

Articles in Press

Item-Fit Analysis in Educational Assessment: Effects of Test Length, Sample Size, and Missing Data under the Partial Credit Model

Mohammed Khresat, Mo'en Alnasraween

PII: 1401.2025

DOI: <https://doi.org/10.35192/jjoas-h.1401.2025>

Reference: JJOAS-H 1401.2025



To appear in: Jordan Journal of Applied Science-Humanities Science Series

Received Date: 08 Jul 2025

Revised Date: 27 Jul 2025

Accepted Date: 07 Aug 2025

Please cite this article as: Khresat, M. & Alnasraween, M. (in press). Item-Fit Analysis in Educational Assessment: Effects of Test Length, Sample Size, and Missing Data under the Partial Credit Model. *Jordan Journal of Applied Science – Humanities Series*. Advance online publication. <https://doi.org/10.35192/jjoas-h.1401.2025>

This is a PDF of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability. This version will undergo additional copyediting, typesetting and review before it is published in its final form. As such, this version is no longer the Accepted Manuscript, but it is not yet the definitive Version of Record; we are providing this early version to give early visibility of the article.

Please also note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.



Item-Fit Analysis in Educational Assessment: Effects of Test Length, Sample Size, and Missing Data under the Partial Credit Model

أثر طول الاختبار وحجم العينة ونسبة البيانات المفقودة في دقة مؤشرات مطابقة الفقرة التقليدية والمعدلة وفقاً لنموذج التقدير الجزئي

Mohammed Khresat^{1*}, Mo'en Alnasraween¹

¹Yarmouk University, Irbid, Jordan

mallamkh@gmail.com*

Abstract

This study is situated within Educational Studies, particularly the field of educational measurement and assessment. It investigates the accuracy of traditional and modified item-fit indices used to evaluate the quality of test items in educational and psychological assessments based on polytomous item response theory models. Specifically, the study compares the performance of the traditional $S-\chi^2$ index and the modified Mimpure- χ^2 index under varying conditions of test length, sample size, and missing data. A simulation design was employed using three levels of test length: 20, 40, and 60 items; three levels of sample size: 250, 500, and 1,000 examinees; and three levels of missing data: 1%, 10%, and 20%. Two sets of simulated polytomous response data were generated using WinGen software. The first dataset was generated according to the Partial Credit Model to evaluate the Type I error rates of the two indices, whereas the second dataset was generated according to the Graded Response Model to estimate statistical power in detecting item misfit. The results showed statistically significant differences between the two indices across most experimental conditions. The modified Mimpure- χ^2 index outperformed the traditional $S-\chi^2$ index by producing lower Type I error rates while maintaining higher statistical power. The findings also indicated that test length, sample size, and missing data percentage influenced the performance of both indices. While the traditional $S-\chi^2$ index performed adequately when the missing data rate was low, particularly at 1%, its accuracy declined as the proportion of missing data increased. In contrast, the modified Mimpure- χ^2 index demonstrated more stable and consistent performance across missing data conditions. These findings contribute to Educational Studies by supporting more accurate test-item evaluation and strengthening the validity and reliability of educational assessment practices. The study recommends using modified item-fit procedures when analyzing polytomous test data with incomplete responses in educational testing contexts.

Keywords: Educational Assessment, Item fit, Missing Data, Simulation Study, Test Validity.

الملخص

هدفت هذه الدراسة إلى تقييم دقة مؤشري مطابقة الفقرة، المؤشر التقليدي ($S - \chi^2$) والمؤشر المعدل ($M_{impute} - \chi^2$)، في الكشف عن الفقرات غير المطابقة في الاختبارات التربوية والنفسية التي يتم تصميمها وفقاً لنماذج نظرية استجابة الفقرة متعددة التدرج، وذلك من خلال دراسة تأثير ثلاثة عوامل لكل عامل ثلاثة مستويات: طول الاختبار (٢٠، ٤٠، ٦٠) فقرة، وحجم العينة (٢٥٠، ٥٠٠، ١٠٠٠) فرد، ونسبة البيانات المفقودة

(١٠٪، ٢٠٪)، وتمت الدراسة باستخدام مجموعتين من البيانات متعددة التدرج (Polytomous)، تم توليدهما باستخدام برمجية (WinGen)، إحداهما وفقاً لنموذج التقدير الجزئي لقياس الأداء من خلال الخطأ من النوع الأول، والثانية وفقاً لنموذج الإجابة المتدرجة لحساب قوة الاختبار. وكشفت النتائج عن وجود فرق دال إحصائياً ($\alpha = 0.05$) بين المؤشرين في أغلب الحالات، حيث تفوق المؤشر المعدل على المؤشر التقليدي في خفض معدلات الخطأ من النوع الأول والحفاظ على قوة اختبار مرتفعة، وأظهرت النتائج أن العوامل الثلاثة تؤثر في أداء المؤشرين، إذ ظهر أن كلا المؤشرين يتأثران وبدرجات متفاوتة بتغير طول الاختبار وحجم العينة، أما بالنسبة لعامل نسبة البيانات المفقودة، فقد أظهرت النتائج أن المؤشر التقليدي أدى بشكل جيد عند النسبة ١٪، ولكنه تراجع بشكل ملفت مع ارتفاع نسبة البيانات المفقودة، بينما قدم المؤشر المعدل أداءً أكثر ثباتاً واستقراراً.

الكلمات المفتاحية: التقييم التربوي، مطابقة الفقرة، البيانات المفقودة، دراسة محاكاة، صدق الاختبار.

المقدمة

الإطار النظري والدراسات السابقة

يسعى كل مقوم للوصول إلى القياس الجيد، وجودة القياس تتطلب أن تكون فقرات الاختبار متطابقة مع النموذج المستخدم. وتلعب مؤشرات مطابقة البيانات دوراً محورياً في تحقيق هذا الهدف، إذ تعمل على تحديد الفقرات التي قد تحتاج إلى تعديل أو حذف لتحقيق اتساق عالٍ بين أداء الفقرات والسمات المستهدفة من عملية القياس. ويؤدي الاعتماد على مؤشرات مطابقة دقيقة إلى تحسين جودة التقديرات المعتمدة على الفقرات، (Orlando & Thissen, 2000).

ومع ازدياد استخدام هذه المؤشرات في التحليل التربوي والنفسي، اتجهت الدراسات إلى البحث في العوامل التي تؤثر على دقتها، إذ وُجد أن كفاءة مؤشرات المطابقة تتأثر بعدة عوامل مثل:

- طول الاختبار: حيث تتأثر دقة المؤشرات بتغير عدد فقرات الاختبار.
 - حجم العينة: فالعينات الصغيرة قد تؤدي إلى تقديرات غير مستقرة أو مضخمة لمؤشرات المطابقة.
 - توزيع القدرة الكامنة: فالتوزيعات غير المتماثلة (skewed) تؤثر على الاستجابات الناتجة.
 - البيانات المفقودة: وهي من أكثر العوامل التي تؤثر على كفاءة المؤشرات.
- وتلعب هذه العوامل دوراً في تحديد دقة المؤشرات وقدرتها على اكتشاف الفقرات التي لا تتطابق مع النموذج المفترض (Kang & Chen, 2008).

وشكلت البيانات المفقودة تحدياً في أبحاث القياس النفسي؛ فخلال مرحلة جمع البيانات، قد تفقد بعض الاستجابات لأسباب متعددة. بعض هذه الاستجابات المفقودة تكون نتيجة لتصميم الاختبار، كما في بعض الاختبارات التي لا يتم فيها تقديم كل الفقرات للمفحوص، مما يسبب فقداً في البيانات. وهناك نوع آخر من فقدان يحدث بسبب سلوك المستجيب أثناء الاختبار، مثل عدم إكمال الاختبار لاعتماده على السرعة، أو تجاهل المستجيب لفقرة معينة بعد قراءتها، وتركها فارغة وعدم الإجابة عليها (Mislevy & Wu, 1996). وقد تم تقديم العديد من المؤشرات الإحصائية لتقييم مدى مطابقة الفقرات للنموذج ضمن نماذج الاستجابة للفقرة (IRT) عند توفر بيانات كاملة. من بين هذه المؤشرات:

- مؤشر Q_1 (Yen, 1981).
- مؤشر G^2 (McKinley & Mills, 1985).
- مؤشر (χ^2, G^{2*}) (Stone, 2000).
- مؤشرا أورلاندو وثيسن ($\chi^2 - S$) و ($G^2 - S$) (Orlando & Thissen, 2000; 2003).

وقد استُخدمت هذه المؤشرات في حالات متعددة، مثل: فحص أخطاء النماذج في البيانات الثنائية أو المتعددة (Chon et al., 2010; Kang & Chen, 2008; Liang & Wells, 2009)، واختبار انتهاك فرضية التدرج (monotonicity) في منحنى الاستجابة (Orlando & Thissen, 2003)، واكتشاف أخطاء بسبب مصفوفة (Q) غير الصحيحة (Wang et al., 2015)، والتحقق من مطابقة الفقرات في النماذج متعددة الأبعاد أو الهرمية (Li & Rupp, 2011; Zhang & Stone, 2008; Zhang et al., 2018). كما استُخدم

اختبار سنهاري (Sinharay, 2005; 2006) لتقييم مطابقة الفقرات ضمن إطار بايزي باستخدام سلاسل ماركوف مونت كارلو، وداخل إطار تحليل البواقي استخدم هابerman (Haberman, 2009; Haberman et al., 2013) البواقي لهذا الغرض، واختبار (Lagrange Multiplier) (Glas & Suárez-Falcón, 2003)، ومؤشر رايس (RISE) (Douglas & Cohen, 2001)، وإحصاءات المعلومات المحدودة (Bartholomew & Leung, 2002; Cai et al., 2006; Maydeu-Olivares & Joe, 2005; Reiser, 2008)، ومؤشرات تعتمد على الجداول الهامشية لتحديد مصدر عدم المطابقة (Liu & Maydeu-Olivares, 2013; 2014). إلا أن هذه الدراسات لم تتناول الحالات التي تحتوي على بيانات مفقودة. وغالبًا ما تعتمد مؤشرات مثل (Q_1) و (G^2) و $(S - \chi^2)$ على مجموع الدرجات أو تقدير θ لتجميع الأفراد. ومع أن $(S - \chi^2)$ و $(S - G^2)$ أثبتا أداءً أفضل مع البيانات الكاملة (Chon et al., 2010; Orlando & Thissen, 2000)، إلا أنهما صمما للتعامل مع البيانات المكتملة، التي تعتمد على المجموع الكلي للدرجات، واستخدامهما في حالة البيانات غير المكتملة قد لا يكون مناسبًا. ولذلك قام كل من زانغ ووانغ (Zhang & Wang, 2022)، بتطوير مؤشرات جديدة من خلال الاعتماد على مؤشري أورلاندو وثيسن $(S - \chi^2)$ و $(S - G^2)$ للتعامل مع البيانات غير المكتملة، لتصبح $(M_{impute} - \chi^2_i)$ و $(M_{impute} - G^2_i)$ ، وسميت بالمؤشرات المعدلة.

وفي هذه الدراسة قام الباحثان باعتماد نموذج التقدير الجزئي (Partial Credit Model) (PCM)، لعمل مقارنة بين مؤشر المطابقة التقليدي $(S - \chi^2)$ ومؤشر المطابقة المعدل $(M_{impute} - \chi^2_i)$ ، حيث تعطى احتمالية استجابة الفرد لفئة معينة من فئات الفقرة حسب المعادلة الرياضية التالية: حيث يشير: $P_k(\theta)$: احتمال استجابة الفرد لفئة معينة من الفقرة، θ : القدرة الكامنة للفرد، δ_p : معلمة الصعوبة، الذي يحدد النقطة التي يحدث فيها الانتقال بين الفئات، M : العدد الكلي لفئات الإجابة.

وتحسب قيمة المؤشر التقليدي $(S - \chi^2)$ ، وقيمة المؤشر المعدل $(M_{impute} - \chi^2_i)$ باستخدام العلاقات الرياضية التالية:

$$S - \chi^2 = \sum_{k=Z_i}^{F-Z_i} \sum_{z=0}^{Z_i} N_k \times \left(\frac{(O_{ikz} - E_{ikz})^2}{E_{ikz}} \right) \quad 1.$$

$$M_{impute} - \chi^2 = \sum_{k=Z_i}^{F-Z_i} \sum_{z=0}^{Z_i} \left(\frac{(f_{ik} - N_k * E_{ikz})^2}{N_k * E_{ikz}} \right) \quad 2.$$

$$E_{ikz} = \left(\frac{\int P_i(z|\theta) \times f^{*i}(k - z|\theta) \times \varphi(\theta) d\theta}{\int f(k|\theta) \times \varphi(\theta) d\theta} \right) \quad 3.$$

حيث يشير (i) إلى رقم الفقرة، (z) فئة الدرجة داخل تلك الفقرة (مثل ٠، ١، ٢، ...، (Z_i) أعلى درجة ممكنة للفقرة (i) ، (F) الدرجة الكلية المثالية للاختبار والتي تُحسب بجمع جميع القيم القصوى (Z_i) لكل الفقرات، (f_{ik}) عدد الأفراد الذين أجابوا بشكل صحيح على الفقرة i في المجموعة k بناءً على البيانات المكتملة، (k) المجموعة الفرعية من المفحوصين، (N_k) عدد المفحوصين ضمن المجموعة (k) ، (O_{ikz}) النسبة الملاحظة من المفحوصين في المجموعة (k) الذين أجابوا بالدرجة (z) على الفقرة (i) ، في حين أن (E_{ikz}) تمثل النسبة المتوقعة لهذه الإجابة وفقًا لنموذج القياس المستخدم، $(P_i(z|\theta))$ احتمال أن يحصل مفحوص يمتلك السمة الكامنة (θ) على الدرجة (z) في الفقرة (i) ، $(f^{*i}(k - z|\theta))$ تمثل توزيع الدرجات الكلية المتوقعة لبقية الفقرات باستثناء الفقرة (i) بشرط السمة الكامنة (θ) ، $(f(k|\theta))$ توزيع الدرجات الكلية الكاملة لجميع الفقرات بشرط (θ) ، $(\phi(\theta))$ دالة الكثافة الاحتمالية الخاصة بتوزيع السمة الكامنة (θ) ، والتي يُفترض في الغالب أنها تتبع توزيعًا طبيعيًا، وأخيرًا، تمثل (θ) نفسها السمة الكامنة لدى المفحوص، والتي تعكس قدرته أو مهارته أو السمة المقاسة في المجال المعرفي أو المهاري موضوع الاختبار.

ويتم حساب كلا المؤشرين حسب الخطوات التالية:

١. ترتيب الأفراد حسب مستويات القدرة المقدرة بعد تقسيمها إلى فئات.
٢. تجميع الأفراد في مجموعات فرعية حسب فئات القدرة.
٣. حساب تكرار العلامات الملاحظة عند كل فئة من فئات القدرة.
٤. حساب النسبة المتوقعة لكل فئة من فئات القدرة باستخدام مراكز الفئات.

٥. المقارنة بين توزيعات الدرجات المرصودة والدرجات المتوقعة، وذلك بمقارنة $(S - \chi^2)$ بتوزيع مربع كاي علماً بأن درجات الحرية هي عدد الفقرات (l) مطروحاً منها عدد معالم الفقرة (m) ثم طرح واحد صحيح $(df=l-1-m)$.

من ناحية أخرى، يتم اتخاذ القرارات الإحصائية ضمن إطار نظرية اختبار الفرضيات، حيث يُفترض في الفرضية الصفرية (H_0) أن الفقرة تتطابق مع النموذج المستخدم. وتشير أدبيات القياس إلى أربع قرارات إحصائية ممكنة عند اختبار الفرضية الصفرية:

١. القرار صحيح (قوة الاختبار): وهو احتمال رفض الفرضية الصفرية (H_0) عندما تكون فعلاً خاطئة.
٢. الخطأ من النوع الأول (α): وهو احتمال رفض الفرضية الصفرية وهي فعلاً صحيحة (أي اعتبار الفقرة غير مطابقة وهي في الحقيقة مطابقة).
٣. القرار صحيح (مستوى الثقة): عدم رفض الفرضية الصفرية (H_0) وهي صحيحة.
٤. الخطأ من النوع الثاني (β): وهو احتمال عدم رفض (قبول) الفرضية الصفرية وهي خاطئة (أي اعتبار الفقرة مطابقة وهي في الحقيقة غير مطابقة).

وتعتمد صحة القرار الإحصائي على القدرة على ضبط الخطأ من النوع الأول والثاني، وينبغي التوازن بينهما. فمن جهة، قد يؤدي خفض مستوى الدلالة (α) إلى اعتبار فقرات على أنها مطابقة للنموذج وهي في الواقع غير مطابقة، بينما قد يؤدي التقليل من (β) إلى حذف فقرات فعلياً مطابقة للنموذج.

الدراسات السابقة

تناولت العديد من الدراسات مؤشرات مطابقة الفقرة للكشف عن الفقرات غير المطابقة للنموذج، ففي دراسة أجراها أورلاندو وثيسين (Orlando & Thissen, 2000)، قدموا فيها مؤشرين جديدين لجودة المطابقة لنماذج نظرية استجابة الفقرة (IRT) للبيانات الثنائية، وهما $(S - \chi^2)$ و $(S - G^2)$ ، ودلت النتائج على كفاءة المؤشرين في الكشف عن الفقرات المطابقة للنموذج المستخدم مقارنة بالمؤشرات الأخرى. وهدفت دراسة لكانغ وتشين (Kang & Chen, 2008) إلى تقييم أداء مؤشر المطابقة الإحصائي $(S - \chi^2)$ لنموذج الاستجابة المتدرجة (Graded Response Model, GRM) وخلصت الدراسة إلى أن مؤشر $(S - \chi^2)$ يُعتبر مؤشراً واعداً لتقييم مطابقة الفقرات في الاختبارات التربوية والنفسية ذات الفقرات متعددة الفئات، وقام شون وآخرون (Chon et al., 2010) بدراسة هدفت إلى مقارنة أداء خمس مؤشرات لمطابقة الفقرات مع النماذج المختلطة لنظرية استجابة الفقرة (IRT)، وأظهرت النتائج أن $(S - \chi^2)$ و $(S - G^2)$ كانت الأكثر كفاءة لمطابقة النماذج مع البيانات ذات التنسيقات المختلطة. وأجرى زانغ وآخرون (Zhang et al., 2018) دراسة هدفت إلى توسيع نطاق استخدام مؤشرات المطابقة المستندة إلى مربع كاي، وأظهرت النتائج أن مؤشر $(S - \chi^2)$ كان الأفضل أداءً. وأجرى زانغ ووانغ (Zhang & Wang, 2022) دراسة هدفت إلى معالجة مشكلة البيانات المفقودة، وأشار الباحثان إلى أن المؤشرين $(S - \chi^2)$ و $(S - G^2)$ ، لا يستطيعان التعامل مع البيانات الناقصة بشكل مباشر بسبب اعتمادهما على الدرجات الكلية المكتملة، ولحل هذه المشكلة، اقترح الباحثان نسجاً معدلة من هذين المؤشرين $(M_{impute} - \chi^2)$ و $(M_{impute} - G^2)$ تعتمد على طرق تعويض البيانات الناقصة (Imputation Methods) بدلاً من استخدام الدرجات الكلية الملاحظة، وأظهرت النتائج أن المؤشرات المعدلة تعطي نتائج مكافئة للمؤشرات الأصلية عندما تكون البيانات مكتملة، لكنها تقدم تحسينات واضحة في حالات البيانات الناقصة.

وبتفحص الدراسات السابقة التي تناولت مؤشرات مطابقة الفقرة لنماذج (IRT)، يتبين أنه لا يوجد أي دراسة عمدت إلى تقييم المؤشرات المعدلة وفقاً لنماذج نظرية استجابة الفقرة متعددة التدرج (Polytomous)، ولذلك تم تقديم هذه الدراسة لتقييم مدى فعالية المؤشر المعدل $(M_{impute} - \chi^2)$ في التعامل مع البيانات المفقودة وفقاً لنموذج التقدير الجزئي (PCM)، وذلك بمقارنته بالمؤشر الأصلي $(S - \chi^2)$.

مشكلة الدراسة وأسئلتها

تعد مطابقة الفقرات للنماذج المفترضة في نظرية استجابة الفقرة (IRT) من الأسس المهمة في بناء الاختبارات وتحليلها، وتستخدم مؤشرات المطابقة التقليدية مثل: $(S - \chi^2)$ و $(S - G^2)$ ، بشكل شائع لتقييم مدى تطابق فقرات الاختبار مع النماذج الرياضية المفترضة، وتعتمد هذه المؤشرات على الدرجات الكاملة للأفراد، مما يجعلها عرضة للتأثر بالبيانات المفقودة، ويؤدي ذلك إلى زيادة التحيز في التقديرات، وضعف قدرة المؤشرات التقليدية على الكشف عن الفقرات غير المطابقة، وانخفاض موثوقية النتائج. بناءً على ذلك، ظهرت الحاجة إلى

تطوير مؤشرات مطابقة فقرة معدلة مثل: $(M_{impute} - X_i^2)$ و $(M_{impute} - G_i^2)$ ، والتي تعتمد على تعويض البيانات المفقودة باستخدام تقنيات تعويض متنوعة، مثل: طريقة التعويض الفردي (Single Imputation)، والتي تم اعتمادها في هذه الدراسة. حيث إن الهدف من المؤشرات المعدلة هو معالجة المشكلات المرتبطة بالبيانات غير المكتملة، وتحسين دقة التقديرات واستقرارها، وبالتالي جاءت هذه الدراسة لقياس مدى كفاءة المؤشرات المعدلة مقارنة بالتقليدية من خلال دراسة أثر ثلاثة عوامل: طول الاختبار، حجم العينة، ونسبة البيانات المفقودة.

أسئلة الدراسة

١. ما أثر طول الاختبار وحجم العينة ونسبة البيانات المفقودة على الخطأ من النوع الأول (α) بواسطة مؤشر المطابقة التقليدي للفقرة ومؤشر المطابقة المعدل وفقاً لنموذج التقدير الجزئي (PCM)؟
٢. ما أثر طول الاختبار وحجم العينة ونسبة البيانات المفقودة على قوة الاختبار بواسطة مؤشر المطابقة التقليدي للفقرة ومؤشر المطابقة المعدل وفقاً لنموذج التقدير الجزئي (PCM)؟

أهداف وأهمية الدراسة

أهداف الدراسة

هدفت هذه الدراسة إلى قياس مدى فاعلية كل من مؤشرات مطابقة الفقرة التقليدية والمعدلة في الكشف عن الفقرات المطابقة وغير المطابقة للنموذج الرياضي المستخدم، في ظل فقدان نسب معينة من البيانات، واختلاف طول الاختبار وحجم العينة.

أهمية الدراسة

تكمن أهمية هذه الدراسة من الحاجة المتزايدة إلى أدوات تحليل دقيقة لمطابقة الفقرات في سياق نظرية استجابة الفقرة (IRT)، وتعد مشكلة البيانات غير المكتملة تحديًا كبيرًا يؤثر على دقة المؤشرات التقليدية مثل: $(S - \chi^2)$ ، مما يجعل النتائج المستخلصة من التحليل غير موثوقة، وتتمثل أهمية الدراسة في اختبار مدى تأثير وجود نسب معينة من البيانات المفقودة واختلاف طول الاختبار وحجم العينة على دقة مؤشرات مطابقة الفقرة المعدلة، مثل: $(M_{impute} - \chi^2)$ ، والتي تعتمد على تعويض البيانات المفقودة قبل حساب المؤشر. وتكمن الأهمية التطبيقية للدراسة في أنها تقدم دليلًا عمليًا حول فاعلية المؤشر المعدل، والتي تعتمد على تعويض البيانات المفقودة، مما يساهم في تحسين إجراءات بناء الاختبارات في تحديد المؤشرات الأنسب التي تحافظ على دقة التقدير الإحصائي تحت ظروف اختبار مختلفة.

التعريفات الإجرائية

مطابقة الفقرة (Item Fit)

هي عملية تحليل مدى توافق استجابات الأفراد على فقرة معينة مع النموذج النظري المستخدم في نظرية الاستجابة للفقرة.

الخطأ من النوع الأول (Type I Error)

يُشير إلى احتمال رفض الفرضية الصفرية (H_0) عندما تكون صحيحة، في سياق الدراسة، يعني هذا اعتبار الفقرة غير مطابقة للنموذج عند مستوى الدلالة ($\alpha = 0.05$) على الرغم من أنها مطابقة.

مؤشر مطابقة الفقرة التقليدي $(S - \chi^2)$

أحد المؤشرات الإحصائية المستخدمة في إطار نظرية استجابة الفقرة (IRT)، ويُستخدم لتقييم مدى مطابقة الفقرة للنموذج الرياضي المفترض دون إجراء أي معالجة للبيانات المفقودة.

مؤشر مطابقة الفقرة المعدل $(M_{impute} - \chi^2)$

مؤشر تم تطويره اعتمادًا على مؤشر مطابقة الفقرة التقليدي، ويأخذ في الحسبان معالجة الفقد في البيانات، بهدف تحسين دقة التقدير والحد من أثر البيانات المفقودة.

قوة الاختبار

احتمال رفض الفرضية الصفرية عندما تكون خاطئة، وحسب الدراسة الحالية فهي احتمالية اعتبار الفقرة غير مطابقة عندما تكون فعلاً غير مطابقة.

نسبة البيانات المفقودة

وتشير إلى الفقرات التي تركت بلا إجابة، والتي قد تنجم عن أسباب تتعلق بتصميم الاختبار أو بالمتن، مثل عدم الفهم أو التعب أو الامتناع عن الإجابة أو انتهاء الوقت.

محددات الدراسة

- اقتصرَت الدراسة على نموذج التقدير الجزئي (PCM).
- اقتصرَت الدراسة على المؤشر التقليدي ($\chi^2 - S$) والمؤشر المعدل ($M_{impute} - \chi^2$).
- اعتمدت الدراسة على بيانات محاكاة مصممة خصيصاً لدراسة العوامل الثلاثة (طول الاختبار، حجم العينة، ونسبة البيانات المفقودة).

الطريقة والإجراءات

اتبعت الدراسة الحالية التصميم التجريبي ($3 \times 3 \times 3$)، اعتماداً على بيانات محاكاة، تم تصميمها وفق مجموعة من الشروط التجريبية التي تنوعت في ثلاثة عوامل مستقلة تصنيفية، ولكل عامل ثلاثة مستويات مختلفة: طول الاختبار (عدد الفقرات)، ($20, 40, 60$) فقرة، وحجم العينة ($250, 500, 1000$) مفحوص، ونسبة البيانات المفقودة ($1\%, 10\%, 20\%$)، وتم تكرار توليد البيانات مئة مرة لكل حالة للحصول على تقديرات مستقرة لمعدلات التحديد الخاطئ (FPRs) (false positive rates) (الخطأ من النوع الأول) ومعدلات الكشف الصحيح (CDRs) (correct detection rates) (قوة الاختبار). وللإجابة عن السؤال الأول، جرى توليد بيانات محاكاة حسب نموذج التقدير الجزئي (PCM) باستخدام برمجية (WinGen)، واستُخدمت برمجية (R) لحذف النسب المطلوبة من البيانات في كل حالة وبشكل عشوائي، وجرى حساب قيمة المؤشر التقليدي بتجاهل الأفراد الذين لديهم إجابات مفقودة، وحساب قيمة المؤشر المعدل بعد تعويض القيم المفقودة بطريقة التعويض الفردي (Single Imputation)، وحساب قيم (P-value) للمؤشرين للحكم على الفقرة بأنها مطابقة أو غير مطابقة للنموذج، وأيضاً وبالاعتماد على نفس البرمجية (R) تم حساب الخطأ من النوع الأول (Type-1 Error) لكل فقرة، ولإيجاد معدل الخطأ من النوع الأول لكل فقرة، قام الباحثان بحساب عدد المرات التي كانت القيمة الاحتمالية لها أقل من (0.05) وقسمتها على (100)، وهو العدد الكلي لمرات الحكم على مطابقة الفقرة، وباستخدام حزمة البيانات الإحصائية (SPSS) تم تحليل البيانات الناتجة باستخدام تحليل التباين الثلاثي للقياسات المتكررة (Three-Way Repeated Measures ANOVA)، لكون تصميم الدراسة يتضمن ثلاثة عوامل مستقلة تصنيفية، والمتغير التابع (الخطأ من النوع الأول) كمي تم قياسه باستخدام مؤشرين طبقاً على نفس البيانات، مما يدل على استخدام قياسات متكررة داخل الحالة الواحدة. وتم التحقق من افتراضات تحليل التباين الثلاثي للقياسات المتكررة من خلال التحقق من التوزيع الطبيعي للقيم داخل كل مجموعة فرعية، بالإضافة إلى افتراض استقلالية المشاهدات، الذي تم ضمانه من خلال التصميم التجريبي للدراسة. ونظراً لأن العامل المكرر اقتصر على مستويين فقط، فإن افتراض الكروية (Sphericity) متحقق تلقائياً دون الحاجة لاختباره.

ولتحديد دلالة الفروق بين المؤشرين عند كل مستوى من مستويات العامل المدروس، تم إجراء اختبار t لعينتين مرتبطتين (Paired Samples t-test)، ولخص الباحثان نتائج معدلات الخطأ من النوع الأول من خلال عرض الجداول الإحصائية والرسوم البيانية التي توفر صورة بصرية واضحة لنتائج المقارنات. وأما بالنسبة للسؤال الثاني، قام الباحثان بالإجابة عليه من خلال توليد بيانات باستخدام نموذج مختلف عن نموذج التقدير الجزئي، لضمان عدم مطابقة الفقرات، حيث تم استخدام نموذج الإجابة المتدرجة (GRM)، مع تكرار المحاكاة (100) مرة لكل حالة، ومن ثم حساب قوة الاختبار لكل فقرة بنفس طريقة حساب الخطأ من النوع الأول لكلا المؤشرين، وتم استخدام نفس طرق التحليل المتبعة في الإجابة على السؤال الأول.

نتائج الدراسة ومناقشتها

النتائج المتعلقة بسؤال الدراسة الأول: وللإجابة على السؤال الأول تم استخدام تحليل التباين الثلاثي للقياسات المتكررة، والجدول (١) يبين قيم الأوساط الحسابية للخطأ من النوع الأول لجميع متغيرات الدراسة.

جدول (١): قيم الأوساط الحسابية للخطأ من النوع الأول للمؤشرين التقليدي والمعدل تبعاً لمتغيرات الدراسة.

طول الاختبار	نسبة البيانات المفقودة	حجم العينة			
		٢٥٠	٥٠٠	١٠٠٠	المعدل التقليدي
٢٠	٪١	٠,٠٥٨	٠,٠٥٤	٠,٠٥٨	٠,٠٥٣
٢٠	٪١٠	٠,١٦٦	٠,١٦١	٠,١٧٦	٠,٠٦٤
٢٠	٪٢٠	٠,٣٨٠	٠,٢٧٧	٠,٣٦٩	٠,١١٥
٤٠	٪١	٠,٠٦٠	٠,٠٥٩	٠,٠٥١	٠,٠٦٨
٤٠	٪١٠	٠,٢٦٩	٠,٢٢٩	٠,٢٤١	٠,٠٥٦
٤٠	٪٢٠	٠,٥٥٢	٠,٥٠٥	٠,٥٥٤	٠,٠٦٨
٦٠	٪١	٠,٠٧٢	٠,٠٨٦	٠,٠٨٢	٠,٠٦٩
٦٠	٪١٠	٠,٧١٢	٠,٥٦٧	٠,١٣٥	٠,٠٨٩
٦٠	٪٢٠	٠,٩٠٣	٠,٨٦٥	٠,١٢٢	٠,٠٩٨

من خلال استعراض القيم الواردة في الجدول ١، يتضح أن هناك فروقاً واضحة في أداء المؤشرين، فبشكل عام، يظهر المؤشر المعدل $(M_{impute} - \chi^2)$ معدلات خطأ أقل وأكثر استقراراً مقارنةً بالمؤشر التقليدي $(S - \chi^2)$ خصوصاً مع ارتفاع نسبة البيانات المفقودة وزيادة طول الاختبار ونقصان حجم العينة، فعندما كانت نسبة البيانات المفقودة منخفضة (١٪)، كان الفرق في معدل الخطأ بين المؤشرين محدوداً وقليلًا نسبيًا، حيث تفوق المؤشر التقليدي على المؤشر المعدل في بعض الحالات، ولكن مع ارتفاع نسبة البيانات المفقودة ظهر انهيار واضح في قيم الخطأ من النوع الأول للمؤشر التقليدي وخصوصاً عندما وصلت نسبة البيانات المفقودة إلى (٢٠٪). ولمعرفة دلالة الفرق بين المؤشرين باختلاف مستويات العوامل المستقلة، تم إجراء تحليل التباين الثلاثي للقياسات المتكررة بتكرار القياس على عامل واحد (العامل التابع) وهو الأوساط الحسابية لقيم الخطأ من النوع الأول لكلا المؤشرين التقليدي والمعدل. ويبين الجدول (٢) نتائج التحليل.

جدول (٢): نتائج تحليل التباين للقياسات المتكررة لقيم الأوساط الحسابية للخطأ من النوع الأول وفقاً للمؤشر المستخدم وطول الاختبار وحجم العينة ونسبة البيانات المفقودة.

مصدر التباين	مجموع المربعات	درجة الحرية	متوسط المربعات	قيمة اختبار F	الدلالة الإحصائية Sig.	الدلالة العملية η^2
طول الاختبار	١١,٩٢٠	٢	٥,٩٦٠	١٢١٧,٠٠١	٠,٠٠٠**	٠,٧٠
حجم العينة	٠,٢١٠	٢	٠,١٠٥	٢١,٤٤٤	٠,٠٠٠**	٠,٠٤
نسبة البيانات المفقودة	٢٠,٠٤٠	٢	١٠,٠٢٠	٢٠٤٥,٩٨٦	٠,٠٠٠**	٠,٨٠
طول الاختبار × حجم العينة	٠,٢٣١	٤	٠,٠٥٨	١١,٧٧٩	٠,٠٠٠**	٠,٠٤
طول الاختبار × نسبة البيانات المفقودة	٤,٦٩٩	٤	١,١٧٥	٢٣٩,٨٩١	٠,٠٠٠**	٠,٤٨
حجم العينة × نسبة البيانات المفقودة	٠,١٢٥	٤	٠,٠٣١	٦,٣٦٥	٠,٠٠٠**	٠,٠٢
طول الاختبار × حجم العينة × نسبة البيانات المفقودة	٠,٢٨٨	٨	٠,٠٣٦	٧,٣٣٩	٠,٠٠٠**	٠,٠٥
الخطأ	٥,١٥٧	١٠٥٣	٠,٠٠٥		٠,٠٠٠**	
الكلي	٤٢,٦٦٩	١٠٧٩			٠,٠٠٠**	
نوع المؤشر	٢٦,٥٣٧	١	٢٦,٥٣٧	٥٥٦,٠٢١	٠,٠٠٠**	٠,٨٤

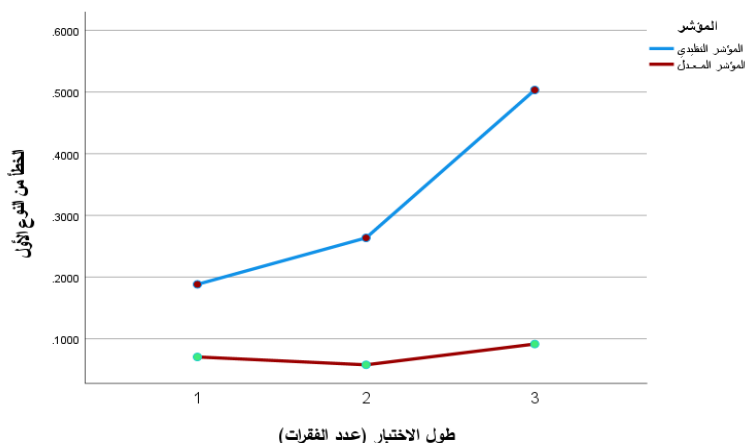
بين المجموعات

نوع المؤشر × طول الاختبار	٧,٩٤٨	٢	٣,٩٧٤	٨٣٢,٦٦٠	٠,٠٠٠**	٠,٦١
نوع المؤشر × حجم العينة	٠,٣٣٥	٢	٠,١٦٨	٣٥,١٤٥	٠,٠٠٠**	٠,٠٦
نوع المؤشر × نسبة البيانات المفقودة	١٧,٣٦٦	٢	٨,٦٨٣	١٨١٩,٣٦٢	٠,٠٠٠**	٠,٧٨
نوع المؤشر × طول الاختبار × حجم العينة	٠,٢٣١	٤	٠,٠٥٨	١٢,١٠٧	٠,٠٠٠**	٠,٠٤
نوع المؤشر × طول الاختبار × نسبة البيانات المفقودة	٤,٤٨٧	٤	١,١٢٢	٢٣٥,٠٤٥	٠,٠٠٠**	٠,٤٧
نوع المؤشر × حجم العينة × نسبة البيانات المفقودة	٠,١٧٦	٤	٠,٠٤٤	٩,٢٢٥	٠,٠٠٠**	٠,٠٣
نوع المؤشر × طول الاختبار × حجم العينة × نسبة البيانات المفقودة	٠,٣٤٤	٨	٠,٠٤٣	٨,٩٩٩	٠,٠٠٠**	٠,٠٦
الخطأ	٥,٠٢٦	١٠٥٣	٠,٠٠٥			
الكلي	٦٢,٤٥	١٠٨٠				

داخل المجموعات

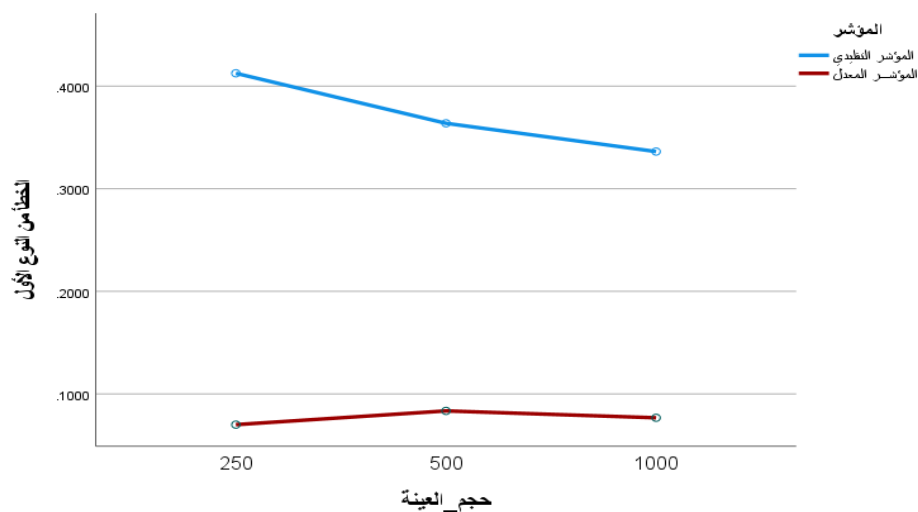
** وتعني دالة إحصائية عند مستوى الدلالة ($\alpha = 0,05$)

أظهرت نتائج الجدول ٢ وجود فروق دالة إحصائية لجميع العوامل المدروسة على قيم الخطأ من النوع الأول، ومن حيث الدلالة العملية فقد تبين أن نسبة البيانات المفقودة كان لها الأثر الأكبر ($\eta^2 = 0,80$)، تليها طول الاختبار ($\eta^2 = 0,70$)، ومن ثم حجم العينة بأثر ضعيف ($\eta^2 = 0,04$)، وذلك حسب معايير كوهين (Cohen, 1988)، حيث أشار إلى أن قوة تأثير العامل المستقل في العامل التابع تكون ضعيفة إذا فسر حوالي ١٪ من التباين الكلي، ومتوسطة إذا فسر ٦٪، وكبيرة إذا فسر أكثر من ١٤٪، أما بالنسبة للتفاعلات، فكانت جميعها دالة إحصائية، فقد كان أبرزها التفاعل بين طول الاختبار ونسبة البيانات المفقودة ($\eta^2 = 0,48$)، مما يدل على أن الجمع بين هذين العاملين يزيد من تأثيرهما على معدل الخطأ من النوع الأول، في المقابل، كانت بقية التفاعلات ذات دلالة إحصائية لكنها بحجم أثر ضعيف. وفيما يتعلق بالتفاعلات، فقد بينت النتائج أن جميع التفاعلات بين العوامل المستقلة كانت دالة إحصائية. كما أظهرت النتائج أن تفاعل نوع المؤشر مع طول الاختبار وحجم العينة ونسبة البيانات المفقودة في جميع التفاعلات كان دالاً إحصائياً وعملياً، مما يشير إلى أن أداء المؤشرين يختلف باختلاف مستوى العامل المستقل، وأيضاً يختلف أداء المؤشرين عند تفاعل نوع المؤشر مع العوامل المستقلة، مع تفاوت في الدلالة العملية بين الحالات. كما أظهرت النتائج فعالية المؤشر المعدل خاصة في الحالات التي تحتوي على نسب عالية من البيانات غير المكتملة، مقارنة بالمؤشر التقليدي. والشكل (١) يوضح أثر التفاعل بين طول الاختبار ونوع مؤشر مطابقة الفقرة المستخدم.



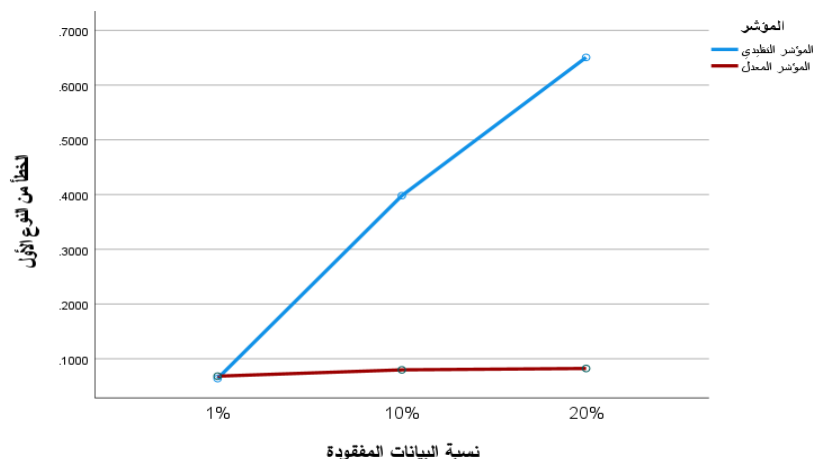
الشكل (١): الرسم البياني للتفاعل بين نوعي مؤشر المطابقة التقليدي والمعدل وطول الاختبار في الخطأ من النوع الأول

ويُظهر الشكل (١) تفاعلاً واضحاً بين نوع المؤشر وطول الاختبار في معدلات الخطأ من النوع الأول، حيث تزداد قيمة الخطأ بشكل ملحوظ مع زيادة عدد الفقرات عند استخدام المؤشر التقليدي، بينما يظل المؤشر المعدل ثابتاً نسبياً ويُظهر مقاومة أكبر لتأثير طول الاختبار، مما يعكس تفوقه في الحفاظ على دقة الاختبار عند زيادة عدد الفقرات. ويبين الشكل (٢) التفاعل بين نوع المؤشر المستخدم (التقليدي والمعدل) وأحجام العينات الثلاثة في تحديد الخطأ من النوع الأول.



الشكل (٢): الرسم البياني للتفاعل بين نوعي مؤشر المطابقة التقليدي والمعدل وحجم العينة في الخطأ من النوع الأول

ويتضح من خلال الشكل (٢) البياني أن متوسطات الخطأ من النوع الأول كانت مرتفعة نسبياً للمؤشر التقليدي مقارنة بالمؤشر المعدل في جميع أحجام العينات، ويلاحظ بوضوح أن المؤشر التقليدي أظهر انخفاضاً تدريجياً في الخطأ من النوع الأول مع ازدياد حجم العينة من (٢٥٠) إلى (١٠٠٠)، في حين بقي المؤشر المعدل منخفضاً ومستقرًا إلى حد كبير عبر جميع أحجام العينات، وتعكس هذه النتيجة استقراراً وفعالية أكبر للمؤشر المعدل في الحفاظ على معدلات منخفضة من الخطأ من النوع الأول. والشكل (٣) يوضح التفاعل بين معدل الخطأ من النوع الأول في نوع المؤشر المستخدم (التقليدي والمعدل) ونسبة البيانات المفقودة وذلك بهدف الكشف عن حساسية كل مؤشر للتغير في نسب البيانات المفقودة.



الشكل (٣): الرسم البياني للتفاعل بين نوعي مؤشر المطابقة التقليدي والمعدل ونسبة البيانات المفقودة في الخطأ من النوع الأول

ويتضح من الشكل (٣) البياني أن قيم متوسطات الخطأ من النوع الأول كانت منخفضة ومستقرة نسبياً للمؤشر المعدل عند مستويات نسبة البيانات المفقودة الثلاث، مما يدل على مقاومته وممانعته للتغيرات الناتجة عن زيادة البيانات المفقودة، في المقابل، يُلاحظ تقارب أداء المؤشر التقليدي من المؤشر المعدل عند نسب الفقد المنخفضة (١٪)، ومع زيادة نسبة البيانات المفقودة ترتفع معدلات الخطأ من النوع الأول بشكل حاد، مما يعكس حساسية هذا المؤشر وضعف أدائه في التعامل مع بيانات فيها الكثير من الاستجابات المفقودة.

وبعد أن أظهرت نتائج تحليل التباين أن طول الاختبار وحجم العينة ونسبة البيانات المفقودة كان لها أثر دالٌّ إحصائيًا، تم إجراء المقارنات البعدية باستخدام اختبار توكي (Tukey HSD) للكشف عن الفروق بين متوسطات الخطأ من النوع الأول عبر مستويات العوامل المستقلة الثلاثة. ويوضح الجدول (٣) دلالة الفرق في متوسط الخطأ من النوع الأول بين كل زوج من مستويات طول الاختبار.

جدول (٣): نتائج اختبار المقارنات البعدية (Tukey HSD) بين مستويات طول الاختبار في متوسط الخطأ من النوع الأول

أطوال الاختبار	الفرق في المتوسط	الخطأ المعياري	الدلالة الإحصائية
(٢٠، ٤٠) فقرة	-٠,٣١٣	٠,٠٠٤٥	< ٠,٠٠١**
(٢٠، ٦٠) فقرة	-٠,١٦٨٠	٠,٠٠٤٣	< ٠,٠٠١**
(٦٠، ٤٠) فقرة	-٠,١٣٦٧	٠,٠٠٣٣	< ٠,٠٠١**

** وتعني دالة إحصائية عند مستوى الدلالة ($\alpha = 0,05$)

وأظهرت نتائج الجدول (٣) أن الفرق بين المتوسطات في الخطأ من النوع الأول كان دالًّا إحصائيًا بين جميع أزواج مستويات طول الاختبار. والإشارة السالبة تعني أن معدل الخطأ من النوع الأول يزداد بزيادة طول الاختبار. والجدول (٤) يوضح دلالة الفرق في متوسط الخطأ من النوع الأول بين كل زوج من مستويات حجم العينة.

جدول (٤): نتائج اختبار المقارنات البعدية (Tukey HSD) بين مستويات حجم العينة في متوسط الخطأ من النوع الأول.

أحجام العينات	الفرق في المتوسط	الخطأ المعياري	الدلالة الإحصائية
(٢٥٠، ٥٠٠)	+٠,٠١٧٧	٠,٠٠٣٧	< ٠,٠٠١**
(٢٥٠، ١٠٠٠)	+٠,٠٣٤٨	٠,٠٠٣٧	< ٠,٠٠١**
(٥٠٠، ١٠٠٠)	+٠,٠١٧١	٠,٠٠٣٧	< ٠,٠٠١**

** وتعني دالة إحصائية عند مستوى الدلالة ($\alpha = 0,05$)

وبلاحظ من نتائج الجدول (٤) أن الفرق بين المتوسطات في الخطأ من النوع الأول كان دالاً إحصائياً بين جميع أزواج مستويات حجم العينة، كما يلاحظ من الجدول أن معدل الخطأ من النوع الأول يقل بزيادة حجم العينة بدليل أن الفرق في المتوسط بين الأزواج موجب. والجدول (٥) يوضح دلالة الفرق في متوسط الخطأ من النوع الأول بين كل زوج من مستويات نسبة البيانات المفقودة.

جدول (٥): نتائج اختبار المقارنات البعدية (Tukey HSD) بين مستويات نسبة البيانات المفقودة في متوسط الخطأ من النوع الأول

نسب البيانات المفقودة	الفرق في المتوسط	الخطأ المعياري	الدلالة الإحصائية
(١٠٪، ١٪)	-٠,١٧٢٨	٠,٠٠٣٦٩	٠,٠٠٠**
(١٠٪، ٢٠٪)	-٠,٣٠٠٤	٠,٠٠٣٦٩	٠,٠٠٠**
(١٠٪، ٢٠٪)	-٠,١٢٧٦	٠,٠٠٣٦٩	٠,٠٠٠**

** وتعني دالة إحصائية عند مستوى الدلالة ($\alpha = 0.05$)

ويوضح الجدول (٥) أن الفرق بين المتوسطات في الخطأ من النوع الأول كان دالاً إحصائياً بين جميع أزواج، نسب البيانات المفقودة، كما يلاحظ من الجدول أن معدل الخطأ من النوع الأول يزداد بزيادة نسبة البيانات المفقودة بدليل أن الفرق في المتوسط بين الأزواج سالب.

ولتحديد الدلالة الإحصائية للفروق بين معدلات الخطأ من النوع الأول لكلا المؤشرين (التقليدي والمعدل) عند كل طول من أطوال الاختبار (٢٠، ٤٠، ٦٠) فقرة، قام الباحثان بإجراء اختبار t لعينتين مرتبطتين، والجدول (٦) يوضح نتائج التحليل.

جدول (٦): نتائج اختبار t لعينتين مرتبطتين للفروق بين مؤشري مطابقة الفقرة عند كل طول من أطوال الاختبار

طول الاختبار	نوع مؤشر مطابقة الفقرة	المتوسط	درجات الحرية	قيمة اختبار t	مستوى الدلالة
٢٠ فقرة	التقليدي	٠,١٨٨	١٧٩	١٣,٩	٠,٠١**
	المعدل	٠,٠٧١			
٤٠ فقرة	التقليدي	٠,٢٦٤	٣٥٩	١٨,٦٨٥	٠,٠١**
	المعدل	٠,٠٥٨			
٦٠ فقرة	التقليدي	٠,٥٠٣	٥٣٩	٢٧,٨٤	٠,٠١**
	المعدل	٠,٠٩٢			

** وتعني دالة إحصائية عند مستوى الدلالة ($\alpha = 0.05$)

وبلاحظ من الجدول (٦) وجود فرق دال إحصائياً بين المؤشرين التقليدي والمعدل عند جميع أطوال الاختبار، حيث ترتفع قيمة الخطأ من النوع الأول بشكل ملحوظ مع ازدياد عدد الفقرات عند استخدام المؤشر التقليدي مقارنة بالمؤشر المعدل، إذ بلغ متوسط الخطأ ٠,٥٠٣ للمؤشر التقليدي مقابل ٠,٠٩٢ فقط للمؤشر المعدل عند طول ٦٠ فقرة، مما يعكس ضعف المؤشر التقليدي في الحفاظ على دقة التقدير مع ازدياد طول الاختبار، ويزر كفاءة المؤشر المعدل في تقليل الخطأ من النوع الأول.

ولتحديد الدلالة الإحصائية للفروق بين معدلات الخطأ من النوع الأول لكلا المؤشرين (التقليدي والمعدل) عند كل مستوى من مستويات حجم العينة (٢٥٠، ٥٠٠، ١٠٠٠) فرد، جرى استخدام اختبار t لعينتين مرتبطتين (Paired Samples t-test)، والجدول (٧) يوضح نتائج هذا التحليل.

جدول (٧): نتائج اختبار t لعينتين مرتبطتين للفروق بين مؤشري مطابقة الفقرة عند كل مستوى من أحجام العينة.

حجم العينة	نوع مؤشر مطابقة الفقرة	المتوسط	درجات الحرية	قيمة اختبار t	مستوى الدلالة
٢٥٠	التقليدي	٠,٤١٢	٣٥٩	١٩,٣٣٥	٠,٠١**
	المعدل	٠,٠٧٠			
٥٠٠	التقليدي	٠,٣٦١	٣٥٩	١٨,٢٧٦	٠,٠١**
	المعدل	٠,٠٨٤			
١٠٠٠	التقليدي	٠,٣٣٦	٣٥٩	١٨,٥٦٢	٠,٠١**
	المعدل				

المعدل	٠,٠٧٧
--------	-------

**** وتعني دالة إحصائية عند مستوى الدلالة ($\alpha = 0.05$)**

وبلاحظ من نتائج الجدول ٧ وجود فرق دال إحصائياً بين مؤشري مطابقة الفقرة التقليدي والمعدل عند جميع مستويات حجم العينة، وعلى الرغم من أن المؤشر التقليدي أظهر انخفاضاً تدريجياً مع زيادة حجم العينة، إلا أن قيم الخطأ من النوع الأول له بقيت مرتفعة بشكل واضح.

ويُظهر الجدول (٨) نتائج اختبار t لعينتين مرتبطتين لدراسة الفروق بين معدلات الخطأ من النوع الأول لكل من مؤشري مطابقة الفقرة التقليدي والمعدل للتحقق من دلالة الفرق بين أداء المؤشرين عند كل نسبة من نسب البيانات المفقودة.

جدول (٨): نتائج اختبار t لعينتين مرتبطتين للفروق بين مؤشري مطابقة الفقرة عند كل نسبة من نسب البيانات المفقودة.

نسبة البيانات المفقودة	نوع مؤشر مطابقة الفقرة	المتوسط	درجات الحرية	قيمة اختبار t	مستوى الدلالة
١%	التقليدي	٠,٠٦٤	٣٥٩	١,٢١٤	٠,٢٢٦
	المعدل	٠,٠٦٨			
١٠%	التقليدي	٠,٣٩٨	٣٥٩	٢٨,٠٩٥	٠,٠٠١**
	المعدل	٠,٠٨٠			
٢٠%	التقليدي	٠,٦٥١	٣٥٩	٤٥,٧٢٤	٠,٠٠١**
	المعدل	٠,٠٨٢			

**** وتعني دالة إحصائية عند مستوى الدلالة ($\alpha = 0.05$)**

وتشير نتائج اختبار t إلى عدم وجود فرق دال إحصائياً بين متوسطات المؤشرين التقليدي والمعدل عند مستوى البيانات المفقودة المنخفض (١%)، مع تفوق للمؤشر التقليدي على المؤشر المعدل عند هذه النسبة، في حين ظهرت فروق ذات دلالة إحصائية واضحة لصالح المؤشر المعدل عند مستويات نسب البيانات المفقودة الأعلى (١٠%، ٢٠%).

النتائج المتعلقة بسؤال الدراسة الثاني: وللإجابة على السؤال الثاني، جرى حساب قيم الأوساط الحسابية لقوة الاختبار لكلا المؤشرين، ويبين الجدول (٩): نتائج التحليل.

جدول (٩): قيم الأوساط الحسابية لقوة الاختبار للمؤشرين التقليدي والمعدل تبعاً لمتغيرات الدراسة.

طول الاختبار	نسبة البيانات المفقودة	حجم العينة			
		٢٥٠	٥٠٠	١٠٠٠	المعدل
٢٠	١%	٠,١٠٤	٠,٠٩٧	٠,١١٠	٠,١٢٥
٢٠	١٠%	٠,١٩٦	٠,١٤٦	٠,١٤١	٠,٣٥٤
٢٠	٢٠%	٠,٤٢٠	٠,٢٨٤	٠,٣٦٦	٠,٦٩٠
٤٠	١%	٠,٠٨١	٠,٠٦٦	٠,٢٠٩	٠,٤٦٨
٤٠	١٠%	٠,١٩٥	٠,٢١٨	٠,٤٧٢	٠,٦٤٨
٤٠	٢٠%	٠,٠٨٦	٠,١٧١	٠,٣٨١	٠,٧١٤
٦٠	١%	٠,٣٦٦	٠,٤١٧	٠,٦٥٤	٠,٤٨١
٦٠	١٠%	٠,٤٨٤	٠,٣٣٩	٠,٥٠١	٠,٤٥٠
٦٠	٢٠%	٠,٠٩٢	٠,٢٩٤	٠,١٤٢	٠,٢٨٨

وبلاحظ من نتائج الجدول (٩) وجود فروق واضحة في قيم الأوساط الحسابية لقوة الاختبار بين المؤشرين التقليدي والمعدل تبعاً لمتغيرات الدراسة الثلاثة، ومن خلال النظر إلى متوسطات قوة الاختبار في الجدول ٩ يتبين أن المؤشر المعدل تفوق بمعدلات الكشف الصحيح عن الفقرات غير المطابقة بأغلبية حالات الدراسة بشكل واضح لا سيما عند زيادة طول الاختبار وارتفاع نسب الفقد في البيانات (١٠%، ٢٠%)، مع بعض التفوق أحياناً للمؤشر التقليدي ($\chi^2 - S$)، كما تظهر بيانات الجدول أن قوة الاختبار لكل من المؤشرين تزداد بشكل عام مع زيادة

حجم العينة وتقل بزيادة طول الاختبار، مع بروز تفوق واضح للمؤشر المعدّل في الظروف الأكثر صعوبة (مثل الاختبار الطويل، وحجم العينة الصغير، ونسبة البيانات المفقودة المرتفعة).

ويبيّن الجدول (١٠) نتائج تحليل التباين الثلاثي للقياسات المتكررة، لقيم الأوساط الحسابية لقوة الاختبار وفقاً للمؤشر المستخدم وطول الاختبار، وحجم العينة، ونسبة البيانات المفقودة، ويهدف هذا التحليل إلى الكشف عن مدى تأثير كل متغير من هذه المتغيرات، والتفاعلات المحتملة فيما بينها، على قوة الاختبار للمؤشرين.

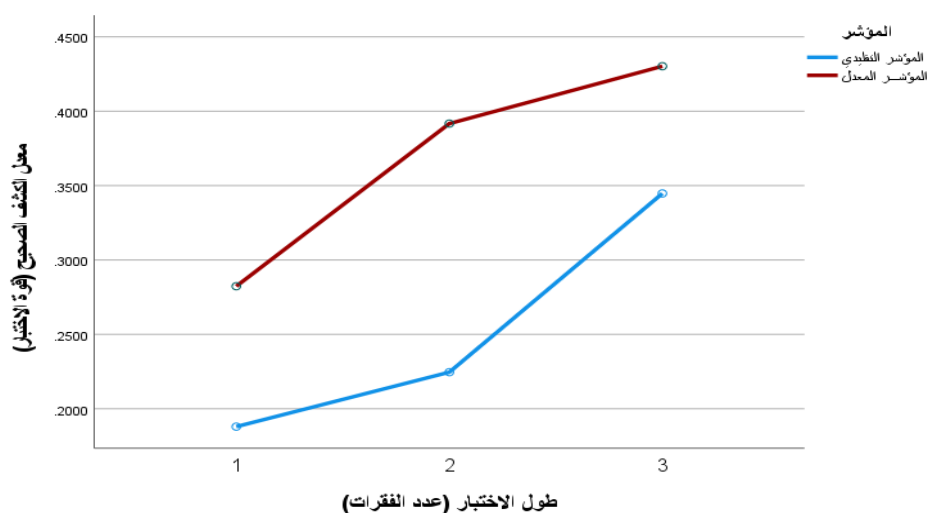
جدول (١٠): نتائج تحليل التباين للقياسات المتكررة لقيم الأوساط الحسابية لقوة الاختبار وفقاً للمؤشر المستخدم وطول الاختبار وحجم العينة ونسبة البيانات المفقودة.

الدلالة العملية η^2	الدلالة الإحصائية Sig.	قيمة اختبار F	متوسط المربعات	درجة الحرية	مجموع المربعات	مصدر التباين	
٠,٤٥	٠,٠٠٠**	٤٢٨,١٧١	٣,٤١٤	٢	٦,٨٢٨	حجم العينة	بين المجموعات
٠,٤٦	٠,٠٠٠**	٤٤٤,١٢٢	٣,٥٤١	٢	٧,٠٨٢	طول الاختبار	
٠,١٤	٠,٠٠٠**	٨٦,٩٩٨	٠,٦٩٤	٢	١,٣٨٧	نسبة البيانات المفقودة	
٠,٦٧	٠,٠٠٠**	٥٢٩,٢٩٦	٤,٢٢٠	٤	١٦,٨٨١	طول الاختبار × حجم العينة	
٠,٠٧	٠,٠٠٠**	١٩,٧٦٨	٠,١٥٨	٤	٠,٦٣٠	حجم العينة × نسبة البيانات المفقودة	
٠,٦٣	٠,٠٠٠**	٤٤٣,٩٢١	٣,٥٤٠	٤	١٤,١٥٨	طول الاختبار × نسبة البيانات المفقودة	
٠,٠٩	٠,٠٠٠**	١٢,٨٥٦	٠,١٠٣	٨	٠,٨٢٠	طول الاختبار × حجم العينة × نسبة البيانات المفقودة	داخل المجموعات
	٠,٠٠٠**		٠,٠٠٨	١٠٥٣	٨,٣٩٦	الخطأ	
	٠,٠٠٠**			١٠٧٩	٥٦,١٨٢	الكل	
٠,٤٤	٠,٠٠٠**	٨١١,٥١٦	٥,٩١٢	١	٥,٩١٢	نوع المؤشر	
٠,٢٠	٠,٠٠٠**	١٣٤,١٥٦	٠,٩٧٧	٢	١,٩٥٥	نوع المؤشر × حجم العينة	
٠,٠٩	٠,٠٠٠**	٥٢,٢٥٨	٠,٣٨١	٢	٠,٧٦١	نوع المؤشر × طول الاختبار	
٠,٣٠	٠,٠٠٠**	٢٢٤,١٤٥	١,٦٣٣	٢	٣,٢٦٦	نوع المؤشر × نسبة البيانات المفقودة	
٠,١٦	٠,٠٠٠**	٥١,٨١٣	٠,٣٧٧	٤	١,٥١٠	نوع المؤشر × طول الاختبار × حجم العينة	
٠,١٨	٠,٠٠٠**	٥٩,٢٤٤	٠,٤٣٢	٤	١,٧٢٦	نوع المؤشر × حجم العينة × نسبة البيانات المفقودة	
٠,١٩	٠,٠٠٠**	٦١,١٤٦	٠,٤٤٥	٤	١,٧٨٢	نوع المؤشر × طول الاختبار × نسبة البيانات المفقودة	

نوع المؤشر ×	٢,٠٧٢	٨	٠,٢٥٩	٣٥,٥٥٤	٠,٠٠٠**	٠,٢١
طول الاختبار ×						
حجم العينة ×						
نسبة البيانات						
المفقودة						
الخطأ	٧,٦٧١	١٠٥٣	٠,٠٠٧			
الكلي	٢٦,٦٥٥	١٠٨٠				

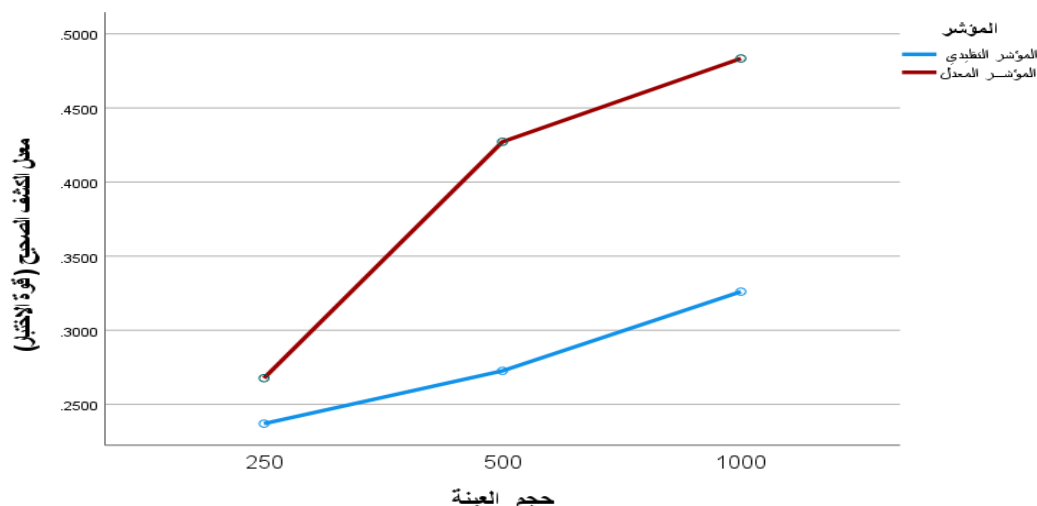
** وتعني دالة إحصائية عند مستوى الدلالة ($\alpha = 0.05$)

وبإلا حظ من الجدول (١٠) وجود فروق ذات دلالة إحصائية لجميع مصادر التباين الرئيسية (طول الاختبار، حجم العينة، نسبة البيانات المفقودة) وكذلك التفاعلات بين جميع العوامل، كما أظهرت النتائج أن تفاعل نوع المؤشر مع طول الاختبار وحجم العينة ونسبة البيانات المفقودة في جميع التفاعلات كان دالاً إحصائياً وعملياً، مما يشير إلى أن أداء المؤشرين يختلف باختلاف مستوى العامل المستقل، وأيضاً يختلف أداء المؤشرين عند تفاعل نوع المؤشر مع العوامل المستقلة، مع تفاوت في الدلالة العملية بين الحالات. كما أظهرت النتائج فعالية المؤشر المعدل خاصة في الحالات التي تحتوي على نسب عالية من البيانات غير المكتملة، مقارنة بالمؤشر التقليدي. والشكل (٤) يوضح أثر التفاعل بين طول الاختبار ونوع مؤشر مطابقة الفقرة المستخدم.



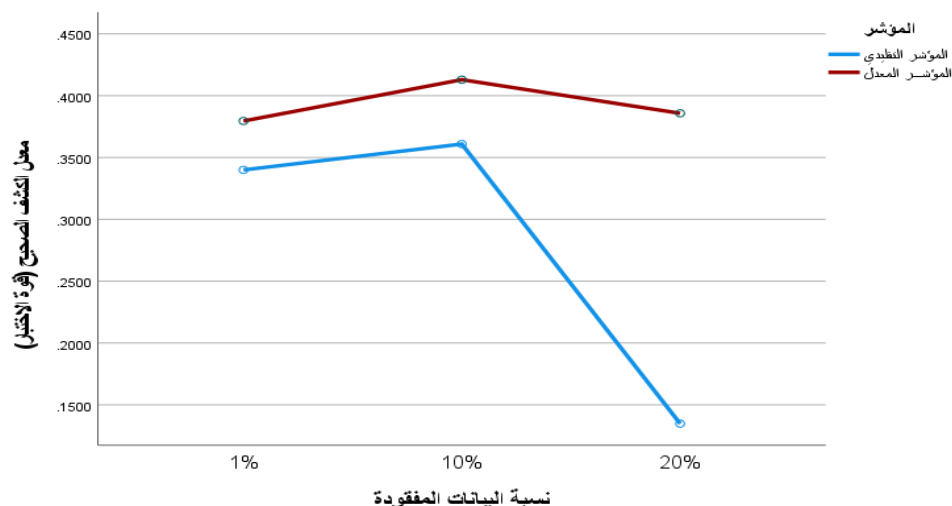
الشكل (٤): الرسم البياني للتفاعل بين نوعي مؤشر المطابقة التقليدي والمعدل وطول الاختبار في قوة الاختبار

ويتضح من الشكل (٤) أن معدلات الكشف الصحيح (قوة الاختبار) لكل من المؤشرين التقليدي والمعدل ترتفع مع زيادة طول الاختبار (من قصير إلى طويل)، ما يؤكد أن قوة الاختبار تتحسن كلما زاد عدد الفقرات الاختبار، وبإلا حظ أن المؤشر المعدل يتفوق عبر جميع مستويات طول الاختبار، إذ كانت معدلات الكشف الصحيح أعلى باستمرار مقارنة بالمؤشر التقليدي، وعلى الرغم من أن كلا المؤشرين أظهرتا تحسناً مع زيادة طول الاختبار، إلا أن المؤشر المعدل بقي الأكثر كفاءة في الكشف الصحيح عبر المستويات الثلاثة بشكل عام. ويبين الشكل (٥) التفاعل بين نوع المؤشر المستخدم (التقليدي والمعدل) وحجم العينة في تحديد معدل الكشف الصحيح (قوة الاختبار).



الشكل (٥): الرسم البياني للتفاعل بين نوعي مؤشر المطابقة التقليدي والمعدل وحجم العينة في قوة الاختبار

ويلاحظ من الشكل (٥) وجود تحسن في قوة الاختبار مع زيادة حجم العينة لكل من المؤشر التقليدي والمعدل، إلا أن هذا التحسن كان أكثر وضوحاً لدى المؤشر المعدل، إذ يُلاحظ أن المؤشر المعدل سجل معدلات أعلى للكشف الصحيح في جميع مستويات حجم العينة. في المقابل، أظهر المؤشر التقليدي تحسناً تدريجياً في قوة الاختبار، ولكن بشكل أبطأ مع زيادة حجم العينة. ويعرض الشكل (٦) التفاعل بين نوعي مؤشر مطابقة الفقرة (التقليدي والمعدل) ونسب البيانات المفقودة لتوضيح تأثير فقد البيانات على قوة الاختبار لكل من المؤشرين:



الشكل (٦): الرسم البياني للتفاعل بين نوعي مؤشر مطابقة الفقرة التقليدي والمعدل ونسبة البيانات المفقودة في قوة الاختبار

ويُظهر الشكل (٦) سلوك كل من المؤشرين التقليدي والمعدل عند تغير نسبة البيانات المفقودة، فقد أظهر المؤشر المعدل أداءً مستقرًا نسبيًا، حيث حافظ على قوة اختبار مرتفعة عند جميع نسب الفقد، مع انخفاض طفيف عند نسبة ٢٠٪، في حين يُلاحظ أن أداء المؤشر التقليدي كان أكثر تقلبًا، إذ انخفضت قوة اختبار بشكل حاد عند نسبة ٢٠٪ من البيانات المفقودة، بعد تحسن طفيف عند ١٠٪، وتدل هذه النتائج على أن المؤشر

المعدل أكثر مقاومة لتأثير البيانات المفقودة، ويُظهر ثباتاً أكبر في قوة الكشف عن الفترات غير المطابقة مقارنةً بالمؤشر التقليدي.

وبعد أن أظهرت نتائج تحليل التباين أن جميع العوامل المستقلة كان لها أثر دالّ إحصائيًا، تم إجراء المقارنات البعدية باستخدام اختبار توكي (Tukey HSD) للكشف عن الفروق بين متوسطات قوة الاختبار عبر مستويات العوامل المستقلة الثلاثة بشكل عام. ويوضح الجدول (١١) دلالة الفرق في متوسط الخطأ من النوع الأول بين كل زوج من مستويات طول الاختبار.

جدول (١١): نتائج اختبار المقارنات البعدية (Tukey HSD) بين مستويات طول الاختبار في متوسط الخطأ من النوع الأول

أطوال الاختبار	الفرق في المتوسط	الخطأ المعياري	الدلالة الإحصائية
(٢٠، ٤٠) فقرة	-٠,٠٧٣٠	٠,٠٠٥٧٦	٠,٠٠٠**
(٢٠، ٦٠) فقرة	-٠,١٥٢٣	٠,٠٠٥٤٣	٠,٠٠٠**
(٤٠، ٦٠) فقرة	-٠,٠٧٩٤	٠,٠٠٤٣٠	٠,٠٠٠**

** وتعني دالة إحصائية عند مستوى الدلالة ($\alpha = 0.05$)

وأظهرت نتائج الجدول (١١) أن الفرق بين متوسطات قوة الاختبار كان دالاً إحصائياً بين جميع أزواج مستويات طول الاختبار. والإشارة السالبة تعني أن معدل قوة الاختبار يزداد بزيادة طول الاختبار. والجدول (١٢) يوضح دلالة الفرق في متوسط الخطأ من النوع الأول بين كل زوج من مستويات حجم العينة.

جدول (١٢): نتائج اختبار المقارنات البعدية (Tukey HSD) بين مستويات حجم العينة في متوسط الخطأ من النوع الأول

أحجام العينات	الفرق في المتوسط	الخطأ المعياري	الدلالة الإحصائية
(٢٥٠، ٥٠٠)	٠,٠٩٧٥	٠,٠٠٤٧١	٠,٠٠٠**
(٢٥٠، ١٠٠٠)	٠,١٥٢٤	٠,٠٠٤٧١	٠,٠٠٠**
(٥٠٠، ١٠٠٠)	٠,٠٥٤٨	٠,٠٠٤٧١	٠,٠٠٠**

** وتعني دالة إحصائية عند مستوى الدلالة ($\alpha = 0.05$)

وبلاحظ من نتائج الجدول (١٢) أن الفرق بين متوسطات قوة الاختبار كان دالاً إحصائياً بين جميع أزواج مستويات حجم العينة. والجدول (١٣) يوضح دلالة الفرق في متوسط الخطأ من النوع الأول بين كل زوج من مستويات نسبة البيانات المفقودة.

جدول (١٣): نتائج اختبار المقارنات البعدية (Tukey HSD) بين مستويات نسبة البيانات المفقودة في متوسط الخطأ من النوع الأول

نسب البيانات المفقودة	الفرق في المتوسط	الخطأ المعياري	الدلالة الإحصائية
(١٠٪، ١٠٪)	-٠,٠٢٧٢	٠,٠٠٤٧١	٠,٠٠٠**
(١٠٪، ٢٠٪)	٠,٠٩٩٥	٠,٠٠٤٧١	٠,٠٠٠**
(١٠٪، ٢٠٪)	٠,١٢٦٦	٠,٠٠٤٧١	٠,٠٠٠**

** وتعني دالة إحصائية عند مستوى الدلالة ($\alpha = 0.05$)

ويوضح الجدول (١٣) أن الفرق بين متوسطات قوة الاختبار كان دالاً إحصائياً بين جميع أزواج نسب البيانات المفقودة.

ويظهر الجدول (١٤) نتائج اختبار t لعينتين مرتبطتين، وذلك للتحقق من دلالة الفرق في أداء المؤشرين (التقليدي والمعدل) في ضوء مستويات طول الاختبار.

جدول (١٤): نتائج اختبار t لعينتين مرتبطتين للفروق بين مؤشري مطابقة الفقرة عند كل طول من أطوال الاختبار

طول الاختبار	نوع مؤشر مطابقة الفقرة	المتوسط	درجات الحرية	قيمة اختبار t	مستوى الدلالة sig
٢٠ فقرة	التقليدي	٠,١٨٨	١٧٩	٦,٨٥٦	٠,٠٠١**
	المعدل	٠,٢٨٢			

٤٠	فقرة	التقليدي	٠,٢٢٥	٣٥٩	١٤,٣٨٢	٠,٠١	**
		المعدل	٠,٣٩٢				
٦٠	فقرة	التقليدي	٠,٣٤٥	٥٣٩	١٠,٦٤٨	٠,٠١	**
		المعدل	٠,٤٣٠				

** وتعني دالة إحصائية عند مستوى الدلالة ($\alpha = 0,05$)

وتشير نتائج الجدول (١٤) إلى وجود فروق ذات دلالة إحصائية بين مؤشر مطابقة الفقرة التقليدي (S) و χ^2 (-) ومؤشر مطابقة الفقرة المعدل ($M_{impute} - \chi^2_i$) عند جميع مستويات طول الاختبار الثلاثة (٢٠، ٤٠، ٦٠) فقرة، حيث كانت قيمة sig لجميع الأطوال أقل من (٠,٠١)، ويلاحظ من الجدول ١٢ أيضاً أن المتوسطات الخاصة بمؤشر مطابقة الفقرة المعدل كانت أعلى من نظيرتها للمؤشر التقليدي في جميع الأطوال.

ويظهر الجدول (١٥) نتائج اختبار t لعينتين مرتبطتين، وذلك لقياس الفرق بين مؤشري مطابقة الفقرة (التقليدي والمعدل) عند كل مستوى من مستويات حجم العينة (٢٥٠، ٥٠٠، ١٠٠٠) مفحوص، بهدف تحديد أي من مستويات حجم العينة يعطي فرقاً دالاً بين المؤشرين.

جدول (١٥): نتائج اختبار t لعينتين مرتبطتين للفروق بين مؤشري مطابقة الفقرة عند كل حجم من أحجام العينة

حجم العينة	نوع مؤشر مطابقة الفقرة	المتوسط	درجات الحرية	قيمة اختبار t	مستوى الدلالة
٢٥٠	التقليدي	٠,٢٣٧	٣٥٩	٣,٥٠٣	٠,٠١
	المعدل	٠,٢٦٨			
٥٠٠	التقليدي	٠,٢٧٣	٣٥٩	١٥,٧٤١	٠,٠١
	المعدل	٠,٤٢٧			
١٠٠٠	التقليدي	٠,٣٢٦	٣٥٩	١٣,٤٤١	٠,٠١
	المعدل	٠,٤٨٣			

** وتعني دالة إحصائية عند مستوى الدلالة ($\alpha = 0,05$)

ويلاحظ من الجدول (١٥) وجود فرق دال إحصائياً بين مؤشري مطابقة الفقرة التقليدي والمعدل عبر جميع مستويات حجم العينة، ويظهر من خلال عمود المتوسطات أن متوسطات المؤشر المعدل كانت أعلى من نظيرتها للمؤشر التقليدي في كل الحالات الثلاثة، ما يشير إلى أداء أفضل للمؤشر المعدل في الكشف عن الفقرات غير المطابقة، كما أن الفروق بين المؤشرين تزداد بوضوح مع زيادة حجم العينة.

وبين الجدول (١٦) نتائج اختبار t لعينتين مرتبطتين للتحقق من وجود فرق دال إحصائياً بين أداء المؤشرين عند كل نسبة من نسب البيانات المفقودة.

جدول (١٦): نتائج اختبار t لعينتين مرتبطتين للفروق بين مؤشري مطابقة الفقرة عند كل نسبة من نسب البيانات المفقودة

نسبة البيانات المفقودة	نوع مؤشر مطابقة الفقرة	المتوسط	درجات الحرية	قيمة اختبار t	مستوى الدلالة
١%	التقليدي	٠,٣٤	٣٥٩	٥,٧٣٧	٠,٠١
	المعدل	٠,٣٨			
١٠%	التقليدي	٠,٣٦١	٣٥٩	٥,١٥٣	٠,٠١
	المعدل	٠,٤١٣			
٢٠%	التقليدي	٠,١٣٥	٣٥٩	٢٣,٧٨١	٠,٠١
	المعدل	٠,٣٨٦			

** وتعني دالة إحصائية عند مستوى الدلالة ($\alpha = 0,05$)

ويلاحظ من نتائج الجدول (١٦) وجود فرق دال إحصائياً بين مؤشري مطابقة الفقرة التقليدي والمعدل عند جميع نسب البيانات المفقودة الثلاث، وتُظهر القيم أن متوسطات المؤشر المعدل كانت أعلى من متوسطات المؤشر التقليدي في جميع الحالات، ويلاحظ أن أكبر فرق بين المتوسطات سُجِّل عند نسبة الفقد ٢٠٪، حيث كان الفرق بين المؤشرين واضحاً مما يعكس تفوق المؤشر المعدل في الكشف عن الفقرات غير المطابقة.

مناقشة النتائج

أظهرت نتائج هذه الدراسة أن البيانات المفقودة وطول الاختبار وحجم العينة تؤثر وبشكل دال إحصائياً في الخطأ من النوع الأول لمؤشرات مطابقة الفقرة التقليدي ($S - \chi^2$) والمعدل ($M_{impute} - \chi^2$)، حيث بينت النتائج أن كلا المؤشرين يتأثران سلباً بزيادة طول الاختبار ونقصان حجم العينة، وهذه النتيجة تتفق مع ما توصلت إليه العديد من الدراسات السابقة مثل دراسة: (Zhang & Wang, 2022; Chon et al., 2010; Kang & Chen, 2008) على صعيد الخطأ من النوع الأول، أما بالنسبة لعامل نسبة البيانات المفقودة، فقد دلت النتائج على تأثير المؤشر التقليدي بشكل كبير، حيث ارتفع معدل الخطأ من النوع الأول بشكل حاد بارتفاع نسبة البيانات المفقودة، على الرغم من الأداء الجيد للمؤشر عند النسبة ١٪، بينما أظهر المؤشر المعدل قدرة أعلى على التعامل مع البيانات المفقودة، فكان أداؤه أكثر استقراراً وثباتاً، كما هو واضح من الشكل ٣، كما أظهرت النتائج فروقاً دالة إحصائية بين أداء المؤشر التقليدي ($S - \chi^2$) والمعدل ($M_{impute} - \chi^2$) في أغلب ظروف التصميم التجريبي، فقد تبين من خلال تحليل الخطأ من النوع الأول أن المؤشر المعدل حافظ بدرجة أكبر على معدلات خطأ منخفضة، حتى في ظل ظروف مثل زيادة طول الاختبار أو ارتفاع نسب البيانات المفقودة، وتتفق هذه النتيجة ما توصلت إليه دراسة (Zhang & Wang, 2022).

وفيما يتعلق بقوة الاختبار، فقد أظهرت النتائج تفوقاً واضحاً للمؤشر المعدل من حيث قدرته على اكتشاف الفقرات غير المطابقة، إذ سجل معدلات كشف صحيحة مرتفعة (CDRs) عبر مختلف مستويات العوامل المدروسة، وبفروق ذات دلالة عملية مقارنة بالمؤشر التقليدي، وهذا يعكس مرونة المؤشر المعدل في التعامل مع مختلف الظروف التجريبية، ويعزز من موثوقيته في تطبيقات القياس التي تتطلب دقة في التمييز بين الفقرات المطابقة وغير المطابقة للنموذج الرياضي المعتمد وتتفق هذه النتيجة مع ما توصلت إليه دراسة (Zhang & Wang, 2022)، وعند النظر في أثر العوامل الثلاثة (طول الاختبار، حجم العينة، نسبة البيانات المفقودة)، كشفت نتائج التحليل أن هذه العوامل تؤثر بدرجات متفاوتة على أداء المؤشرين، حيث تبين أن طول الاختبار وحجم العينة يشكلان عاملين حاسمين في تعزيز أو إضعاف كفاءة المؤشرين، فكلما زاد طول الاختبار وحجم العينة، تحسن أداء المؤشرين عمومًا، إلا أن المؤشر التقليدي كان أكثر تأثراً بانخفاض هذه القيم، ما يدل على أنه أقل مقاومة للظروف التجريبية الصعبة، أما بالنسبة لعامل نسبة البيانات المفقودة، فقد كشفت النتائج أن ارتفاع هذه النسبة أضعف أداء المؤشر التقليدي بشكل واضح، بينما حافظ المؤشر المعدل على درجة من الاستقرار النسبي، مما يؤكد كفاءته عند التعامل مع بيانات غير مكتملة، ويمكن تفسير هذه الفروقات من منظور البنية الإحصائية لكل من المؤشرين، إذ يعتمد المؤشر التقليدي على المجموع الكلي لدرجات الفرد، دون اعتبار لحالات الفقد، بينما يأخذ المؤشر المعدل في الاعتبار القيم المحسوبة بالتعويض، ما يمنحه قدرة أعلى على التقدير في ظل غياب جزئي للمعلومات.

وعند المقارنة بنتائج الدراسات السابقة، فإن ما توصلت إليه هذه الدراسة يتفق مع ما أشار إليه زانغ ووانغ (Zhang & Wang, 2022) حول أهمية تطوير مؤشرات تتعامل بمرونة مع البيانات غير المكتملة.

وفي ضوء ما سبق، يمكن القول إن المؤشر المعدل يوفر أداة أكثر فاعلية لتقييم مطابقة الفقرات، ويمثل خطوة متقدمة نحو تحسين جودة أدوات القياس التربوي والنفسي، لا سيما في البيئات التي تتصف بتعقيد البيانات أو ضعف في ظروف التصميم التجريبي.

التوصيات والمقترحات

١. اعتماد المؤشر المعدل في التطبيقات العملية التي تهدف إلى تقييم مطابقة الفقرات في اختبارات القياس التربوي والنفسي، خاصة في ظل فقدان نسبة كبيرة من البيانات.
٢. تطبيق المؤشرات المعدلة على نماذج متعددة الأبعاد لنظرية الاستجابة للفقرة، مما يساهم في تعميم نتائج المؤشرات وتقييم أدائها تحت نماذج أكثر تنوعاً.
٣. دراسة أثر اختلاف طبيعة الفقرات (صعوبة، تمييز، عدد البدائل) على دقة المؤشرات المختلفة.

الشكر والتقدير

نتوجه بخالص الشكر إلى المراجعين ومحرري التحرير على تعليقاتهم القيمة على المسودة الأولية لهذه الورقة البحثية. كما نعبر عن امتناننا لكل من ساهم في إنجاز هذا البحث، سواء بالملاحظات العلمية أو الدعم الفني أو المساعدة في استخراج البيانات

الإقرارات الأخلاقية

تم عمل هذا البحث وفقاً لمبادئ النزاهة البحثية والأخلاقيات العلمية. وتم التعامل مع جميع مصادر ومواد البحث بمسؤولية وبما يضمن حقوق الملكية الفكرية. وتم الالتزام بمبادئ الدقة والشفافية والمصادقية لضمان سلامة النتائج.

بيان توافر البيانات

تتوفر مجموعات البيانات التي تم توليدها وتحليلها خلال هذه الدراسة عند الطلب من الباحثين.

التمويل

لم يتم تلقي أي تمويل لإجراء هذا البحث.

بيان تضارب المصالح

يقر المؤلفان أنه ليس لديهم أي تضارب في المصالح.

مساهمات المؤلفين

قام الباحث الأول بتصميم الدراسة، وتحليل البيانات، وكتابة المسودة الأولى. بينما قام الباحث الثاني بتدقيقها وإبداء الملاحظات حيالها.

الموافقة على النشر

المؤلفان موافقان على نشر هذا العمل بصورته النهائية في المجلة.

استخدام الذكاء الاصطناعي

تم استخدام أدوات الذكاء الاصطناعي لتحسين بعض الصياغات فقط.

References

- Bartholomew, D. J., & Leung, S. O. (2002). A goodness-of-fit test for sparse 2^p contingency tables. *British Journal of Mathematical and Statistical Psychology*, 55(1), 1–15. <https://doi.org/10.1348/000711002159617>
- Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2^p tables. *British Journal of Mathematical and Statistical Psychology*, 59(1), 173–194. <https://doi.org/10.1348/000711005X66419>
- Chon, K. H., Lee, W. C., & Dunbar, S. B. (2010). A comparison of item fit statistics for mixed IRT models. *Journal of Educational Measurement*, 47(3), 318–338. <https://doi.org/10.1111/j.1745-3984.2010.00116.x>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Douglas, J., & Cohen, A. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement*, 25(3), 234–243. <https://doi.org/10.1177/01466210122032046>
- Glas, C. A., & Suárez-Falcón, J. C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, 27(2), 87–106. <https://doi.org/10.1177/0146621602250530>
- Haberman, S. J. (2009). Use of generalized residuals to examine goodness of fit of item response models. *ETS Research Report Series*, 2009(1), i–17. <https://doi.org/10.1002/j.2333-8504.2009.tb02172.x>

Articles in Press

- Haberman, S. J., Sinharay, S., & Chon, K. H. (2013). Assessing item fit for unidimensional item response theory models using residuals from estimated item response functions. *Psychometrika*, 78(3), 417–440. <https://doi.org/10.1007/s11336-012-9305-1>
- Kang, T., & Chen, T. T. (2008). Performance of the generalized S-X² item fit index for polytomous IRT models. *Journal of Educational Measurement*, 45(4), 391–406. <https://doi.org/10.1111/j.1745-3984.2008.00071.x>
- Li, Y., & Rupp, A. A. (2011). Performance of the S- χ^2 statistic for full-information bifactor models. *Educational and Psychological Measurement*, 71(6), 986–1005. <https://doi.org/10.1177/0013164410392031>
- Liang, T., & Wells, C. S. (2009). A model fit statistic for generalized partial credit model. *Educational and Psychological Measurement*, 69(6), 913–928. <https://doi.org/10.1177/0013164409332222>
- Liu, Y., & Maydeu-Olivares, A. (2013). Local dependence diagnostics in IRT modeling of binary data. *Educational and Psychological Measurement*, 73(2), 254–274. <https://doi.org/10.1177/0013164412453841>
- Liu, Y., & Maydeu-Olivares, A. (2014). Identifying the source of misfit in item response theory models. *Multivariate Behavioral Research*, 49(4), 354–371. <https://doi.org/10.1080/00273171.2014.910744>
- Maydeu-Olivares, A., & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2ⁿ contingency tables: A unified framework. *Journal of the American Statistical Association*, 100(471), 1009–1020. <https://doi.org/10.1198/016214504000002069>
- McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement*, 9(1), 49–57. <https://doi.org/10.1177/014662168500900105>
- Mislevy, R. J., & Wu, P. K. (1996). Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing. *ETS Research Report Series*, 1996(2), i–36. <https://doi.org/10.1002/j.2333-8504.1996.tb01708.x>
- Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24(1), 50–64. <https://doi.org/10.1177/01466216000241003>
- Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X²: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, 27(4), 289–298. <https://doi.org/10.1177/0146621603027004004>
- Reiser, M. (2008). Goodness-of-fit testing using components based on marginal frequencies of multinomial data. *British Journal of Mathematical and Statistical Psychology*, 61(2), 331–360. <https://doi.org/10.1348/000711007X204215>
- Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement*, 42(4), 375–394. <https://doi.org/10.1111/j.1745-3984.2005.00021.x>

Articles in Press

- Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology*, 59(2), 429–449. <https://doi.org/10.1348/000711005X66888>
- Stone, C. A. (2000). Monte Carlo-based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement*, 37(1), 58–75. <https://doi.org/10.1111/j.1745-3984.2000.tb01076.x>
- Wang, C., Shu, Z., Shang, Z., & Xu, G. (2015). Assessing item-level fit for the DINA. *Applied Psychological Measurement*, 39(7), 525–538. <https://doi.org/10.1177/0146621615583050>
- Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5(2), 245–262. <https://doi.org/10.1177/014662168100500212>
- Zhang, B., & Stone, C. A. (2008). Evaluating item fit for multidimensional item response models. *Educational and Psychological Measurement*, 68(2), 181–196. <https://doi.org/10.1177/0013164407301547>
- Zhang, X., & Wang, C. (2022). Modified item-fit indices for dichotomous IRT models with missing data. *Applied Psychological Measurement*, 46(8), 705–719. <https://doi.org/10.1177/01466216221099999>
- Zhang, X., Wang, C., & Tao, J. (2018). Assessing item-level fit for higher order item response theory models. *Applied Psychological Measurement*, 42(8), 644–659. <https://doi.org/10.1177/0146621618762740>