الاستخــدامات الخاطئــة للدلالــة الإحصائيــة وحلــول مقترحــة

الباحث: إبراهيم بن سعيد بن حميد الوهيبي
الجامعة الإسلامية العالمية الماليزية
ishalwahaibi@gmail.com

أ.د. داود بن عبد الملك الحدابي
الجامعة الإسلامية العالمية الماليزية
dawood@iium.edu.my

أ.م.د.حسين بن علي بن طالب الخروصي
جامعة السلطان قابوس
hussein393500@gmail.com

الملخص

هدفت الدراسة الحالية إلى توضيح جوانب القصور المختلفة التي تعاني منها الدلالة الإحصائية في تفسير نتائج الدراسات والبحوث، وتحديد سوء استخدام الدلالة الإحصائية في تفسير النتائج واقتراح بعض الحلول النظرية لتطوير استخدام الدلالة الإحصائية، أظهرت النتائج أن أهم جوانب القصور في الدلالة الإحصائية هو تأثرها بحجم العينة والطبيعة الثنائية لنتائج الدلالة الإحصائية، وكذلك منطقية الفرضية الصفرية، وثّقت الدراسة أيضًا الاستخدامات الخاطئة للدلالة الإحصائية في الدراسات والبحوث والتي من أهمها استخدام نتائج الدلالة الإحصائية كأساس لشرح أهمية النتائج، والإشارة إلى الأهمية العملية للنتائج، والتحقق من صحة الفرضية الصفرية، واقترحت الدراسة مجموعة من الحلول للاستخدامات الصحيحة للدلالة الإحصائية مثل عدم الاعتماد على النتائج في تقييم البحوث والدراسات، والصياغة المناسبة لأسئلة البحث وفرضياته، وذلك باستخدام مؤشرات ذات أهمية عملية وقدرة إحصائية باستخدام تحليل المنفعة العملية أو الدلالة الإكلينيكية.

الكلمات المفتاحية: الدلالة الإحصائية، سوء تفسير (استخدام) الدلالة الإحصائية، قصور الدلالة الإحصائية، تطوير الدلالة الإحصائية

The misuses of the statistical significance and suggested solutions

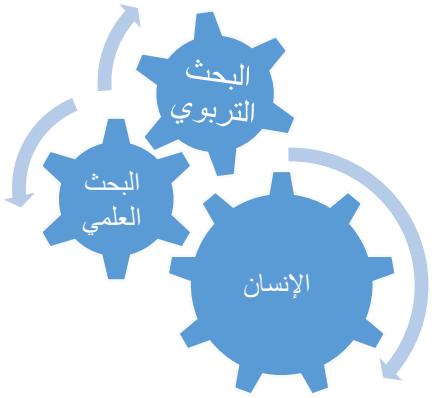
Abstract:

The present study aimed at clarifying the various shortcomings of the statistical significance in interpreting the results of studies and research,

identifying the misuses of the statistical significance in the interpretation of the results, and proposing some theoretical solutions to enhance the use of statistical significance. Results showed that the most important shortcoming in the statistical significance are related to the effect of the sample size on the statistical significance, the binary status of the results of the statistical significance, and the logic of the null hypothesis. The study has also documented some of the misuses of the statistical significance in research such as the use of statistically significant results as a basis to explain the importance of the results, indicate the practical significance of the results, and validate the null hypothesis. The study proposed a range of solutions for the correct uses of the statistical significance such as not to rely on the results in the evaluation of research and studies, appropriate phrasing of the research questions and hypotheses, using indicators of practical significance and statistical power, and using analysis of practical benefit or clinical significance.

Keywords: statistical significance, misuses of the statistical significance, shortcomings of the statistical significance, improvement of statistical significance

إذا كان البحث العلمي في المجالات التطبيقية يحظى باهتمام كبير لا يخفى على ذي فكر وبصيرة؛ فإن البحث العلمي في المجالات التربوية أو ما يسمى بالبحث التربوي يعتبر جزءا مهما من منظومة البحث العلمي فهو النواة المكونة للبحث العلمي، نظرا لكونه يهتم ببناء العقول البشرية، وتنمية الإنسان الذي هو مؤسس العلوم التطبيقية وصانعها، ويصف الباحثون العلاقة بين الإنسان والبحث التربوي والبحث العلمي بأنها علاقة تبادلية دائرية؛ فالبحث التربوي يستمد طرائقه وأساليبه من البحث العلمي، ويطبقها في بناء شخصية الإنسان وإيجاد الحلول لمشكلاته، ليقوم الإنسان بالبحث العلمي في شتى ميادين الحياة، والشكل (1) يوضح هذه العلاقة.



الشكل(1): العلاقة بين الإنسان والبحث التربوي والبحث العلمي

للبحث التربوي أهمية كبيرة نظرا لأهمية المجال الذي يبحث فيه وهو المجال التربوي، وكذلك لأهمية إيجاد الحلول العلمية السليمة لمشكلات المجتمع التربوية، والتي لا تقل أهمية عن مشكلاته السياسية والاقتصادية والاجتماعية (قسيس، وآخرون، 2008).

والبحث التربوي كما عرفته الجمعية الأمريكية للبحوث التربوية American Educational Research Association (AERA,2018) هو "مجال الدراسة العلمي الذي يدرس عمليات التعليم والتعلم والصفات الإنسانية والتفاعلات والمنظمات والمؤسسات التي تشكل النتائج التعليمية"، ويعرفه حمداوي(2014، ص9) أيضا بأنه "ذلك البحث الذي يدرس الظواهر التربوية والقضايا الديداكتيكية، وكل ما يرتبط بها من مواضيع نفسية، واجتماعية، وفلسفية، وسياسية، واقتصادية، وإدارية، ولسانية، وتاريخية، وبيولوجية، ..."، ومن خلال هذين التعريفين يلاحظ الباحثون بأن محور البحث التربوي يتمركز حول الإنسان وسلوكياته المختلفة، ويعرفونه بذلك العلم الذي يبحث في تأثير ممارسات الإنسان وأفعاله التربوية في مختلف الجوانب الحياتية والعلمية والثقافية والاجتماعية، وتطويرها للأفضل، ووضع الحلول والبدائل لكل ما يواجهها من عقبات ومشكلات.

والبحث التربوي له أهمية عظمى في صنع السياسة التعليمية والتربوية وتوجيهها الوجهة الصحيحة في مختلف المراحل التعليمية، كما أنه يسهم بدرجة كبيرة في حل كل ما يعوقها من مشكلات وصعوبات؛ لذلك لا بد من أن تتميز نتائج الأبحاث التربوية بالدقة والاتقان، حيث أن أي خلل في هذه النتائج سوف يؤثر بشكل أو بآخر في بناء القرارات التربوية الصحيحة أو معالجة المشكلات التربوية.

وحتى تكون البحوث التربوية ذات جودة لا بد من صياغة الأسئلة البحثية في صيغة فروض علمية تساعد على التفكير بعمق في نتائج الدراسة المتوقعة، وتوضح المقصود من الأسئلة البحثية(عباس، 2013)، ويتم ترجمة الفروض العلمية إلى فرضيات إحصائية بلغة قياسات المجتمع (التل وأبو زينة والبطش، 2007)، نظرا لعدم القدرة على قياس الفرضيات العلمية بصفة مباشرة، حيث أنها لا تحدد مقدار العلاقة بين المتغيرات أو الأثر، وإنما تتوقعهما فقط أو تتنبأ بهما (الدليمي وصالح، 2014)، وتمثل الفرضيات الإحصائية الصيغة الرياضية للفرضيات العلمية، وهي "عبارة عن جملة أو عدد من الجمل تعد باستخدام بعض النماذج الإحصائية ذات العلاقة ببعض خصائص مجتمع البحث، والتي تستخدم من أجل تأكيد العلاقات أو السببية أو الارتباط بين المتغيرات" (خضر، 2013)، وتنقسم إلى قسمين:

الفرضية الصفرية (Null Hypothesis): تسمى أيضا فرضية العدم، وهي فرضية التساوي، أي أن متوسط العينة (μ_0) يساوي متوسط المجتمع (μ) التي أخذت منه(السيد،2009)، أي أن ($\mu_0 = \mu$) أو ($0=\mu - \mu_0$)، وغالباً ما يكون الفرضية الصفرية عكس ما يعتقده الباحثون فعلاً (Lane,2013)، ومن الأمثلة على الفرضية الصفرية: "لا توجد فروق ذات دلالة إحصائية عند مستوى دلالة أقل من (0.05) بين متوسطات طلبة كليتي العلوم والآداب في الاتجاه نحو الرياضيات"، أو "لا توجد علاقة ذات دلالة إحصائية عند مستوى دلالة أقل من (0.05) بين قلق الامتحان والدافعية نحو التعلم لدى طلبة الجامعة الإسلامية العالمية "، ومما ينبغي التنبه له بأن مصطلح "الصفر أو العدم" في الفرضية الصفرية لا يعني دائما أن الفرق بين متوسط العينة (الإحصاءة) ومتوسط المجتمع (المعلمة) أو الفرق بين متوسطي المتغيرين يساوي صفرا، فقد يكون الفرق يساوي رقم أخر مثل "600" (Lane,2013)، ويرمز للفرضية الصفرية بالرمز (H_0). أي أن ($H_0: \mu_0 = \mu$) أو ($H_0: \mu - \mu_0 = 0$). أو ($H_0: \mu_1=\mu_2=...=\mu_k$)، حيث k تشير إلى عدد المجموعات، الجدير بالذكر بأن الفرضية الصفرية اقترحها الإحصاء البريطاني الشهير فيشر كما ذكر ذلك الدليمي وصالح(2014، ص356).

الفرضية البديلة (Alternative Hypothesis) : هي فرضية عدم التساوي التي توضح اختلاف متوسط العينة عن متوسط المجتمع، أي إن متوسط العينة (μ_0) لا يساوي متوسط المجتمع (μ) (السيد،2009)، أي أن ($\mu_0 \neq \mu$)، ومن الأخطاء الشائعة عند الباحثين في الفرضية البديلة عند تعدد المجموعات أنهم يعتقدون أن تكون جميع المتوسطات غير متساوية بمعنى $\mu_1 \neq \mu_2 \neq ... \neq \mu_k$، لكن الصحيح أنه لا يلزم عدم التساوي بين كل المتوسطات، وإنما يمكن القول بأنه ليس كل شئ متساوي، فقد يكون الفرضية البديلة تساوي .(STAT502,2018) $\mu_1=\mu_2 \neq ... \neq \mu_k$

والفرضية البديلة إما أن يكون فرضية غير موجهة Non-Directional Hypothesis بمعني الاعتقاد بوجود علاقة أو وجود فروق لكن بدون تحديد نوع هذه العلاقة إيجابية أم سلبية، ولا تحديد الفروق لصالح أحد المتوسطين، ومن أمثلتها: "توجد علاقة بين استخدام نظام البصمة الإلكتروني في العمل والرضا الوظيفي" ، أو "طريقة تدريس الطلاب خارج الغرفة الصفية تؤثر على تحصيلهم الدراسي"، بدون تحديد نوع هذا التأثير هل هو إيجابي أو سلبي.

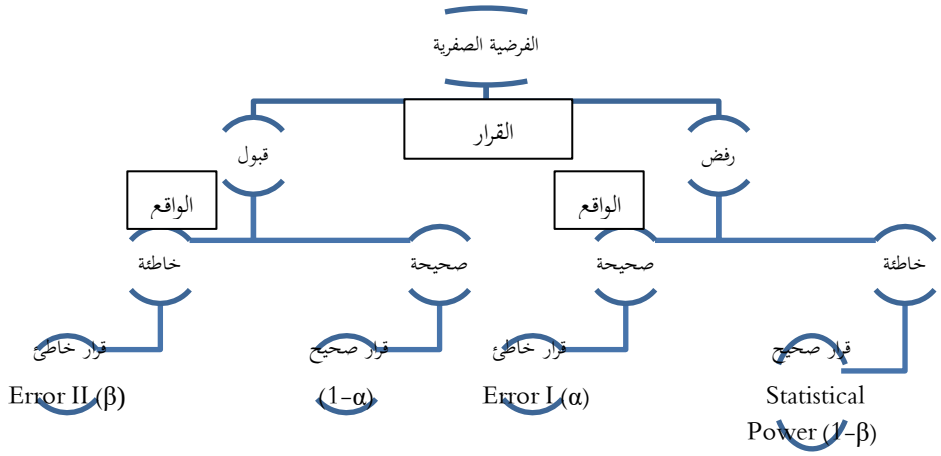
وقد يكون الفرضية البديلة فرضية موجهة Directional Hypothesis وتشمل مستويين فإما أن تكون فرضية موجهة إيجابية، كالقول بوجود علاقة إيجابية بين الدافعية نحو التعلم والتحصيل الدراسي، بمعنى أن الدافعية نحو التعلم تزيد من التحصيل الدراسي، وإما أن تكون فرضية موجهة سلبية كمثل القول بأن طريقة التدريس المستخدمة أدت إلى عزوف الطلاب عن المادة وبالتالي انخفضت اتجاهاتهم نحو المادة، ويرمز للفرض البديل بالرمز (H_1).

وبناء على ما تم توضيحه من حالات الفرضية البديلة؛ فإن الفرضية العلمية أو الفرضية البحثية هو إحدى هذه الحالات الثلاث للفرضية البديلة، وعليه فإن صور الفرضية البديلة هي:

($H_{1:}\ \mu_0 \neq \mu$) أو ($H_1: \mu > \mu_0$) أو ($H_1: \mu < \mu_0$).

ومن خلال خبرة الباحثين في التحليل الاحصائي لدراسات الماجستير والدكتوراه بمركز الخوارزمي للدراسات والاستشارات التربوية، ومن خلال اطلاعهم على العديد من الدراسات التربوية؛ فإنهم يلاحظون أن أغلب الباحثين –إن لم يكن كلهم– يستخدمون الفرضية الصفرية (H_0) في دراساتهم، ويبنون عليها اختباراتهم الإحصائية، ولا يتطرقون إلى الفرضيات البديلة، على الرغم من أن الإطار النظري للدراسة والدراسات السابقة المرتبطة بها هي التي تحدد ما إذا كان للباحث أن يستخدم فرضيات صفرية أم فرضيات بديلة في دراسته، ويصف الدليمي وصالح (2014، ص363) هذا الأمر بأنه موضة خاطئة شاعت في البحوث النفسية والتربوية والاجتماعية، ونادا بضرورة التوقف عنها، كما نبه على ذلك أيضا د. أحمد إبراهيم خضر(2013) في صفحته الإلكترونية حيث قال: "لجأ بعض الباحثين إلى الفروض غير الموجهة – ومنها الفرضيات الصفرية – كحيلة هروبية يتخلصون بها من الجهد المعرفي اللازم لبناء إطار نظري سليم للبحث".

وعندما يقوم الباحثون باختبار فرضيات دراسته؛ فإن القرار الذي يتم التوصل إليه حول رفض أو قبول الفرضية الصفرية يتخذ إحدى الحالات الأربعة التي يوضحها الشكل (2) الآتي:



* من تصميم الباحثين

الشكل (2): حالات اتخاذ القرار الإحصائي بناء على الفرضية الصفرية(H_0)

ومن خلال الشكل (2) يمكن تلخيص حالات اتخاذ القرار الإحصائي بناء على الفرضية الصفرية(H_0) في الآتي:

- قرار خاطئ: عندما يتم رفض الفرضية الصفرية(H_0) وهي في الواقع صحيحة (كالقول بأن هناك فروق مع أن الواقع يدل على عدم وجود فروق) (التل وأبو زينة والبطش، 2007)، ويسمى هذا القرار الخاطئ بالخطأ من النوع الأول Error I، كما يسمى ب"مستوى الدلالة Significance Level" ويرمز له بالرمز (α).

- قرار خاطئ: عند الفشل في رفض الفرضية الصفرية(H_0) وهي في الواقع خاطئة (عند الفشل في إيجاد فروق مع أن الواقع يدل على وجود فروق)، ويسمى بالخطأ من النوع الثاني Error II، كما يسمى بـ "مميز الفاعلية Characteristic Operating" (السواح،2007)، ويرمز له بالرمز (β)

- قرار صحيح: عندما يتم رفض الفرضية الصفرية(H_0) وهي في الواقع خاطئة، (كمثل القول بوجود فروق والواقع يدل على ذلك)، ويسمى ب "القوة الاحصائية للاختبار Statistical Power" ويرمز له (1-β)

- قرار صحيح: عند الفشل في رفض الفرضية الصفرية(H_0) وهي في الواقع صحيحة، (كالقول بعدم وجود فروق بين المجموعات والواقع يدل على ذلك)، ويسمى ب "درجة الثقة confidence Level"، ويرمز له $(1-\alpha)$.

وأشار الثبيتي(2008) إلى قلة الثقة لدى متخذو القرار التربوي في مصداقية نتائج الدراسات التربوية، مما أدى إلى وجود فجوة بين كثرة الدراسات والأبحاث العلمية وقلة تطبيق نتائجها في الميدان التربوي، ومن خلال تتبع الأسباب التي أدت إلى هذا الأمر؛ نجد أن الكثير من الباحثين لا يدركون المفاهيم الصحيحة للأساليب الإحصائية التي يستخدمونها في دراساتهم والتي يبنون عليها قراراتهم في تفسير نتائج هذه الدراسات(باهي،2010)، حيث أنهم يكتفون بنتائج الدلالة الإحصائية فقط في اختبار فرضياتهم البحثية، ولا يدركون أن إيجاد الدلالة الإحصائية ما هي إلا خطوة أولى في تحليل النتائج، وتتبعها خطوات أخرى أكثر أهمية من الدلالة الإحصائية، ويجب أن ترتكز عليها القرارات العليمة في تفسير النتائج(سلامة،2004).

ولقد أساء الباحثون في استخدام مفهوم الدلالة الإحصائية، في تفسير النتائج، حيث ذكر حسن(2008) من خلال اطلاعه على (Nickerson,2000 و Thompson,1996) جملة من الأخطاء التي يقع فيها الباحثون بسبب سوء فهمهم لهذا المفهوم الإحصائي، ومن بين هذه الأخطاء ما يلي:

- الخلط بين الدلالة الإحصائية للنتائج ودلالتها التربوية والعملية، ومقدار تأثير المتغير المستقل في المتغير التابع (نصر الله، 2016).
- الاعتقاد بأن الدلالة الإحصائية للنتائج تجعها ذات أهمية، وأن عدم الدلالة تدل على عكس ذلك(Gagnier & Morgenstern,2017).
- وجود دلالة إحصائية لنتائج ما لا يعنى تكراراها مرة أخرى في حالة تطبيق الدراسة على عينات أخرى من نفس المجتمع.
- عدم إدراك الباحثون أن الدلالة الإحصائية مرتبطة بحجم العينة، فقد تكون النتائج ذات دلالة إحصائية نظرا لكبر حجم العينة، وليس بسبب قوة تأثير المعالجات على المتغيرات التابعة(Aron & Aron,1994; نصار،2006؛ بابطين، 2002؛ عباس،2013 ؛ Nix & Barnette,1998; Huck,2009; Gunn,2019).

لذلك عندما ظهرت الاستخدامات الخاطئة للدلالة الإحصائية افترق الباحثون إلى فريقين، فريق عارض استخدام الدلالة الإحصائية ونادى بالابتعاد عنها نظرا لضعفها وقصورها، وفريق تمسك بها وعزا ما تعانيه من قصور إلى سوء استخدام الباحثون لها وليس قصور في ذاتها، فمن الدراسات المؤيدة للدلالة الإحصائية دراسة هيج Haig(2017) والتي تناولت طبيعة ومكان اختبارات الدلالة الإحصائية في العلوم بشكل عام، في دراسات علم النفس بشكل خاص، حيث نبعت المشكلة الرئيسة لهذه الدراسة من القبول الواسع وغير الناقد لاختبارات الفرضية الصفرية (NHST)، والذي يعتبر مزيجًا لا يمكن الدفاع عنه للأفكار المقتبسة من تفكير العالم فيشر Fisher حولها، ومن الطريقة البديلة لكل من نيمان وبيرسون Neyman and Pearson، وسعت الدراسة إلى تصحيح أوجه القصور في هاتين الطريقتين، وتوصلت الدراسة إلى ضرورة أن يستفيد علم النفس من منظورين مهمين وأكثر حداثة في اختبارات الدلالة الإحصائية، وهما المنظور الجديد لمؤيدوا العالم فيشرFisher ، ومنظور الخطأ الإحصائي، يسعى المنظور الجديد لمؤيدوا العالم فيشرFisher إلى تحسين الطريقة الأصلية لفيشرFisher ، ويرفض العناصر الرئيسية للطريقة البديلة لـ نيمان وبيرسون، وفي المقابل فإن منظور الخطأ الإحصائي يعتمد على نقاط القوة لكلا الطرق الإحصائية، ويقترح هيج Haig أن تكون هذه التوقعات الأحدث على اختبارات الدلالة الإحصائية بمثابة تحسن واضح في اختبارات الفرضية الصفرية NHST ، لا سيما الخطأ الإحصائي، كما تدعوا هذه الدراسة إلى أن تلعب اختبارات الدلالة الإحصائية دورًا مفيدًا، وإن كان محدودًا في البحوث النفسية.

كما قدم كل من هوبارد وأرمسترونج Hubbard & Armstrong(2006) دراسة بعنوان "لماذا لا نعرف حقيقة معنى الدلالة الإحصائية؟"، أوضحا فيها بأنه يوجد خلط في الكتب العلمية والدراسات والبحوث بين مفهومين لهما دلالة إحصائية وهما قيم p ومستويات α بشكل عام، على الرغم من أن لكل منهما تفسيرات مختلفة تمامًا، مما أدى إلى سوء فهم لمعنى الدلالة الإحصائية، حيث أن العديد من الباحثين لديهم معرفة غير مؤكدة بما تعنيه الدلالة أو الأهمية الإحصائية حقًا، هل هي قيم p أو مستويات α أو المعيار p <α أو كل ما سبق؟، وأوضحت هذه الدراسة أن سبب نشوء هذا الالتباس هو المزج الخاطئ لمدرستين من الاختبارات الإحصائية الكلاسيكية، وهما مدرستي Fisher و Neyman and Pearson، ولكل منهما مقاييس مختلفة ذات دلالة إحصائية، وقدمت عدة اقتراحات للمعلمين والباحثين حول كيفية التغلب عليه، وقدمت الدراسة توصيتين هما أولاً: إذا كان اختبار الدلالة الإحصائية يجب إدراجه في المنهج، فإن الاختلافات بين نموذجي Fisher و N–P تتطلب شرحًا واسعا ومبسطا، ويجب على الطلاب أن يكونوا على دراية بما يعنيه الدلالة الإحصائية بالضبط، ولا يعتمدوا على البرامج الإحصائية الحاسوبية للإبلاغ عن

مجموعة الاختبارات الإحصائية دون أن فهم معناها، ثانياً: وهو الأهم يجب أن ندرس توفير فواصل الثقة حول إحصائيات العينة ومؤشرات الدلالة العملية، ودراسة ما إذا كانت فواصل الثقة ذات الصلة تتداخل عبر دراسات مماثلة في تكرار منهجي.

ومن الدراسات النظرية دراسة ليتل (2001)Little والتي هدفت إلى عرض نظرة تاريخية عن مفهوم الدلالة الإحصائية، وأشارت إلى أن مفهوم الدلالة الإحصائية يعتبر من أكثر المفاهيم في العلوم الإجتماعية الذي حظي باهتمام كبير وواسع، فلم يكن هناك مفهوم آخر حظي بنفس هذا الاهتمام، وأوضحت إلى أنه يجب أن يشمل محو الأمية العلمية الإدراك بأن ممارسة إضفاء الشرعية على النتائج الإحصائية عن طريق الدلالة أو الأهمية الإحصائية، والتي يتم توحيدها حاليًا عند مستوى (0.05)، ليست نتيجة منطق أساسي للرياضيات ولكن تم بناؤه وإعادة بنائه اجتماعيًا استجابة للظروف الاجتماعية التاريخية.

أما الدراسات التي انتقدت الدلالة الإحصائية فكثيرة، ومن ضمنها دراسة سلامة(2004) التي أوضح فيها بأن استخدام الدلالة الإحصائية ما هو إلا مرحلة أولية في تحليل نتائج البحوث يجب أن تتبعها خطوات أخرى، وانتقد الباحثين الذين يعتمدون على الدلالة الإحصائية فقط، وبين بأن مجرد استخدام الدلالة الإحصائية يعد ديكورا رقميا أكثر من كونه ضرورة بحثية، ثم تطرق إلى الدلالة العملية وذكر بأن مفهومها ليس جديدا في البحث العلمي، إلا أن استخدامها في البحوث التربوية قليلة، وشرح مؤشرات الدلالة العملية لبعض الاختبارات الإحصائية.

كما أن هناك دراسات بحثت عن المشكلات التي تواجه الدلالة الإحصائية، واقترحت الحلول البديلة لهذه المشكلات، ومن ضمن تلك الدراسات دراسة بابطين(2002) بحثت في مشكلات الدلالة الإحصائية، والكشف عن واقع تلك المشكلات في الرسائل العلمية في جامعة أم القرى، وتوصلت الدراسة إلى مجموعة من النتائج، منها: أن أهم مشكلات الدلالة الإحصائية هي استخدام نتائج الدلالة الإحصائية كتفسير لأهمية النتائج، وكتفسير لتأثير المعالجة، واستخدام قيم (P) لتقدير حجم التأثير، وكتفسير لاحتمالية الفرضية الصفرية، إضافة إلى التحفير في النتائج الدالة إحصائيا، والطبيعة الثنائية (دال، غير دال) لنتائج الدلالة الإحصائية، كما توصلت الدراسة إلى أهم المفاهيم التي يمكن أن تقدم حلولا لمشكلات الدلالة الإحصائية، والتي من ضمنها استخدام الدلالة العملية، والقوة الإحصائية، وتقدير حدود الثقة، وتحليل الإعادة، وأوصت الدراسة بالتركيز في مقررات الإحصاء الاستدلالي على فلسفة الأساليب الإحصائية، والاهتمام بأساليب تقدير المؤشرات مقارنة بأساليب اختبار الفرضيات.

مشكلة الدراسة:

نظرا للانتشار الواسع للاستخدامات الخاطئة للدلالة الإحصائية في تفسير النتائج، حيث يذكر (cumming,2012,p25) بأن Michael Oakes أجرى دراسة في ثمانينات القرن الماضي (1986) حول مدى فهم قيم P–value والاستخدام الصحيح لها في تفسير النتائج، وتكونت أداة دراسته من (6) أسئلة، وخلص الدراسة إلى أن عدد الأشخاص الذين أجابوا إجابة صحيحة على جميع الأسئلة الستة هم (3) فقط من إجمالي العينة (70) بنسبة (4.3%) فقط، وعلى الرغم من قِدم هذه التفسيرات الخاطئة للدلالة الإحصائية إلا أنها لا تزال تراوح مكانها، ولم تختفي من أوساط الأكاديميين، فبعد ما يقرب من ثلاثة عقود من دراسة Michael Oakes؛ قام بزينا وسوندرس Bezzina & Saunders(2014) بتطبيق دراسة هدفت إلى البحث إلى مدى انتشار المفاهيم الخاطئة الإحصائية بين الأكاديميين وخاصة في البحوث التجارية، واشتملت أداة الدراسة على (30) فكرة خاطئة في تفسير الدلالة الإحصائية، تم توزيعها على عينة الدراسة المكونة من (166) أكاديميا، استجاب منهم (80) شخصا على كامل الاستبانة، وخلصت النتائج إلى أن التفسيرات الخاطئة للدلالة الإحصائية لم يقتصر على الطلبة فقط، وإنما يتعدى ذلك إلى الأكاديميين أيضا.

وحيث أن دراسة كل من (Michael Oakes,1986 المذكور في cumming,2012) و(Bezzina & Saunders,2014) تم تطبيقيهما على عينة أجنبية، قام الباحثون بتطبيق دراسة استكشافيه حول مفهوم الدلالة الإحصائية، ومدى انتشار استخداماتها الخاطئة في تفسير نتائج الدراسات، لمعرفة مدى انتشار مثل هذه المفاهيم الخاطئة لاستخدام الدلالة الإحصائية في تفسير نتائج الدراسات والبحوث في الأوساط العربية؛ وتكونت أداة الدراسة من (10) فقرات فقط، وتم تطبيق الدراسة الاستكشافية على عينة قصدية بلغت (60) شخصا من التربويين حملة الدكتوراه والماجستير داخل سلطنة عمان وخارجها، وخلصت نتائجها أن المتوسط الحسابي للإجابات الصحيحة بلغ (41.8%) فقط، ولم يحصل أي شخص من أفراد العينة على إجابات صحيحة بنسبة (100%) في الاستبانة، وحصل شخص واحد فقط على متوسط حسابي بنسبة (90%)، كما أن (61.67%) من أفراد العينة لم يعرفوا الإجابة على سؤال واحد على الأقل من أسئلة هذه الدراسة، مما يؤكد على أن أفراد العينة لديهم سوء فهم حول الدلالة الإحصائية في تفسير نتائج الدراسات والبحوث.

لذا فإن هذه الدراسة تسعى إلى تقصي الاستخدامات الخاطئة للدلالة الإحصائية واستعراض جوانب القصور فيها، واقتراح الحلول المناسبة للاستخدام الصحيح للدلالة الإحصائية في تفسير نتائج الدراسات والبحوث، وتحديدا فإن الدراسة تسعى إلى الإجابة على الأسئلة الآتية:

1- ما جوانب القصور في الدلالة الإحصائية في تفسير نتائج الدراسات العلمية؟

2- ما الاستخدامات الخاطئة للدلالة الإحصائية؟

3- ما الحلول المقترحة لتطوير الاستخدام الصحيح للدلالة الإحصائية؟

أهداف الدراسة:

سوف تسعى الدراسة الحالية إلى تحقيق الأهداف الآتية:

1. توضيح جوانب القصور التي تعاني منها الدلالة الإحصائية في تفسير نتائج الدراسات والبحوث.

2. معرفة الاستخدامات الخاطئة للدلالة الإحصائية في تفسير النتائج.

3. اقتراح بعض الحلول النظرية لتطوير استخدام الدلالة الإحصائية.

أهمية الدراسة:

تستمد هذه الدراسة أهميتها من أهمية تبصير الباحثين والمهتمين بالإحصاء بالاستخدامات الخاطئة للدلالة الإحصائية، وجوانب قصورها في تفسير نتائج الدراسات والبحوث، وذلك لتجنبها في تحليل بيانات دراساتهم، كما أن الحلول المقترحة في هذه الدراسة قد يكون لها تأثير كبير في تصحيح المسار لاستخدام الدلالة الإحصائية في تفسير نتائج الدراسات.

نتائج الدراسة:

الإجابة على السؤال الأول:

ما جوانب القصور في الدلالة الإحصائية في تفسير نتائج الدراسات العلمية؟

قام فيشر عام 1922م، بتطوير منهجية محددة لاختبار بيانات البحوث وتقييم نتائجها، وتمثلت هذه المنهجية في اختبار فرضية واحدة صفرية Nill hypothesis ثنائية باستخدام القيمة p كقوة للإحصاء، ولم

يقم بتطوير أو دعم الفرضيات البديلة، والأخطاء من النوع الأول (α) ومن النوع الثاني (β) في اختبارات الدلالة الإحصائية، أو مفهوم القوة الإحصائية، وحاول معاصران رياضيان للعالم فيشر، وهما جيرزي نيمان وإيجون شارب بيرسون Neyman–Pearson تحسين منهجية فيشر Fisher، إلا أنهم انتهى بهما المطاف في تطوير نظرية جديدة، والتي تتمحور حول اختبار فرضيات إحصائية متعددة، وهي الفرضية الأساسية والفرضيات البديلة، مما أدى إلى ظهور خلاف بين فيشر Fisher، وكل من Neyman–Pearson، ونظرا لكون كلتا النظريتين بينهما أوجه تشابه كافية يمكن الخلط بينها بسهولة، ظهرت عام 1940م بواسطة Lindquist منهجية جديدة تمثلت في اختبارات الدلالة الإحصائية (NHST)، هي الأكثر شيوعًا المستخدمة لاختبار البيانات في الوقت الحالي، وهي عبارة عن دمج لنظريتي Fisher و–Neyman Pearson، وعلى الرغم من أن الهدف الأساسي من اقتراح Lindquist لاختبارات الدلالة الإحصائية (NHST)، هو محاولة الاستفادة من نظريتي Fisher و Neyman–Pearson، والتخلص من جوانب الخلاف بينهما، إلا أن هذا الدمج غير موفق ويعتبر من العلوم الزائفة، التي تسيء للعلم أكثر من إفادته، كما أن هذا الدمج ينفي بشكل فعال الفوائد التي يمكن جنيها من نظريتي Fisher و Neyman–Pearson؛ كما أنه يبطئ التقدم العلمي(carver,1987)، وعلى ذلك أثيرت انتقادات عديدة ضده على مدار ما يقارب من قرن من الزمن (Gigerenzer, 2004; Perezgonzalez, 2014, 2015)، ومن ضمن أسباب نقد اختبارات الدلالة الإحصائية هو أنها تعاني من بعض القصور لتأثرها بعدة عوامل تؤدي إلى حدوث سوء فهم في تفسير نتائجها، ومن أهم تلك العوامل ما يلي:

1) تأثر الدلالة الإحصائية بحجم العينة:

أثبت التدراسات والتجارب بأن النتائج الدالة إحصائيا ليست دائما تدل على وجود فروق حقيقية بين المتغيرات، وإنما قد تكون بسبب حجم العينة الكبير(Gunn,2019 ؛ عباس،Huck,2009;2013)، فزيادة حجم العينة يؤدي إلى تقليل قيمة p value تدريجيا إلى أن تساوي الصفر، حتى مع عدم (أو قلة) وجود أثر حقيقي للمتغيرات المستقلة، كما أن أهمية اختبارات الدلالة الإحصائية تقل مع زيادة حجم العينة، في تجارب العينة الكبيرة، خاصة تلك التي تتضمن متغيرات متعددة، ويتضاءل دور اختبارات الدلالة الإحصائية لأنه حتى الفروق الصغيرة غير ذات المغزى غالباً ما تكون ذات دلالة إحصائية (Nix & Barnette,1998)، ولقد قام كل من نصار(2006) و Kellow(1998) الموضح في دراسة بابطين(2002) بتجربة أمثلة افتراضية توضح انخفاض مستوى الدلالة الإحصائية عند مستوى دلالة أقل من (0.05) كلما زاد حجم العينة، مع تثبيت حجم الأثر، وفي كلا الدراستين تم استخدام تحليل التباين الأحادي

One Way Anova لبحث الفروق بين ثلاث مجموعات، وتثبيت حجم الأثر الذي تم الحصول عليه من خلال حجم العينة المستخدم في التجربة الأولى، حيث استخدم نصار (2006) مربع ايتا (η^2) لقياس حجم الأثر، وثبته عند (0.229)، أما Kellow(1998) فاستخدم مقياس نسبة التباين المفسر (r^2) لقياس حجم الأثر، وثبته عند (0.333)، وبعد إجراء تجارب عدة بزيادة أحجام العينات؛ توصلت كلتا الدراستين إلى أن احتمالية رفض الفرضية الصفرية يزيد بزيادة حجم العينة.

كما ضرب هوك Huck(2009) مثالا لتأثر الدلالة الإحصائية بحجم العينة حيث توصل عند حجم عينة (10) بلغ معامل الارتباط (0.55) وهو ذات دلالة إحصائية عند مستوى دلالة أقل من (0.05)، أما عند زيادة حجم العينة إلى (5000) فكان معامل الارتباط ضعيف جدا يكاد ينعدم حيث بلغ (0.03)، وعلى الرغم من صغره إلا أنه ذو دلالة إحصائية عند مستوى دلالة أقل من (0.05).

وحيث أن الباحثَين نصار(2006) و Kellow(1998) استخدما في أمثلتهما نفس الاختبار الإحصائي وهو تحليل التباين الأحادي؛ فإن الباحثين لهذه الدراسة قاموا بإعادة تجربتهما على مثال افتراضي لكن بتغيير الاختبار الإحصائي، حيث استخدم الباحثون اختبار ت T-Test لعينتين مستقلتين، وذلك للتأكيد على أثر حجم العينة في رفض الفرضية الصفرية وقبول الفرضية البديلة، حيث بحث الباحثون أثر التدريس باستخدام الاستراتيجية التفاضلية في الاتجاه نحو الرياضيات، لطلاب معهد العلوم الإسلامية، وقد افترض حجم العينة بثمانية طلاب، أربعة طلاب لكل مجموعة من مجموعتي الدراسة(التجريبية التي استخدمت الاستراتيجية التفاضلية في التدريس، والمجموعة الضابطة)، ويوضح الجدول (1) نتائج الطلاب المفترضة لمتوسطات الاتجاه نحو الرياضيات.

الجدول (1): نتائج طلاب المفترضة لمتوسطات الاتجاه نحو الرياضيات.

الطلاب	المجموعة التجريبية				المجموعة الضابطة			
	1	2	3	4	1	2	3	4
المتوسط الحسابي	4.3	4.0	4.1	4.0	4.0	4.0	4.2	4.0
المتوسط الحسابي للمجموعة	4.10				4.05			

ومن خلال الجدول (1) نجد أن الفرق بين متوسطتي المجموعتين التجريبية والضابطة في الاتجاه نحو الرياضيات يعتبر ضئيل جدا حيث بلغ (0.05)، وقد تم استخدام اختبار ت للتأكد من عدم دلالة هذه الفروق، وبينت النتائج عدم وجود فروق ذات دلالة إحصائية متوسطتي المجموعتين التجريبية والضابطة في

الاتجاه نحو الرياضيات، حيث بلغت قيمة (ت) (0.577) وقيمة P-value (0.585) وهي غير دالة عند مستوى دلالة أقل من (0.05)، كما بلغت الدلالة العملية أو حجم الأثر باستخدام مؤشر d (0.471)، وهو مؤشر متوسط، يدل على أن استخدام الاستراتيجية التفاضلية في التدريس له أثر متوسط في زيادة الاتجاه نحو الرياضيات عند الطلاب. وتم تثبيت قيمة الدلالة العملية (0.471) وزيادة حجم العينة في المجموعتين، من (4) طلاب في كل مجموعة إلى (48) طالب في كل مجموعة، وإجراء اختبار (ت) في كل مرة، والجدول (2) يوضح نتائج هذه الاختبارات.

الجدول(2) نتائج اختبارا ت لمتوسطات الاتجاه نحو الرياضيات المفترضة عند تغيير حجم العينة (16، 32، 64، 80، 96)، وتثبين قيمة الدلالة العملية (مؤشر d) عند (0.471)

المجموعة	العدد	المتوسط الحسابي	الانحراف المعياري	درجات الحرية	حجم العينة	قيمة ت	مستوى الدلالة	القرار الإحصائي	الدلالة العملية (مربع ايتا)	الدلالة العملية (مؤشر d)
التجريبية	4	4.10	0.14	6	8	0.577	0.585	غير دالة	0.053	0.471
الضابطة	4	4.05	0.10							
التجريبية	8	4.10	0.13	14	16	0.882	0.393	غير دالة	0.053	0.471
الضابطة	8	4.05	0.09							
التجريبية	16	4.10	0.13	30	32	1.291	0.207	غير دالة	0.053	0.471
الضابطة	16	4.05	0.09							
التجريبية	32	4.10	0.12	62	64	1.856	0.068	غير دالة	0.053	0.471
الضابطة	32	4.05	0.09							
التجريبية	40	4.10	0.12	78	80	2.082	0.041	دالة	0.053	0.471
الضابطة	40	4.05	0.09							
التجريبية	48	4.10	0.12	86	88	2.186	0.032	دالة	0.053	0.471
الضابطة	48	4.05	0.09							

*من إعداد الباحثين

يتضح من الجدول (2) أن قيمة الدلالة العملية (مؤشر d) ثابتة عند قيمة (0.471)، بمعنى أنها لم تتأثر بزيادة حجم العينة من (16) إلى (96)، وأن قيمة اختبار (ت) زادت بزيادة حجم العينة، من (0.577) عند حجم عينة (8)، إلى أن بلغت (2.186) عند حجم عينة (96)، مما أدى إلى أن قيمة مستوى الدلالة P-value انخفضت تدريجيا من (0.585) عند حجم عينة (16)، وهي غير دالة عند مستوى دلالة أقل من (0.05) إلى أن وصلت إلى (0.041) عند حجم عينة (80)، وهي ذات دلالة

إحصائية عند مستوى دلالة أقل من (0.05)، بمعنى أن الفرضية الصفرية تم رفضها عندما بلغ حجم العينة (80) على الرغم من أن حجم الأثر ثابت ولم يتغير، وبالتالي فإن هذا المثال يؤكد ما توصل إليه الباحثون السابقون من أن الدلالة الإحصائية ليست جازمة بوجود علاقة قوية بين المتغير المستقل أو المتغيرات المستقلة والمتغير التابع أو المتغيرات التابعة، فقد تكون بسبب كبر حجم العينة فقط.

2) الطبيعة الثنائية لنتائج الدلالة الإحصائية:

نظرا لوجود مستويين فقط لتفسير نتيجة الدلالة الإحصائية (دالة إحصائيا، غير دالة إحصائيا) يُشعر بأن هناك خطًا فاصلًا دقيقًا يفصل بين رفض أو قبول الفرضية الصفرية، وحيث أن المستوى المتعارف عليه في الدلالة الإحصائية هو (0.05) أو (0.01)، فهذا جعل القيمة (0.05) لها قدسية وأهمية بالغة في الاختبارات الإحصائية فهي الفيصل في قبول أو عدم قبول النتائج، مما يوقع في سوء فهم التفسير الصحيح للنتائج التي يتم الحصول عليها (بابطين، 2002؛ Schneider,2013)، فمثلا إذا قام باحثان بإجراء نفس الدراسة وتوصل أحدهما إلى أن نتيجة قيمة الاحتمال P-value والتي تمثل الدلالة الإحصائية تساوي (0.049)، بينما نتيجة الدراسة الآخرى بلغت (0.051)، فهنا يتم الحكم على الدراسة الأولى بأنها جيدة ونتائجها ذات دلالة إحصائية، أما الدراسة الثانية فنتائجها ليست دالة إحصائا، وليس لها أي اعتبار مع أن الفرق بينهما في قيمة الاحتمال P-value يساوي (0.002) فقط، وقد يكون تأثير المعالجات في الدراسة الثانية أفضل منها في الدراسة الأولى، وذكر Perezgonzalez(2015) إلى أن Johnstone(1987) دعا إلى عدم جمود مستوى القبول للدلالة الإحصائية عند مستوى (0.05) بحيث يمكن اعتبار القيم (0.049) و (0.051) لهما نفس الدلالة الإحصائية حول مستوى (0.05).

هذا وأكدت الجمعية الأمريكية للإحصاء (ASA) في مبادئها التي أصدرتها عام 2016م حول استخدام قيمة الاحتمال p-value، على عدم الاستناد على ثنائية P-value كمقياس في اتخاذ القرارات، حيث نص المبدأ الثالث على أن "يجب أن لا تستند الاستنتاجات العلمية والقرارات الاقتصادية أو السياسية فقط على ما إذا كانت قيمة P تمر بعتبة محددة"، وإنما يجب على الباحثين الأخذ بعوامل أخرى لإصدار قراراتهم النهائية، كتصميم الدراسة، وجودة القياسات، وصحة الافتراضات،..إلخ (Wasserstein & Lazar,2016, p131).

3) منطقية الفرضية الصفرية:

تُصاغ الفرضية الصفرية على أساس عدم وجود فروق أو علاقات بين متغيرات الدراسة(معالم المجتمع)، أو أن الفروق بين المتغيرات تساوي صفر، وهذا الأمر مخالف للواقع، حيث أن كل شيء في هذا الوجود له علاقة أو ارتباط بالأشياء الأخرى ولو بالقدر اليسير جدا(Kirk, 2003, p86)، خاصة في العلوم الإنسانية والنفسية والاجتماعية، فلا يكاد توجد صفة أو سمة معينة غير مرتبطة بغيرها من الصفات، لذا فإن بناء الفرضية الصفرية بصيغة النفي التام للفروق أو العلاقات ليس له أساس منطقي، ويشير كوهن(1994) كما في Schneider(2013, p10) إلى "أن الفرضية الصفرية غير قابلة للتصديق، ورفضها لا فائدة منه" وذلك لأنها غير موجودة في الواقع أصلا، كما أن الفرضية الصفرية لا بد من رفضها عند حجم معين من العينة، مهما كان الفرق بين المتغيرات صغير جدا لدرجة أن يُقال عنه بأنه "فرق تافه" (Mayo,2006, p809).

الإجابة عن السؤال الثاني:

ما الاستخدامات الخاطئة للدلالة الإحصائية في تفسير النتائج؟

أدى الانتشار الواسع لاختبارات الدلالة الإحصائية في مختلف الدراسات النفسية والاجتماعية إلى سوء استخدامها في تفسير النتائج، ولعل السبب في ذلك يعود إلى أن أساتذة مناهج البحث نادرا ما يعلمون طلابهم كيفية تفسير نتائجهم بطرق مفيدة لغير الإحصائيين(Ellis,2010, p3)، ولقد تنبه العلماء إلى سوء استخدام الدلالة الإحصائية في تفسير النتائج منذ ثلاثينيات القرن الماضي، حيث ذكر Daniel,1997 المذكور في بابطين(2002) أن تايلور 1931م ذكر بأن "التفسيرات التي عادة ما تستنتج من الدراسات الحديثة تُشير بوضوح إلى ميلنا للاعتقاد والتصور بأن الدلالة الإحصائية تكافأ الدلالة الاجتماعية، هذان المصطلحان مختلفان بالضرورة ويجب ألا يُخلطا، الفروق الدالة إحصائيًا ليست دائما دالة اجتماعيًا ، والفروق غير الدالة إحصائيًا ربما تكون دالة اجتماعيًا"، وذكر كارفور carver(1978) بأن مصدر التفسيرات الخاطئة للدلالة الإحصائية يكمن في الفئات الثلاث الآتية:

- P=0.05 تعني أن نسبة (5%) فقط من النتائج ناتجة عن الصدفة، أو أن (95%) منها لم تكون ناتجة عن الصدفة.
- أن النتائج موثوقة بنسبة (95%)، وأن نسبة تكرارها تبلغ (95%).

- احتمال صحة فرضية البحث يبلغ (0.95).

وعارض كارفور carver(1978) استخدام اختبارات الدلالة الإحصائية، وأشار إلى أن البحث التربوي سيكون في وضع أفضل إذا تم التوقف عن استخدام اختبارات الدلالة الإحصائية، واستمرت الانتقادات حول سوء استخدامات الدلالة الإحصائية، حتى وقتنا هذا، لما تعانيه من قصور في ذاتها، ولمحدوديتها في تفسير النتائج العلمية، من مثل دراسات (Welge-Crow,1990; Shaver,1992; Thompson,1997; Kaufman,1998; Nix & Barnette, 1998; Hubbard & Lindsay,2008; Kmetz,2019; Hurlbert & Levine & Utts,2019; Johnson,2019; Amrhein & Greenland & McShane;2019; Ioannidis,2019; Greenland,2019)، ودعوا إلى ضرورة تصحيح المسار للدلالة الإحصائية، وحدد نيكس وبارنيت في دراسته النظرية Nix & Barnette (1998) الانتقادات الأكثر صخبا لاختبارات الدلالة الإحصائية التي ظهرت في الأدب على مدى السنوات الـ(50) الماضية التي سبقت دراسته أي في الفترة (1948-1998)، وذكر (cumming,2012, p26) بأن "الاستخدامات الخاطئة للدلالة الإحصائية منتشرة بقوة في أوساط الباحثين، والطلبة، وحتى بين أساتذة الإحصاء في أقسام علم النفس".

ولقد لخص (Goodman,2008) التفسيرات الخاطئة للدلالة الإحصائية في (12) جزئية، وبسبب مخاوف سوء فهم الدلالة الإحصائية، وقيمة الاحتمال p-value، والتفسيرات الخاطئة لها في البحوث التطبيقية، دعت الجمعية الأمريكية للإحصاء (ASA) عام 2014م لعقد اجتماعات مع مجموعة من الإحصائيين والخبراء من مجموعات متنوعة من التخصصات لصياغة بيان حول سياسة استخدام قيم P واختبار الفرضيات (Grabowski2016,)، وخلال عام 2016م أصدرت الجمعية (ASA) بيانًا رسميًا تحذر من التفسيرات الخاطئة للدلالة الإحصائية، واشتمل البيان على ستة مبادئ (Wasserstein & Lazar,2016; Gagnier & Morgenstern,2017; Thomas & et al.,2017, Amrhein & Greenland & McShane, 2019)، حول التفسير الصحيح للدلالة الإحصائية وقيم الاحتمال P-value، واختبار الفرضيات وصنع القرار في العلوم المختلفة، والسياسات، ومن أهم التفسيرات الخاطئة لنتائج الدلالة الإحصائية ما يلي:

1) استخدام نتائج الدلالة الإحصائية كأساس لتفسير أهمية النتائج:

تمثل هذه الجزئية من أهم التفسيرات الخاطئة للدلالة الإحصائية وأكثرها انتشارًا وشيوعًا، التي وقع فيها أغلب الباحثون خاصة المبتدئين منهم، حيث يتم التركيز على أهمية الدلالة الإحصائية واعتبارها الهدف الأساس لنتيجة الاختبارات الإحصائية المستخدمة، فاعتبار النتائج الدالة إحصائيًا مهمة في ذاتها وعند مقارنتها بالنتائج غير الدالة إحصائيا (Gagnier & Morgenstern,2017, p1602; Schneider,2013)، وقد قام Ziliak & McCloskey(2004) بمراجعة (137) دراسة منشورة في المجلة الاقتصادية الأمريكية American Economic وتوصلا إلى أن (82٪) خلطوا بين الدلالة الإحصائية والأهمية الاقتصادية، وأن النتيجة الإيجابية للدلالة الإحصائية للدراسة تُشعر الباحث بالأريحية والطمأنينة بأن نتائج دراسته جيدة ومهمة وتستحق التقدير، والنشر(Mayo,2006, p809)، وأن عدم الوصول إلى مستويات الدلالة الإحصائية المقبولة يؤدي إلى التقليل من قيمة الدراسة بل ورفضها في بعض الأحيان(نصر الله، 2016)، وقد هيمن هذا الفكر الخاطئ على بعض المحكمين للدراسات وبعض المجلات المحكمة (Wasserstein & Lazar,2016, p131; Amrhein & Greenland & McShane, 2019, p306)، وأدى بهم إلى التحيز للدراسات ذات الدلالة الإحصائية، ورفض الدراسات التي لا تُظهر نتائجها دلالة إحصائية للمعالجات المدروسة، (Masicampo & Lalande,2012)، بل وصل الأمر إلى تفضيل الدراسات التي دلالاتها الإحصائية عند مستوى دلالة أقل من (0.01) على التي أعتبر مستوى (0.05) للدلالة الإحصائية، بغض النظر عن الدلالة العلمية وحجم التأثير، لذلك ظهرت في الفترة الأخيرة دعوات تنادي باحتضان "عدم الدلالة الإحصائية"، أو "عدم اليقين" (Johnson,2019)، في إشارة إلى أن النتائج التي لا تظهر دلالة إحصائية قد تكون لها أهمية عملية وينبغي النظر إليها والاستفادة منها.

2) استخدام نتائج الدلالة الإحصائية كأساس للدلالة العملية:

ضرب هوك Huck(2009) مثالا للخلط بين الدلالة الإحصائية والدلالة العملية بمثل الذي ينفق الوقت والقدرة والمال في شيء ما ويتوقع أن أفعاله تحدث فرق كبيرا، فتوفر الأمور الأساسية كالوقت والمال والقدرة لإنجاز أمر ما لا تعني أن ذلك الأمر قد تم انجازه بالفعل، ويشبه الباحثون العلاقة بين الدلالة الإحصائية والدلالة العملية بعلاقة التحصيل الدراسي بالذكاء، فقد يعتقد البعض أن الحصول على درجة عالية في مادة ما يدل على ذكاء عالي لدى الطالب، مع أن هذا الأمر ليس صحيحا على إطلاقه، حيث أن الحصول على درجة عالية في مادة ما قد تكون بسبب مساعدة الآخرين، أو بسبب الغش في الامتحان أو بسبب سهولة الاختبار، أو بسبب تحيز الأستاذ مع الطالب أو تساهله في التصحيح، فالحصول على مستوى عالي في التحصيل مؤشر على الذكاء فقط وليس جزما بذلك، كما أن الحصول على مستوى متدني

في التحصيل لا يدل دائما على قلة الذكاء أو انخفاضه، لذا يعتبر متغير التحصيل أو الذكاء في الإحصاء من المتغيرات الفئوية وليست من المتغيرات النسبية، حيث الحصول على صفر في التحصيل لا يدل على انعدام المعرفة تماما.

هذا وجاء الخلط بين الدلالتين الإحصائية والعملية بسبب الاعتقاد بأهمية النتائج الدالة إحصائيا، مما أدى بالباحثين إلى اعتبار أن الدلالة الإحصائية دليلًا لوجود تأثير المعالجات التجريبية أو تأثير المتغيرات المستقلة على المتغيرات التابعة مع أن الصحيح عدم الجزم بصحة ذلك (Goodman,2008,p137; Huck,2009; Grabowski,2016; Gagnier&Morgenstern,2017, p1602)، وقد تكون الدلالة الإحصائية سببها زيادة حجم العينة المستخدمة في الدراسة، كما ذكر الباحثون سابقا بأن الدلالة الإحصائية تتأثر بحجم العينة، فجميع الاختبارات الإحصائية سيكون لها دلالة إحصائية عند حجم معين من العينة، كما أن الدلالة الإحصائية لا تهتم بجودة النتائج وأثرها في الواقع، وجُل اهتمامها هو ما إذا كانت النتائج بسبب الصدفة أو تقلبات المعاينة، وذكرت الجمعية الإمريكية للإحصاء(ASA) في مبدأها الخامس بأن قيمة الاحتمال P-value لا تقيس حجم الأثر أو أهمية النتائج، فالقيم الصغيرة لـ P-value لا تشير بالضرورة إلى أهمية عالية أو حجم تأثير كبير(Wasserstein & Lazar,2016, p132).

ومن الأخطاء الشائعة أيضا هو اعتبار قيمة الدلالة الإحصائية لمؤشر لحجم الأثر في المعالجات التجريبية(عباس،2013)، فمثلا إذا كانت قيمة P-value تبلغ (0.058) فإنه يتم التعبير عند النتائج بأنها "تقترب من الدلالة الإحصائية"، وإذا كانت قيمة P-value تساوي (0.000)، فيُعبر عن النتائج بأنها "مرتفعة الدلالة."

3) استخدام الدلالة الإحصائية للدلالة على صحة الفرضية الصفرية:

يشير كل من هوك و جودمان (Huck,2009; Goodman,2008, p136) إلى أن "من المفاهيم الخاطئة حول استخدام قيمة الاحتمال p-value هو أن إذا كانت p=0.05 فإن ذلك يدل على أن الفرضية الصفرية لديها فرصة 5% لتكون صحيحة"، وإذا كانت P=0.3 فإن الفرضية الصفرية لديها فرصة30% لتكون صحيحة، وهذا ليس بصحيح فالدلالة الإحصائية لا تقيس مدى صحة أو خطأ الفرضية الصفرية، وإنما توضح احتمال الحصول على نتائج مختلفة عن الفرضية الصفرية (بابطين،2002)، ولقد أوضح المبدأ الثاني من مبادئ الجمعية الأمريكية للإحصاء (ASA) لسؤء استخدام الاختبارات الإحصائية في تفسير

النتائج أن "قمية الاحتمال p-value لا تقيس احتمالية الفرضيات أو احتمال أن تكون البيانات ناتجة عن الفرص العشوائية الوحيدة" (Gagnier & Morgenstern,2017, p1601).

الإجابة على السؤال الثالث:

ما الحلول المقترحة لتطوير الاستخدام الصحيح للدلالة الإحصائية؟

اختبارات الدلالة الإحصائية شأنها شأن أي نوع من أدوات التقويم للمعالجات التجريبية وشبة التجريبية في مختلف الدراسات الإنسانية أو التطبيقية، لها إيجابياتها وسلبياتها، فهي بالرغم من النقد الحاد ضدها منذ حوالي ما يقارب القرن من الزمن إلا أنها لا تزال متمسكة بأهميتها، ولم يتم التوصل إلى بدائل حاسمة تغني عنها بالكلية، لذلك الطريق الأمثل في التعامل معها هو إزالة ما شابها من أخطاء في استخدامها، ووضعها في مكانها الصحيح في تفسير النتائج، مع مراعاة جوانب القصور فيها، وتدعيمها بطرق أخرى مكملة لها، وقدم شالفر Shaver(1992) ثلاث استراتيجيات لزيادة تفسير نتائج اختبارات الدلالة الإحصائية، تتضمن الإستراتيجية الأولى تقييم نتائج اختبارات الدلالة الإحصائية في سياق حجم العينة، وتضمنت الإستراتيجية الثانية تفسير حجم التأثير كمؤشر لأهمية النتيجة، فيما أكدت الإستراتيجية الثالثة على التفسير بناءً على الاحتمال المقدر لتكرار النتائج، وقد أشار Wasserstein & Lazar(2016, p132) بأنه "في ضوء سوء الاستخدام لقيم الاحتمال P-value، والمفاهيم الخاطئة السائدة المتعلقة بها، يفضل بعض الإحصائيين استخدام طرق أخرى بديلة أو مكملة للدلالة الإحصائية P-value ، مثل الثقة والمصداقية، فترات التنبؤ، طرق بايزي، نسب الاحتمال، عوامل بايز، وغيرها من الطرق مثل النمذجة نظرية القرار ومعدلات الاكتشاف الخاطئة، وتعتمد كل هذه المقاييس والطرق على افتراضات أخرى، لكنها قد تعالج بشكل مباشر حجم التأثير، أو ما إذا كانت الفرضية صحيحة"، ويرى الباحثون أن الحلول المقترحة الأخرى للمسار الصحيح للدلالة الإحصائية ما يلي:

1) عدم الاعتماد على النتائج في تقييم البحوث والدراسات:

تم مؤخر نشر نموذج جديد يستند على سياسة "التقييم الأعمى للنتائج" للدراسات المقدمة إلى المجلات، حيث لا يتم إعطاء النتائج المنشورة في الدراسات أي وزن في القرار فيما يتعلق بملاءمة الدراسة للنشر، في خطوة للتخفيف من الاعتماد المفرط على اختبارات الدلالة الإحصائية، وسوء تفسيراتها، وينص النموذج على إعطاء الوزن بشكل حصري لـ:

(أ) المرحلة الأولى تقييم أهمية الأسئلة البحثية للدراسة، الذي تم تناولها في الدراسة، بحيث يتم توزيع مقدمة الدراسة ومنهجيتها فقط على لجنة التحكيم والمراجعة في المجلة، بغرض التقييم واتخاذ قرار مؤقت بشأن قبول أو رفض الدراسة للنشر، وفي حالة موافقة اللجنة على هذه الأسئلة البحثية، يتم الانتقال إلى المرحلة الثانية، أما في حالة رفض الاسئلة البحثية فيتم رفض البحث أو الدراسة كاملة، بدون النظر إلى النتائج.

(ب) المرحلة الثانية يتم تسليم الدراسة كاملة إلى لجنة التحكيم والمراجعة في المجلة لتقييم ومراجعة الأساليب المستخدمة وجودتها للإجابة عليها على الأسئلة البحثية، بما في ذلك مدى ملاءمة طرق تحليل البيانات (Locascio,2017; Locascio,2019) أما نتائج الدراسة التي تم الحصول عليها فلا يكون لها نصيب من التحكيم والتقييم.

ويرى الباحثون بأن هذا النموذج جيد كمرحلة أولية في طريق التخفيف من حدة الاستخدامات الخاطئة للدلالة الإحصائية، حيث يُعطى الباحثين متنفسا للتعبير عن نتائج دراساتهم بأريحية تامة وعدم الضغط عليهم للتحيز و ليَّ أعناق نتائجهم لتخطي عتبة مستوى الدلالة أقل من (0.05)، ويطمأنهم بإمكانية نشر دراساتهم وبحوثهم، إذا توفرت فيها المنهجية البحثية العلمية الصحيحة، وجودة أسئلتها البحثية، بغض النظر عن النتائج التي تم الحصول عليها، وفي المقابل فإن هذا النموذج لا يعالج القصور التي تعاني منه الدلالة الإحصائية من ثنائية النتيجة، ومشكلة العتبة السحرية الـ(0.05)، وغيرها التي تم ذكرها سابقا، لذلك لا بد من بذل المزيد من الجهد في إيجاد بدائل أخرى أكثر ملائمة توفر أفضل الحلول لاستخدامات الدلالة الإحصائية.

2) تطوير صياغة الأسئلة البحثية والفرضيات:

يبدو أن طبيعة الأسئلة البحثية المستخدمة في مختلف الدراسات هي التي أدت إلى سوء استخدام الدلالة الإحصائية في تفسير النتائج، حيث تقتصر صياغة الأسئلة بالبحث عن الفروق ذات الدلالة الإحصائية بين المتوسطات أو وجود العلاقة بين المتغيرات فقط، والتي تعتمد في إجابتها على رفض الفرضيات الصفرية، ولا تهتم تلك الأسئلة بالبحث عن قوة العلاقة، أو مقدار تأثير المتغيرات المستقلة في المتغيرات التابعة، وهذه النوعية من الأسئلة هي السائدة في مختلف الدراسات سواء كانت اطروحات دكتوراه أو رسائل ماجستير أو دراسات محكمة، وأوضح بابطين(2002) في دراسته لمشكلات الدلالة الإحصائية في رسائل الماجستير في كلية التربية بجامعة أم القرى؛ أن نسبة الأسئلة البحثية المرتبطة بالبحث عن الفروق الدالة إحصائيا أو عن وجود االعلاقة بين المتغيرات بلغت (61,2%)، وهي نسبة عالية، لذلك يقترح

الباحثون تطوير توجهات الدراسات والأبحاث بحيث يتم استبدال الأسئلة الباحثة عن الفروق فقط بهكذا تساؤلات بـ"ما مقدار التأثير للمتغيرات المستقلة في المتغيرات التابعة" أو "ما قوة العلاقة بين المتغيرات"، وهذه النوعية من الاسئلة البحثية سوف تسهم في الحد من سوء تفسير نتائج الدلالة الإحصائية، وتُلزم الباحثين من عدم الاكتفاء باختبارات الدلالة الإحصائية فقط في الإجابة عن تساؤلات أبحاثهم، والبحث عن بدائل أخرى تُوصلهم إلى الإجابة لتساؤلاتهم.

3) البدائل المكملة لاختبارات الدلالة الإحصائية:

الإحصاء علم واسع وسوء استخدام نظرياته أدى إلى خلق النقد الموجه إلى جزء منه وهو "اختبارات الدلالة الإحصائية"، وإلا فإن العلماء المنظرين لعلم الإحصاء أوضحوا لكل فرع من هذا العلم له استخداماته الصحيحة، وبينوا أن اختبارات الدلالة الإحصائية هي مرحلة أولية في الإجابة عن التساؤلات البحثية، والدلالة الإحصائية ليست مشكلة في حد ذاتها، وإنما تكمن المشكلة في سوء استخدامها في تفسير النتائج (Grabowski,2016)، وأكد ماكلين وإرنست McLean & Ernest(1998) بأنهم يدعمون الباحثين الآخرين الذين يوصون بأن اختبار الدلالة الإحصائية يجب أن يكون مصحوبًا بمؤشرات الدلالة العملية للمعالجات، وقابلية تكرارها، لذا فمن يريد الوصول إلى التفسير الصحيح لنتائج دراساته، فإن عليه استخدام الاختبارات والمقاييس المكملة لاختبارات الدلالة الإحصائية، ويمكن تلخيص هذه المقاييس في الآتي (بابطين،2002;2012 ,Cumming ;Grabowski,2016).

- مقاييس الدلالة العملية وحجم الأثر Effect Size.
- اختبار القوة الإحصائية Statistical Power.
- تقديرات فترة الثقة Confidence Interval.
- تحليل الإعادة Replication.

4) تحليل المنفعة العملية Practical Benefit:

يقدم Pogrow(2019) طريقة مبتكرة للتخفيف من مساوئ استخدام الدلالة الإحصائية، وهو نهجًا يعتمد على المنفعة العملية بدلاً من الأهمية الإحصائية أو العملية، وتهدف المنفة العملية إلى معرفة: (أ) كيف كان أداء أفراد المجموعة التجريبية بالفعل؟، (ب) هل كانت النتائج الفعلية لأفراد المجموعة التجريبية

أفضل بشكل ملحوظ من النتائج الحالية لهم؟، بعبارة أخرى فإن المنفعة العملية تقيس الفائدة المحتملة للمعالجة في نتائج أبسط وفعالة للمجموعة التجريبية (فقط)، ولا تهتم بمعرفة أداء المجموعة الضابطة، ويقول Pogrow(2019) إن هذا النهج مفيد بشكل خاص لتقييم ما إذا كانت التدخلات في المنظمات المعقدة (مثل المستشفيات والمدارس) فعالة ولها جدوى عملية.

وهذه الطريقة ليست بالحديثة حيث أنها تشابه بدرجة كبيرة جدا الدلالة الإكلينيكية التي قام كل من (نصار،2017؛ Peterson,2008) بإجراء بحوث حولها، وتمثل الدلالة الإكلينيكية مرحلة متقدمة ومكملة للدلالات الإحصائية والعملية، حيث يشير "مفهوم الدلالة الإكلينيكية للنتائج وخاصة تلك المستخدمة للتصاميم التجريبية إلى وجود فروق ليس فقط دالة إحصائيا أو / ودالة عمليا بين المجموعات التجريبية والضابطة بل إلى أن البرامج المستخدمة في تلك الدراسات فعالة من حيث تغيير واقع حال أفراد المجموعة أو المجموعات التجريبية من حالة إلى حال أخرى من المتوقع أن تكون هي الأفضل أو المرغوب بها" (نصار،2017، ص353).

التوصيات والمقترحات:

من خلال نتائج هذه الدراسة يقدم الباحثون مجموعة من التوصيات والمقترحات كالآتي:

- نشر الثقافة الإحصائية الصحيحة لدى الباحثين بشكل خاص، وأساتذة الجامعات والمهتمين بالإحصاء بشكل عام.
- التعمق في تدريس الإحصاء الاستدلالي في مقررات برامج الماجستير والدكتوراه في الجامعات والكليات، وخاصة فيما يخص الدلالة الإحصائية واستخداماتها الصحيحة، ومؤشرات الدلالة العملية، وعدم الاكتفاء بالتطبيقات الإحصائية للإحصاء الاستدلالي.
- تدريس مقرر الإحصاء الاستدلالي لطلبة مرحلة البكالوريوس في كليتي التربية والآداب.
- دراسة مجموعة البدائل عن الدلالة الإحصائية التي قدمها بعض الإحصائيين والتوصل إلى مقترح تطوير استخدام الدلالة الإحصائية في تحليل بيانات الدراسات والبحوث.
- نشر الدورات التدريبية في الاستخدامات الصحيحة للدلالة الإحصائية.

- إعادة النظر في معايير إجازة اطروحات الدكتوراه ورسائل الماجستير بكليات التربية والآداب، بحيث يُشترط تزويد كل اختبار إحصائي يتم استخدامه في تحليل النتائج بمؤشرات حجم الأثر.

- اقتراح إنشاء مركز الاستشارات الإحصائية يتبع عمادة الدراسات العليا بالجامعات، يسعى إلى نشر الثقافة الإحصائية للباحثين من الهيئة التدريسية والطلبة، وتوجيههم ومساعدتهم في الجوانب الإحصائية، من خلال الدورات التدريبية، والنشرات الدورية، والاستشارات الإحصائية، كما يكون من اختصاصات هذا المركز اعتماد اطروحات الدكتوراه ورسائل الماجستير التي يعدها الطلبة من مختلف الكليات، بحيث يعتمد الأساليب الإحصائية التي يقترحها الطالب في مخطط دراسته قبل مناقشة المخطط، والمرحلة الثانية يعتمد التحليل الإحصائي النهائي للدراسة.

المراجع:

بابطين، عادل أحمد(2002). مشكلات الدلالة الإحصائية في البحث التربوي وحلول بديلة. رسالة ماجستير غير منشورة، كلية التربية بجامعة أم القرى، السعودية.

باهي، مصطفى حسين إبراهيم (2010). العلاقة بين الدلالة الاحصائية وحجم التأثير في البحوث التربوية والنفسية. مستقبل إعداد المعلم في كليات التربية وجهود الجمعيات العلمية في عمليات التطوير بالعالم العربي– كلية التربية، جامعة حلوان، مصر ، مج2، 415–444.

التل، سعيد، وأبو زينة، فريد كامل، والبطش، محمد وليد(2007). مناهج البحث العلمي (تصميم البحث والتحليل الإحصائي). الأردن، عمّان: دار الميسرة للنشر والتوزيع والطباعة، ط1.

الثبيتي، علي حامد (2008). تصاميم البحوث العلمية ودورها في صدق نتائج الدراسات التربوية. رسالة الخليج العربى– السعودية، 29(108).

حسن، عبدالمنعم أحمد(2008). أوجه القصور في استخدام مؤشرات الدلالة العملية في البحوث التربوية والنفسية، دراسات في المناهج وطرق التدريس، 134، 14–39.

حمداوي، جميل(2014). البحث التربوي مناهجه وتقنياته. لبنان، بيروت: دار الكتب العلمية.

خضر، أحمد إبراهيم (2013). فروض البحث : ماهيتها وأنواعها وشروطها ومصادرها. شبكة الألوكة/ موقع الدكتور أحمد إبراهيم خضر/ صناعة الرسالة العلمية، /51442/0/https://www.alukah.net/web/khedr، تاريخ الاسترجاع 2018/12/12

الدليمي، عصام حسن أحمد، و صالح، علي عبدالرحيم (2014). البحث العلمي أسسه ومناهجه. الأردن، عمان: دار الرضوان للنشر والتوزيع، ط1.

سلامة، حسن علي حسن (2004). الدلالة الإحصائية والدلالة العلمية في البحوث التربوية. *المجلة* التربوية–جامعة سوهاج – كلية التربية، مصر ج20، 3–14.

السواح، نادر شعبان إبراهيم (2007). الإسهام في الإحصاء التطبيقي. مصر، الإسكندرية: الدار الجامعية.

السيد، ياسر أحمد، (2009). الإحصاء التطبيقي. مصر، الإسكندرية: مكتبة بستان المعرفة.

عباس، عبدالقادر(2013). أساسيات البحث العلمي: كتابة التقارير. مصر، القاهرة: دار الكتاب الحديث، ط1.

عباس، عبدالقادر(2013). طبيعة البحث العلمي والدلالة الإحصائية. مصر، القاهرة: دار الكتاب الحديث، ط1.

قسيس، منال، وجبر، أحمد فهيم، وأبوسمرة، محمود أحمد(2008). معوقات توظيف البحوث التربوية من وجهات نظر أعضاء الهيئة التدريسية في كليات التربية في الجامعات الفلسطينية. مجلة إتحاد الجامعات العربية–الأردن ، ع 50.

نصار، يحيى حياتي بكر(2017). الدلالة الاكلينيكية للبحوث النفسية والتربوية المستخدمة للتصاميم التجريبية :دراسة تحليلية. مجلة الدراسات التربوية والنفسية – جامعة السلطان قابوس، كلية التربية، سلطنة عمان، 11(2).

نصار، يحيى حياتي(2006). استخدام حجم الأثر لفحص الدلالة العملية للنتائج في الدراسات الكمية. مجلة العلوم التربوية والنفسية – البحرين، 7(2)، 35–59.

نصر الله، عمر، (2016). أساسيات مناهج البحث العلمي وتطبيقاتها. الأردن، عمّان: دار وائل للنشر والتوزيع، ط1.

AERA (2006), “Standards for reporting on empirical social science research in AERA publications,” . Retrieved from https://www.aera.net/Portals/38/docs/12ERv35n6_Standard4Report%20.pdf

Amrhein, V, Greenland, S , & McShane,B (2019). Retire statistical significance. NATURE,567 , 305–307

Aron,A., & Aron,E. N. (1994). Statistics for psychology. Prentice Hall, Englewood Cliffs, New Jersey 07632.

Bezzina, F., & Saunders, M. N. (2014). The pervasiveness and implications of statistical misconceptions among academics with a special interest in business research methods. Electronic Journal of Business Research Methods, 12(2) , 103-114.

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

Cumming,G.(2012). Understanding The New Statistics: effect Sizes, confidence ,intervals, and meta-analysis. Routledge Taylor & Francis Group, New York. United States of America.

Ellis,Paul D.(2010). The Essential Guide to Effect Sizes Statistical Power, Meta-Analysis, and the Interpretation of Research Results. Cambridge University Press, New York, First published.

Gagnier, Joel, Morgenstern, Hal(2017). Misconceptions, Misuses, and Misinterpretations of P Values and Significance Testing . The Journal of Bone and Joint Surgery, 99(18), 1598 -1603, DOI: http://dx.doi.org/10.2106/JBJS.16.01314.

Gigerenzer, G. (2004). Mindless statistics. J. Soc. Econ. 33, 587-606. Doi:10.1016/j.socec.2004.09.033.

Goodman, S. N. (2008). A dirty dozen: twelve p-value misconceptions. Seminars in Hematology,45(3), 135-140.

Grabowski ,Beatrice (2016). "P < 0.05" Might Not Mean What You Think: American Statistical Association Clarifies P Values. Journal of the National Cancer Institute.108(8), djw194, Retrieved from http:// https://doi.org/10.1093/jnci/djw194.

Greenland, S. (2019). Valid p-Values Behave Exactly as They Should: Some Misleading Criticisms of p-Values and Their ResolutionWith s-Values. The American Statistician, 73:sup1, 106-114. Doi:10.1080/00031305.2018.1529625.

Gunn, H,J,(2019). Evaluation of Five Effect Size Measures of Measurement Non-Invariance for Continuous Outcomes, Graduate Theses and Dissertations, A Dissertation Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy ,ARIZONA STATE UNIVERSITY.

Haig, B. D. (2017). Tests of Statistical Significance Made Sound. Educational and Psychological Measurement,77(3). 489-506, DOI: 10.1177/0013164416667981

Hubbard, R. ; Armstrong, J. S (2006). Why We Don't Really Know What Statistical Significance Means: Implications for Educators. Journal of Marketing Education,28(2),114-120, DOI: 10.1177/0273475306288399

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, DOI: 10.1177/0959354307086923, Retrieved from http://tap.sagepub.com

Huck, S.W. (2009). Statistical Misconceptions. Psychology Press Taylor & Francis Group LLC, New York, NY 10016.

Hurlbert, S., Levine, R., & Utts, J. (2019). Coup de Grâce for a Tough Old Bull: 'Statistically Significant' Expires. TheAmerican Statistician, 73:sup1, 352-357. Doi:10.1080/00031305.2018.1543616.

Ioannidis, J. (2019). What Have We (Not) Learnt From Millions of Scientific Papers With p-Values?. The American Statistician, 73:sup1, 20-25. Doi: 10.1080/00031305.2018.1447512.

Johnson .V. E.(2019). Is it the end of 'statistical significance'? The battle to make science more uncertain. Retrieved from https://theconversation.com/is it the end of statistical significance the battle to make science more uncertain.114161.

Kaufman, Alan S.(1998). Introduction to the Special Issue on Statistical Significance Testing. Research in the Schools,5(2).

Kirk,R.E. (2003). The importance of effect magnitude, in S.F. Davis (editor), Handbook of Research Methods in Experimental Psychology. Oxford, UK: Blackwell, 83–105.

Kmetz, J. (2019). Correcting Corrupt Research: Recommendations for the Profession to Stop Misuse of p-Values. The American Statistician, 73:sup1, 36-45. Doi: 10.1080/00031305.2018.1518271.

Lane, David. M. (2013) Introduction to Statistics: An Interactive eBook.

Little, J.(2001). UNDERSTANDING STATISTICAL SIGNIFICANCE: A CONCEPTUAL HISTORY, J. TECHNICAL WRITING AND COMMUNICATION,31(4), 363-372.

Locascio, J. (2017). Results Blind Science Publishing. *Basic and Applied Social Psychology ,39(5)*,239–246. Doi:10.1080/01973533.2017.1336093.

Locascio, J. J. (2019). The Impact of Results Blind Science Publishing on Statistical Consultation and Collaboration. The American Statistician, 73:sup1, 346-51. DOI: 10.1080/00031305.2018.1505658.

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. The Quarterly Journal of Experimental Psychology, 65(11), 2271–2279.

Mayo, D. (2006). Philosophy of Statistics. In: S. Sarkar & J. Pfeifer (Eds), The Philosophy of Science: An Encyclopedia. Routledge: London, 802-815.

McLean, J., & Ernest, J. M.(1998).The Role of Statistical Significance Testing in Educational Research. Research in the Schools. 5(2), 15-22.

Nix ,T. W. , & Barnette ,J. J. (1998). The Data Analysis Dilemma: Ban or Abandon. A Review of Null Hypothesis Significance Testing. Research in the Schools, 5(2), 3-14.

Perezgonzalez, J. D. (2014). A reconceptualization of significance testing. Theor. Psychol. 24(6), 852- 859. Doi:10.1177/0959354314546157.

Perezgonzalez, J. D.(2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. Frontiers in Psychology.6, 1- 11.

Peterson, L. (2008). "Clinical" Significance: "Clinical" Significance and "Practical" Significance are NOT the Same Things. Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, February 7.

Pogrow, S. (2019). How Effect Size (Practical Significance) Misleads Clinical Practice: The Case for Switching to Practical Benefit to Assess Applied Research Findings. The American Statistician, 73:sup1, 223-234,Doi: /10.1080/00031305.2018.1549101.

Schneider, J.W.(2013). Caveats for using statistical significance tests in research assessments. Journal of Informetrics, 7(1), 50-62.

Shaver, James P.(1992). What Statistical Significance Testing Is, and What It Is Not. Paper Presented at Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).

STAT 502(2018). The 7 Step Process of Statistical Hypothesis Testing. Penn State Eberly College Of Science. Retrieved from https://onlinecourses.science.psu.edu/stat502.

Thomas R.L., Barach P.R., Wilkinson J.D., Farooqi A.A., Lipshultz S.E.(2017). The danger of relying on the interpretation of p-values in single studies:

Irreproducibility of results from clinical studies. Progress in Pediatric Cardiology, 44 , 57-61.

Thompson, B. (1997). Rejoinder: Editorial Policies Regarding Statistical Significance Tests: Future Comments. Educational Research. 26,(5),29- 32.

Wasserstein RL, Lazar NA(2016). The ASA's statement on p-values: context, process, and purpose. Am Stat.70(2),129-33.

Welge-Crow, Patricia A., & Others (1990). Looking Beyond Statistical Significance: Result Importance and Result Generalizability. Paper Presented at Annual Meeting of the American Psychological Society (Dallas, TX, June 9,1990).

Ziliak, S.T. and D.N. McCloskey (2004). Size matters: The standard error of regressions in the American Economic Review. Journal of Socio-Economics, 33(5), 527-546