلمات المجتمع المرتبط بالدراسة يساوي صفر، ولكن الحقيقة المعروفة منذ فترة طويلة للإحصائيين المهنيين وفي نظر "ميهل" "Meehl"، هو أن فرضية العدم هي غير صالحة في مجال العلوم السلوكية والاجتماعية، والسبب في ذلك أنه من غير المحتمل وجود متغيرين في العلوم السلوكية والتربوية غير مرتبطين ببعضهما بعلاقة من درجة معينة، فالمتغيرات غالبا مرتبطة ببعضها البعض في أجزاء كثيرة من علم النفس مما أدى إلى تشعب إيجابي كبير في العديد من المتغيرات ترتبط بشكل إيجابي مع بعضها البعض لدرجة كبيرة (Haig, 2017)، لذلك فإن إجراء الدراسات البحثية باستخدام افتراضية استقلال هذه المتغيرات وكذلك افتراض الباحث لوجود أو عدم وجود علاقات بين المتغيرات والتوصل إلى قبول أو رفض هذه الفروض لا يعني أكثر من الانغماس في الشكلية العلمية وضياع لوقت وجهد الباحث دون جدوى (السعيد، 2009، ص127) ، فالظواهر النفسية ذات طبيعة معقدة تحتوي متغيرات وعوامل متعددة، وذلك من حيث كون هذه المتغيرات والعوامل قد تكون مسببات أو مدخلات أو على الأقل متغيرات مستقلة (الصياد، 1985، ص 212) ، وهكذا في مجال علم النفس، حيث التجارب "الحقيقية "غالبا ما تكون غير ممكنة، فإن الحصول على حجم عينة معقول يجعل تحقيق نتيجة ذات دلالة إحصائية هي النتيجة المحتملة للدراسة الميدانية (Haig, 2017)

أهم الانتقادات الموجهة للدلالة الاحصائية

أولا: حساسيتها لحجم العينة Sensitivity to sample size

لعل الانتقاد الأكبر الموجه للدلالة الإحصائية، والمعترف به على نطاق واسع في أدبيات الاحصاء هو حساسيتها لحجم العينة كما ذكر باحثون منهم (Boster, 2002; Cohen, 1990; Royall, 1986)، فعندما يكون حجم العينة صغيرا تكون التأثيرات قوية لكنها قد تكون غير دالة،

المجلة الدولية للأبحاث التربوية- جامعة الإمارات العربية المتحدة
المجلد (43) العدد (2) يوليو 2019

وهنا نقع في خطأ من النوع الثاني ، بدلا من ذلك، عندما يكون حجم العينة كبيرا قد تكون التأثيرات تافهة لكنها تثير الاعجاب لأنها دالة احصائيا، على سبيل المثال، ر (ن:20) = 0.40 لكنها غير دالة عند مستوى دلالة 0.05 ، في حين أن ر (ن:1000) = 0.07 لكنها دالة إحصائيا.

وهكذا فإن قيمة الدلالة الإحصائية بالنسبة للفرضية يعكس حجم العينة، والحصول أو عدم الحصول على الدلالة الإحصائية يرجع إلى حجم العينة إضافة إلى عوامل أخرى، وهذا يمكن أن يؤدي إلى رفض نتائج قد تكون هامة عند عينة صغيرة والأخذ بعلاقات تافهة مع عينات كبيرة (Levine et al., 2008).

يرى (Royall, 1986) في هذا الصدد أن الدلالة الاحصائية في العينات الصغيرة والكبيرة متناقضة تماما، وهذا ما يجعل من الصعب الحصول على قيمة ثابتة وموضوعية وذات معنى، إذ أنها تتوقف فقط على حجم العينة، ويشير(Hubbard & Lindsay, 2008) إلى أن الدلالة الاحصائية غير مفيدة في العينات الكبيرة، فعندما يكون حجم العينة كبيرا سيكون لفرضية العدم قيمة احتمالية ضئيلة.

وذلك لأنه يمكن الحصول على فروق دالة إحصائيا بموجب فحص الفرضيات الإحصائية لمتوسطات متقاربة القيم فقط لأن الباحث قد اختار عينة كبيرة في دراسته، ومن المعروف أن قيمة الخطأ المعياري Standard Error تقل كلما كبر حجم العينة، وبالتالي تكون الدلالة الإحصائية دالة لحجم العينة، أو لمستوى الدلالة، بدلاً من أن تكون دالة للفروق العملية بين المتوسطات (practical differences)، أي أنه يمكن للباحث عند اعتقاده بوجود علاقة ضعيفة بين متغيرات دراسته (المستقلة و التابعة) أن يزيد حجم عينة دراسته للحصول على نتائج دالة إحصائياً ، فالباحث – أي باحث – يستطيع غالبا رفض الفرض الصفري، ويعني وجود دلالة إحصائية عن طريق زيادة حجم العينة زيادة كافية، هنا يختلط الأمر على الباحث هل الدلالة الإحصائية تعني إن المعالجة أو المتغير العامل (المتغير المستقل) تحت الدراسة لها تأثير على الناتج Outcome (المتغير الناتج) أم أن هذه الدلالة الإحصائية جاءت نتيجة لكبر حجم العينة (البارقي، 2012، ص 30) .

وهكذا فإن جل الباحثين تقريبا يعتمدون محك الدلالة الإحصائية وحدها دون مراعاة شدة العلاقة أو قوتها، والأمران (الدلالة الإحصائية وشدة العلاقة) غالبا لا يجتمعان، فقيم معاملات الارتباط صندوق أسود لا تنصاع إلى القراءة المباشرة، ولذلك يستحسن تحويلها إلى معاملات تحديد بتربيع قيم معاملات الارتباط حتى يسهل تصورها أي قراءتها، ويسهل تبين شدتها، فالارتباط الذي قدره 0.30يساوي عند تربيعه 0.09 أي 9%، ومعناه أن العلاقة تفسر نسبة مئوية مقدارها تسعة بالمائة من مجمل التباين، والارتباط الذي قدره 0.20 يساوي عند تربيعه 0.04 ، أما باقي التباين أو المعلومات التي تحتوي عليها العلاقة (96%) يكاد يكون كل التباين أو المعلومات، على الرغم من أن هذه العلاقة دالة إحصائيا إذا كان حجم العينة 100 فردا.(تيغزة، 2017، ص 2)

مستوى الدلالة والتعسفية

قام كل من (Johnson, 1999; William, 2000) بالتذمر من مستوى الدلالة الاحصائية التي تنطوي على استخدام نقطة قطع تعسفية، فالباحثون لا ينبغي أن يكونوا ملزمين بقيمة $\alpha =$ 0.05.، ذلك أن الكثير من الناس يسيئون استخدام الدلالة الاحصائية عندما يتخذون قرارا برفض أو فشل دراساتهم، وربما يرجع ذلك فقط إلى إرث نايمان وبيرسون.

وهذا ما جعل (Coolidge, 2012) يطرح التساؤل التالي: "هل الله حقا يحب مستوى الدلالة 0.05 أكثر من مستوى 0.06؟" ، وذلك نتيجة للاتجاه السائد لدى الباحثين والطلبة لاعتبار مستوى الدلالة الأكبر من 0,5 ليس له أهمية.

فاختبار الدلالة الإحصائية ينفذ بعد تحديد مستوى الدلالة (α)، ومستوى (α) هي احتمالية أن تؤخذ النتيجة المعطاة بسبب أخطاء المعاينة، بمعنى احتمالية ارتكاب الخطأ من النوع (I) ، وهو رفض الفرضية الصفرية وهي صحيحة، ومستويات الدلالة (α) التقليدية 0.05 أو 0.01 المستعلمة من قبل الباحثين هي تقليد عشوائي، عندما كان من الصعب حساب القيمة الدقيقة لـ (P) لإحصاء الاختبار، بدلاً من ذلك استخدم النّاس الجدول لقيم إحصاءة الاختبار المقابلة لقيم قليلة ومختارة وعشوائية لـ (P) وهي 0.05، وفي بعض الأوقات 0.01. هذه القيم أصبحت مقدّسة كعتبة حدية للقيم التي تحدّد الدلالة الإحصائية (النعيمي ، عبد الله، 2012، ص 95)، ويرى (Robinson & Wainer, 2001) أن "فيشر" "Fisher" (1925) ذكر بنفسه أنه لا يوجد عمل علمي لديه مستوى ثابت من مستوى الدلالة في كل الاوقات، وفي جميع الظروف، وقال أنه بدلا من ذلك على الباحث أن يعطي رأيه في كل حالة على حدى في ضوء بياناته وأفكاره.

ماذا لو أن p = 0.06؟

ناقش "هوبكيتز" "Hopkins" (1997) مفهوم مستويات الدلالة الاحصائية بقوله: ما هو الخاص في القيمة 0,05؟ في الحقيقة لا شيء، شخص ما قرر أن هذه القيمة معقولة ، ولذا فإننا جامدون عندها (بابطين، 2001، ص22)، في حين تذمر (Johnson, 1999; William, 2000) من اعتبار أن النتائج التي مستوى دلالتها أقل من 0,05 دالة فيما أن القيم الأكبر منها حتى لو كانت 0.06غير دالة.

كما ذكر سابقا، فإن فيشر وضع مستوى دلالة 0.05 كمعيار للدلالة الاحصائية، لكنه بم يضع لها حدا ثابتا بل دعا الباحثين إلى تقويم الحالات الفردية بالرجوع للبيانات والاطار النظري، إذا لا بد من تفسير قيم مستوى الدلالة في السياق الذي تم اجراء فيه الدراسة.

احصائيا، النتائج ذات الدلالة الإحصائية توفر أساس الإنجاز العلمي، لكن بخصوص النتائج التي تكون فيها الدلالة الاحصائية أكبر من 0.05 ولكن لا تزال قريبة منها، أي أقل من 0.25، يقترح "توكي" "Tukey" (1991) أن نستخدم كلمات أخرى غير كلمة دالة وغير دالة لوصف ترددنا في الرهان على اتجاه فروق أو علاقة حقيقية، على سبيل المثال، إذا كانت العلاقة الارتباطية أكبر من 0.05 ولكن أقل من 0.15، يمكن أن نقول أن العلاقة تميل في اتجاه معين، وإذا كانت قيمة الارتباط أكبر من 0.15 ولكن أقل من 0.25، يمكن أن نقول أن هناك تلميح حول وجود علاقة وهكذا (Robinson & Wainer, 2001).

الحلول البديلة

في عام 1996، عقدت "فرقة العمل المعنية بالاستدلال الإحصائي Task Force on Statistical Inference" (TFSI). التابعة "لجمعية علم النفس الأمريكية" "American Psychological Association" (APA) اجتماعا للرد على الجدل الدائر منذ فترة طويلة حول اختبارات الدلالة الإحصائية وتوضيح طرق بديلة، أين كان يأمل البعض أن "فرقة العمل المعنية بالاستدلال الإحصائي" "TFSI" سوف توصي بفرض حظر على اختبارات الدلالة الإحصائية في مجلات علم النفس.

المجلة الدولية للأبحاث التربوية- جامعة الإمارات العربية المتحدة
المجلد (43) العدد (2) يوليو 2019

لقد نوقش هذا الحظر في عدد خاص من مجلة "العلوم النفسية" "Psychological Science" (1997) ، وكذا مجلة "البحث في المدارس" "Research in the Schools" (1998)، وكتاب حرره "هارلو، مولايك، وستيغر" "Harlow, Mulaik, and Seiger" (1997) بعنوان: "ماذا لو لم تكن هناك اختبارات الدلالة الاحصائية؟ "، كما نوقشت مشاكل اختبارات الدلالة الاحصائية والحلول البديلة ايضا في عدد خاص من مجلة "تجارب التعليم" " Experimental Education"(1993)(Kline, 2004) .

وقد دعت "الجمعية الأمريكية للإحصاء" "ASA" (2016) في بيان لها الباحثين إلى حسن استخدام قيمة P-value في الدراسات الإحصائية، والوقف الفوري لاستخدامها كأداة شبه وحيدة لاتخاذ القرار، حيث يقول "Wasserstein" المدير التنفيذي للجمعية الأمريكية للإحصاء: "إن الدلالة الاحصائية لن تكون أبدا بديلا عن المنطق العلمي"، فيما قالت "Jessica Utts" رئيسة الجمعية الامريكية للإحصاء أن بيان "ASA" هو توجيه البحوث نحو "فترة ما بعد عصر "0.05، إذ مع مرور الوقت، يبدو أن قيمة الدلالة الاحصائية أصبحت بوابة للنشر على الأقل في بعض التخصصات(Wasserstein & Lazar, 2016).

إن هده المناقشة الجدية لمثل هذا الحظر يعكس تتويجا لسنوات عديدة من الاستياء من اختبار الدلالة الإحصائية، وإن الحديث عن فرض حظر على الدلالة الإحصائية ربما يأتي بمثابة مفاجأة للباحث العادي، إذ كيف سيكون من السهل أن يتم الإبلاغ عن نتائج الاختبارات الإحصائية للمقالات والأبحاث التي نشرت في علم النفس وما يتصل بها من التخصصات في السنوات الخمسين الماضية (Kline, 2004)، لكن بالرغم من كثرة الانتقادات التي وجهت لاستخدام الدلالة الاحصائية في البحوث النفسية والتربوية مازالت هي السائدة في كل بحوثنا، وإن كان من الصعب التخلي عن استخدام الدلالة الاحصائية فإنه من الضروري استخدام بدائل كعناصر تعضد الدلالة الإحصائية فيكون اتخاذ القرار بذلك أكثر دقة، وأفضل في الاستدلال بصفة عامة، ومن هذه البدائل يمكن أن نذكر:

أولا: حدود الثقة Confidence Intervals

يقوم الباحثون في بعض الأحيان بحساب حدود أو فترات الثقة بالإضافة إلى اختبار الدلالة الاحصائية لبيانات البحث ، وتمد حدود الثقة الباحث بطريقة لتقدير قيمة الأصل وذلك بناء على البيانات المعروفة من العينة ، وتمثل حدود الثقة أحد فروع الاستدلال الاحصائي حيث يتمكن الباحث من بناء استدلالات من إحصاءات العينة إلى الأصول التي اشتقت منها (السعيد، 2002، ص169).

ثانيا: فترة الثقة Confidence Interval

هي وسيلة أخرى لوصف البيانات، وتوضح فترة الثقة كيف يمكن للمقاييس الوصفية مثل المتوسط والانحراف المعياري والخطأ المعياري للمتوسط أن تقوم بتمثيل المجتمع الحقيقي، تقوم الدراسات بعمل افتراضات حول المجتمع باستخدام عينة من هذا المجتمع، إن قيمة p-value الاحتمالية تحدد ما إذا كانت إحدى المجموعات تختلف عن أخرى وذلك بالاعتماد على متوسط العينة، ولكنها تفشل في إظهار مدى هذا الاختلاف، وإذا قمنا بسحب عينة أخرى من نفس المجتمع فإننا سوف نحصل على متوسط مختلف، لذلك فإن فترة الثقة تعطينا تقديراً لمدى أكبر من القيم التي تمثل المتوسط الحقيقي للمجتمع، إن أكثر فترات الثقة شيوعاً هي فترة الثقة 95% وفترة الثقة 99% وفترة الثقة 90%، ولكن الأولى هي الأكثر استخداماً في الدراسات ، وتعني أنه لو تم دراسة المجتمع فإن متوسط المجتمع سوف يقع ضمن نطاق فترة الثقة هذه في 95% من الحالات وذلك باستخدام بيانات العينة.(أبو عودة، د.س، ص 12)

فإذا كانت اختبارات الدلالة الاحصائية تزودنا فقط بمعلومة حول ما إذا كانت الصدفة هي التفسير أو عدم التفسير للفروق الملاحظة، فإن استعمال حدود الثقة يعتبر بديلا مناسبا لاختبارات الدلالة الاحصائية، والسبب أن كلا الطريقتين تزودان الباحث بنفس النتيجة (رفض أو عدم رفض الفرضية الصفرية)، علما أن حدود الثقة تزود الباحث بمعلومات إضافية لا تقدمها اختبارات الدلالة الاحصائية ، مما يساعد الباحث في تفسير أفضل للنتائج .(بابطين، 2001، ص 88)

ثالثا: قوة الاختبار الاحصائي

تكمن قوة الاختبار الاحصائي في قدرته على رفض الفرضية موضع البحث عندما تكون في واقع الأمر خاطئة، وتعتمد قيمتها (قوة الاختبار) بشكل مباشر على احتمال ارتكاب الخطأ من النوع الثاني، حيث أن قوة الاختبار (1- بيتا)، لذلك فمن المهم أن نتذكر أن قوة الاختبار لا تعني رفض الفرضية الصفرية بشكل مطلق بل تعني في حقيقة الأمر مدى قدرة البيانات المجمعة لدراسة ما ومدى كفاءة تصميمها على رفض فرضية الدراسة. (النجار، 2006، ص266)

رابعا: حجم الاثر

إن حجم الأثر هو مصطلح يستخدم لوصف عائلة من المؤشرات والتي تقيس حجم تأثير المعالجة، ويكون حجم الأثر مختلفا عن اختبارات الدلالة ، وذلك لأن مقاييس حجم الأثر تركز على أهمية النتيجة عمليا وتسمح بإجراء مقارنة بين الدراسات من خلال قدرة الباحثين على الحكم من خلال مستوى الدلالة العملية للنتائج المعروضة.(البارقي، 2012، ص 32)

وحجم الأثر أو التأثير effect size هو مصطلح احصائي يدل على مجموعة كبيرة من المقاييس الاحصائية التي يمكن للباحث في العلوم التربوية والاجتماعية والنفسية والاحصائية استخدامها للتعرف على الاهمية العلمية للنتائج التي اسفرت عنها بحوثه ودراساته ، ويهتم بصفة خاصة ببيان مقدار الاثر الذي تحدثه المتغيرات المستقلة(المعالجات التجريبية) في المتغير أو المتغيرات التابعة التي يقوم عليها تصميم البحث.(باهي، 2010، 1392)

خاتمة

لقد أضحى مستوى الدلالة الاحصائية الذي أفرزته الممارسات الاحصائية منذ أكثر من خمسة عقود من قبل الباحثين تقليدا عشوائيا، وأصبح يشكل حاليا أداة ذات فائدة محدودة ما لم تدعم بالدلالة العملية أو ما يسمى بحجم التأثير ، وتحديد فترات الثقة وحدود الثقة وكذا قوة الاختبار الاحصائي.

وفي ظل غياب التوجيه الكافي من قبل المؤسسات المختلفة كالجمعية الامريكية لعلم النفس وكذا المجلات العلمية التي تقوم بنشر البحوث النفسية والتربوية فإن الإصلاح الإحصائي في علم النفس له مستقبل غير مؤكد، فبالرغم من أن الحجج والانتقادات الموجهة للدلالة الاحصائية مقنعة وجهود الكثيرين متفانية لا يزال من الصعب أن نشعر بالتفاؤل حول مستقبل علم النفس في ظل هذه الممارسات الاحصائية الخاطئة.

قائمة المراجع

-أبو عودة، هشام.(د.س). **تقييم التحليل الإحصائي للمنشورات الدوائية**. كلية الصيدلة. جامعة الملك سعود.

-الرفاعي، أحمد وصبرى ،غنيم نصر محمود. (2005). **التحليل الاحصائي للبيانات باستخدام .spss كلية التربية، جامعة الزقازيق، قسم علم النفس التربوى**. تطوير برامج التعليم العالي النوعي في مصر والوطن العربي في ضوء متطلبات عصر المعرفة.

-المطيري، ريم دانه والديحاني، نوره المطيري. (د.س). **اختبارات الفروض**، إشراف الدكتورة منى توكل. تم استرجاعه في 01.04. 2018 على الرابط https://faculty.mu.edu.sa/download.php?fid=135647

-الصياد، عبد العاطي أحمد. (1985). النماذج الإحصائية في البحث التربوي والنفسي والعربي بين ما هو قائم وما يجب أن يكون. **رسالة الخليج العربي**،16، السنة الخامسة، ص 211- 252.

-البارقي، طلال هيازع حسن. (2012). **واقع الدلالة الإحصائية والدلالة العملية للبحوث المنشورة بمجلة جامعة أم القرى للعلوم التربوية والاجتماعية والإنسانية في المدة 1425- 1430.** رسالة ماجستير غير منشورة. جامعة أم القرى.

-الحداد، سعدة احمد محمد. (2006). **قوة الاختبارات الاحصائية وواقع الدلالة الاحصائية والدلالة العملية في بحوث مجلة الهيئة القومية للبحث العلمي في الجماهيرية العظمى**. رسالة ماجستير غير منشورة. جامعة أم درمان الاسلامية.

-السعيد، رضا مسعد. (2009). **الإحصاء النفسي والتربوي (نماذج وأساليب حديثة)**. ط1. القاهرة: دار الزهراء للنشر والتوزيع .

-النجار، عبدالله بن عمر. (2006). دراسة تحليلية لقوة الاختبار الإحصائي في البحوث الإدارية المنشورة. المجلة العلمية لجامعة الملك فيصل، **العلوم الانسانية والادارية**، مجلد 7، عدد 2، ص 261 - 293.

-النعيمي، ضرغام جاسم و عبدالله، هديل داهي. (2012). دراسة تحليلية لبعض المفاهيم الاحصائية في اختيار حجم العينة ومستوى الدلالة الاحصائية. **مجلة كلية التربية الأساسية، جامعة بابل** ، العدد7، ص 417- 401.

-بابطين، عادل أحمد. (2001). **مشكلات الدلالة الإحصائية في البحث التربوي وحلول بديلة.** رسالة ماجستير غير منشورة. كلية التربية. جامعة أم القرى.

-باهي، مصطفى حسين. (2010). **العلاقة بين الدلالة الاحصائية وحجم التأثير في البحوث التربوية والنفسية**. القاهرة. المؤتمر العلمي السادس عشر لإعداد المعلم وتنميته.

-تيغزة، امحمد بوزيان. (د.س). **التهافت على استعمال ما يدعى بصدق الاتساق الداخلي (Validity of internal consistency)ارتباط الفقرات ببعدها وارتباط الأبعاد بالاختبار ككل ومواطن قصورها**. جامعة وهران 2، الجزائر. تم استرجاعه في 01.04. 2018 على الرابط

https://www.facebook.com/mhamed.tighezza.7/posts/549680868558198

- تيغزة، امحمد.(2017). توجهات حديثة في تقدير صدق وثبات درجات أدوات القياس: تحليل نظري تقويمي وتطبيقي. **مجلة العلوم النفسية والتربوية**، 4 (1)، 7-29.

-حسن، محمد عزت عبد الحميد. (2016). الدلالة الإحصائية والدلالة العملية في البحوث. **مجلة كلية التربية النوعية**، العدد1، المجلد 1، ص 1-19.

-سلامة، حسن علي حسن. (2004). الدلالة الإحصائية والدلالة العلمية في البحوث التربوية. **المجلة التربوية، مصر**، العدد 20 ، ص 3 - 14.

- نوتزو، ريجينا. (2014). أخطاء إحصائية، قِيَم المؤشر «بي P-values، المقياس الذهبي» للصحة الإحصائية، ليست جديرة بالثقة كما يَفترض الكثير من العلماء. **Nature**، doi:10.1038/506150a Published online 31 Mar

Bakan, D. (1966). The test of significance in psychological research. *Psychological bulletin, 66*(6), 423.

Boster, F. J. (2002). On making progress in communication science. *Human Communication Research, 28*(4), 473-490.

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

Cohen, J. (1990). Things I have learned (so far). *American psychologist, 45*(12), 1304.

Cohen, J. (1995). The earth is round (p<. 05): Rejoinder.

Coolidge, F. L. (2012). *Statistics: A gentle introduction*: Sage Publications.

Daniel, L. G. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *Research in the Schools, 5*(2).

Dar, R. (1987). Another look at Meehl, Lakatos, and the scientific practices of psychologists.

Denis, D. J. (2003). Alternatives to null hypothesis significance testing. *Theory & Science, 4*(1), 21.

Gigerenzer, & murray, D. (1987). Cognition as Intu-itive Statistics: Erlbaum, Hillsdale, NJ.

Haig, B. D. (2017). Tests of statistical significance made sound. *Educational and Psychological Measurement, 77*(3), 489-506.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (2013). *What if there were no significance tests?* : Psychology Press.

Hubbard, R. (2015). *Corrupt research: The case for reconceptualizing empirical management and social science*: Sage Publications.

Hubbard, R., & Lindsay, R. M. (2008). Why P values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology, 18*(1), 69-88.

Johnson, D. H. (1999). The insignificance of statistical significance testing. *The journal of wildlife management*, 763-772.

Kline, R. B. (2004). Beyond significance testing: Reforming data analysis methods in behavioral research. *APA Books*.

Levine, T. R., Weber, R., Park, H. S., & Hullett, C. R. (2008). A communication researchers' guide to null hypothesis significance testing and alternatives. *Human Communication Research, 34*(2), 188-209.

McLEAN, J. E., & Ernest, J. M. (1998). The role of statistical significance testing in educational research. *Research in the Schools, 5*(2).

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of science, 34*(2), 103-115.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of consulting and clinical Psychology, 46*(4), 806.

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

Morrison, D. E., & Henkel, R. E. (2006). *The significance test controversy: A reader*: Transaction Publishers.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological methods, 5*(2), 241.

Reese, R. A. (2004). Does significance matter? *Significance, 1*(1), 39-40.

Robinson, D. H., & Wainer, H. (2001). ON THE PAST AND FUTURE OF NULL HYPOTHESIS SIGNIFICANCE TESTING 1. *ETS Research Report Series, 2001*(2), i-20.

Royall, R. M. (1986). The effect of sample size on the meaning of significance tests. *The American Statistician, 40*(4), 313-315.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician, 70*(2), 129-133.

William, L. T. (2000). Null hypothesis testing: problems, prevalence, and an alternative. *J. Wildl. Manage, 64*(4), 912-923.

Ziliak, S., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*: University of Michigan Press.