OPEN ACCESS
تاريخ الاستلام: 16/ 05/ 2025م
تاريخ القبـول: 25/ 05/ 2025م

مجلة الآداب
للدراسات النفسية والتربوية

أثر حجم العينة وطول الاختبار على دقة تقدير معالم الفقرات وفقًا لنموذج التقدير الجزئي المعمم ونموذج دلتا لتقدير الدرجات

دلال عبد الرحمن محمد العويدي*
Dalal.88@live.com

أ.د إسماعيل بن سلامة البرصان**
ibursan@ksu.edu.sa

الملخص

تهدف الدراسة إلى فحص أثر كل من حجم العينة وطول الاختبار على دقة تقدير معالم الفقرات ضمن نموذج التقدير الجزئي المعمم(GPCM) ، ونموذج دلتا لتقدير الدرجات (DSM)، تم استخدام المنهج التجريبي على المحاكاة، إذ تم توليد بيانات افتراضية لأحجام عينات متفاوتة (500، 1000، 5000)، واختبارات بأطوال (5، 10، 15) فقرات، وتم تحليل البيانات باستخدام برنامجي (R وDelta)، كما تم تقييم دقة تقدير المعالم وفق ثلاثة مؤشرات: التحيز، متوسط مربع الخطأ، ومعامل الارتباط. كما أجري تحليل التباين الثلاثي (Three Way ANOVA) باستخدام (SPSS) لدلالة الفروق في دقة التقدير لكل ظرف من ظروف الدراسة والتفاعل بينهما؛ وقد أظهرت النتائج تفوّق نموذج (GPCM) على (DSM) في دقة تقدير صعوبة فئات الاستجابة ومعلم التمييز. كما بيّنت النتائج أن نوع النموذج كان العامل الوحيد ذا الأثر الدال إحصائيًا على دقة تقدير صعوبة فئات الاستجابة، في حين لم يكن لحجم العينة وطول الاختبار أو التفاعل بينهما أثر دال. أما بالنسبة لمعلم التمييز، فكان للنموذج وطول الاختبار أثر دال، بينما لم يظهر حجم العينة تأثيرًا دالًا. كما تبين وجود دلالة للتفاعل بين حجم العينة وطول الاختبار فقط.

الكلمات المفتاحية: أثر حجم العينة، طول الاختبار، دقة تقدير معلم الفقرات، نموذج التقدير الجزئي المعمم، نموذج دليا لتقدير الدرجات

---

* طالبة دكتوراه في مسار القياس والتقويم بقسم علم النفس- كلية التربية- جامعة الملك سعود. المملكة العربية السعودية

** أستاذ القياس والتقويم - بقسم علم النفس- كلية التربية- جامعة الملك سعود -المملكة العربية السعودية

للاقتباس: العويدي، دلال عبد الرحمن محمد؛ البرصان، إسماعيل بن سلامة. (2025). أثر حجم العينة وطول الاختبار على دقة تقدير معالم الفقرات وفقًا لنموذج التقدير الجزئي المعمم ونموذج دلتا لتقدير الدرجات ، *مجلة الآداب للدراسات النفسية والتربوية*، 7(2)، 89-125.

OPEN ACCESS
Received: 16-05-2025
Accepted: 25-05-2025

مجلة الآداب
للدراسات النفسية والتربوية

Effect of Sample Size and Test Length on Item Parameter Estimation Accuracy according to Generalized Partial Credit Model and the Delta Scoring Model

Dalal Abdulrahman Mohammed Al-Owaidi [*]
Daalowidy@pnu.edu.sa

Prof. Dr. Ismail bin Salamah Al-Bursan [**]
ibursan@ksu.edu.sa

Abstract

This study aims to examine sample size and test length effects on item parameter estimation under the Generalized Partial Credit Model (GPCM) and the Delta Scoring Model (DSM). A comparative analytical approach based on simulation was employed, wherein simulated data were generated for varying sample sizes (500, 1000, 5000) and test lengths (5, 10, 15 items). The data were analyzed using R and Delta software. Estimation accuracy was evaluated using three indicators: bias, mean square error (MSE), and correlation coefficient. A three-way ANOVA was conducted using SPSS to assess statistically-significant of differences in estimation accuracy across study conditions and their interactions. Results demonstrated the superiority of the GPCM over the DSM in estimating response category difficulty and discrimination parameters accuracy. The model type was the only statistically significant factor affecting the accuracy of response category difficulty estimation, whereas neither sample size, test length, nor their interaction exhibited significant effects. For the discrimination parameter, both the model and test length showed significant effects, while sample size did not. Additionally, a statistically significant interaction was observed between sample size and test length.

Keywords: Sample size effect, test length, accuracy of item parameter estimation, Generalized Partial Credit Model (GPCM), Delta Scoring Model (DSM)

[*] PhD Scholar in Measurement and Evaluation, Psychology Department, College of Education, King Saud University, Saudi Arabia
[**] Professor of Measurement and Evaluation, Psychology Department, College of Education, King Saud University, Saudi Arabi.

Cite this article as:: Al-Owaidi, Dalal Abdulrahman Mohammed. & Al-Bursan, Ismail bin Salamah (2025). Effect of Sample Size and Test Length on Item Parameter Estimation Accuracy according to Generalized Partial Credit Model and the Delta Scoring Model. *Journal of Arts for Psychological & Educational Studies* 7(2) 89-125

مقدمة الدراسة:

اهتم علم النّفس بدراسة السّلوك الإنساني بهدف فهمه وتفسيره وضبطه والتنبؤ به في المستقبل. فالمتتبع لتاريخ علم النّفس الحديث يجد جهود ويليام فونت عام (1800) الذي أخرج علم النّفس من قالب التأمل الفلسفي والملاحظات غير الممنهجة إلى علم يعتمد على القياس ضمن ظروف مضبوطة بدقة. تلتها جهود فرنسيس جالتون الذي بين أنّ القدرات العقلية يمكن أن تتوزع اعتداليًا، وأنّ من الممكن وصف أيّ سلوك إنساني من خلال متوسطه الحسابي وانحرافه المعياري. كما أنتج جالتون مفهوم الارتباطات الذي بنى على أثره تلميذه بيرسون معامل الارتباط (*r*) ثم اختبار (*chi-square*).

فالقياس الدّقيق لقدرات الأفراد وخصاص الفقرات هو حجر الأساس لكل محاولات علماء القياس لتقديم أدق تقدير للعلاقة بين السمة الكامنة (القدرة) للفرد والاستجابة الصحيحة على الفقرة. وقد مرّ تحقيق هذا الهدف بعدد من المحاولات التي أنتجت نماذج نظرية الاستجابة للفقرة(Item Response Theory (IRT؛ التي حلّت بعض تحديات نظرية القياس الكلاسيكية Classic Test Theory (CTT). إذ تفترض نظرية القياس الكلاسيكية (CTT) وجود درجة ملاحظة للفرد (X) يمكن الحصول عليها بجمع مكوّنين افتراضيين، وهما الدرجة الحقيقية (T) ودرجة الخطأ العشوائي (E). وعلى الرّغم من سهولة إجراءات نظرية القياس الكلاسيكية (CTT) ووضوح نتائجها إلّا أنّها أُنتقدت في بعض النّواحي؛ ومنها اعتماد خصائص فقرات الاختبار على خصائص عيّنة الأفراد التي أُجريَ الاختبار عليها، كما أنّ أداء الأفراد على الاختبار يختلف باختلاف فقرات الاختبار، بالإضافة إلى أنّها بحكم الواقع تفترض تساوي تباين أخطاء القياس لدى جميع الأفراد في عينة الاختبار، وغيرها من الانتقادات التي انعكست على دقة النتائج التي يتم التوصل لها عند تطبيق نظرية القياس الكلاسيكية (CTT)، الأمر الذي أسهم في تطوير أساليب القياس وظهور نظرية الاستجابة للفقرة (IRT) بفضل جهود لورد (Lord, 1952) إذ اهتمت بشكل أساس في تحليل بيانات الاختبارات التّربوية والنّفسية بمنهجية مغايرة ومتقدمة عن نظرية القياس الكلاسيكية (CTT)؛ فحَررت نظرية الاستجابة للفقرة (IRT) الاختبارات النّفسية والتّربوية من هذه الافتراضات؛ وأصبحت الخصائص السيكومترية للفقرات مستقلة عن خصائص عيّنة الأفراد، وخصائص عيّنة الفرد متحررة من الخصائص السيكومترية للاختبار، وهو ما يعرف باللاتغاير في تقديرات معالم الفقرات المكوّنة للاختبار بتغير شكل توزيع قدرة الأفراد (person free)، وكذلك اللاتغاير في تقديرات معلمة القدرة عند الفرد بتغير

معالم الفقرات التي أجابوا عليها (Item free)، علاوة على تقدير الخطأ المعياري في التّقدير حول موقع كل شخص من خلال تحديد فترات الثقة، الأمر الذي أسهم في توظيف نظرية الاستجابة للفقرة (IRT) بشكل كبير في معادلة درجات الاختبارات وإنشاء بنوك الأسئلة، وإجراء الاختبارات التكيفية ومعالجة تحيز الفقرات. تتألّف نظرية الاستجابة للفقرة (IRT) من عدد من النماذج الأساسية وهي بالواقع دوال رياضية احتمالية تنقسم من حيث قابلية بياناتها للتدريج إلى نموذجين أساسيين:

1. النماذج ثنائية التدريج (Dichotomous Models): وهو النموذج الذي يستند على نوعين من الاستجابات، إمّا صحيح أو خاطئ ويتم ترميزها بـ (0، 1) إذ يحصل المختبر على الدرجة (1) عند الاستجابة الصحيحة، وعلى الدرجة (0) عند الاستجابة الخاطئة. ينقسم النموذج الثنائي كما أورده Hambleton & Swaminathan (1985) إلى ثلاثة نماذج أساسية هي:

- النموذج اللوجستي أحادي المعلم ((1PL) 1 parameter Logistic Model): هو أبسط نماذج نظرية الاستجابة للفقرة (IRT) وأكثرها شيوعًا وهو نموذج مكافئ لنموذج راش. الذي يتنبأ باحتمالية الاستجابة الصّحيحة باستخدام معالم الصعوبة (b) فقط لكل فقرة من فقرات الاختبار، إذ تعرف الصّعوبة بأنّها: نقطة الانعطاف في دالة الاستجابة للفقرة التي يتغير عندها اتجاه ميل المنحنى، كما يفترض النموذج اللوجستي أحادي المعلم (1PL) أن معلمة التمييز (a) ثابتة لجميع العناصر (دي أيالا، آر، 2017).
- النموذج اللوجستي ثنائي المعلم ((2PL) 2 parameter Logistic Model): وهو تعميم رياضي للنموذج (1PL) من خلال تغيير قيمة (a) عبر فقرات الاختبار، تتنبأ دالة النموذج اللوجستي ثنائي المعلم (2PL) باحتمالية الاستجابة الصّحيحة للفرد على فقرات الاختبار باستخدام معلمة الصعوبة (b) والتمييز (a) لكل فقرة من فقرات الاختبار. وتُعرف معالم الصعوبة بأنّها: نقطة انقلاب المنحنى، كما يصف معلم التمييز قدرة الفقرة على التمييز بين الأفراد على متصل السمة وبازدياد قيمة (a) يصبح ميل الدالة أكثر انحدارًا بالتي أكثر قدرة على التمييز بين الأفراد. (دي أيالا، آر، 2017).
- النموذج اللوجستي ثلاثي المعلم ((3PL) 3 parameter Logistic Model): وهو النموذج الأكثر عمومية من الصيغتين السابقتين إذ يضيف عامل الصدفة من خلال معامل التخمين (c)، تتنبأ دالة النموذج اللوجستي ثلاثي المعلم (3PL) باحتمالية الاستجابة الصّحيحة للفرد على فقرات الاختبار باستخدام معلمة الصعوبة (b) والتمييز (a) والتخمين (c) لكل فقرة من فقرات الاختبار. يتم تحديد معلم الصعوبة في نقطة على مقياس القدرة عندما يكون احتمال الإجابة الصّحيحة واقعًا في نصف المدى بين التخمين و(1)، كما يصف معلم التمييز قدرة الفقرة على التّمييز بين الأفراد على

متصل السمة، وبازدياد قيمة (a)؛ يصبح ميل الدالة أكثر انحدارًا بالتي أكثر قدرة على التمييز بين الأفراد. (دي أيالا، آر، 2017)

2. النماذج متعددة التدريج (Polychotomous Models):

a. ظهرت النماذج متعددة التدريج لمعالجة البيانات التي تحتوي على أكثر من درجتين محتملتين؛ كما في مقاييس تقدير ليكرت (التقييم على المقياس من 1 إلى 5) أو كما في المسائل الرياضية المعقدة التي يجب فيها الحصول على نتائج وسيطة متعددة لحل المهمة بأكملها بشكل صحيح، فعند تدريج هذه المهمة بطريقة ثنائية التدريج سوف يتم فقد كمية كبيرة من المعلومات، إذ إنّ الطالب الذي يحل جميع الخطوات المتوسطة ويفشل فقط في الخطوة الأخيرة سيحصل على النتيجة نفسها التي يحصل عليها الشخص الذي لم يتمكن حتى من اتخاذ الخطوة الأولى، بالتالي يُعد منح الطالب الذي تمكن من الحل الجزئي رصيدًا جزئيًا من الدرجة الكلية هو أقرب تقدير لقدرة الطالب. تُعرف النماذج التي تتعامل مع هذا النوع من الاستجابات الترتيبية بالنماذج متعددة التدريج، ويمكن تصنيفها كما أوردها دي أيالا، آر، (2017):

2. نماذج راش لبيانات ترتيبية متعددة التدريج :

هي تعميمات رياضية لنموذج راش ثنائي التدريج تفترض هذه النماذج ثبات معامل التمييز عبر الفقرات (a=1) إذ يتم تقدير معالم الفقرات والأفراد من خلال تقديرات مواقع الأشخاص وصعوبات العناصر والعتبات الخاصة بكل فقرة، ومنها:

a. نموذج التقدير الجزئي Partial Credit Model (Masters, 1982): يفترض أنّ احتمالية اختيار فئة (k) على فئة (k-1) تحكمها نموذج الاستجابة ثنائية التدريج، ولتحليل البيانات التي تتضمن عددًا من فئات الاستجابة المتعددة، التي تتضمن عددًا من الخطوات للوصول إلى الحل الصحيح؛ إذ يمنح الفرد جزءًا من الدرجة الكلية نتيجة حصوله على جزء من الحل، كما يختلف مقدار الصّعوبة النّسبية عبر فئات الاستجابة، ويمكن تقدير العتبات لكل فقرة من فقرات الاختبار بشكل مستقل؛ وهي لا تكون - بالضرورة - متساوية أو مرتبة بشكل ثابت.

b. نموذج مقياس التقدير Rating Scale Model (Andrich, 1978): يُعد RSM حالة خاصة من النموذج الجزئي PCM؛ إلّا أنّه يفترض أن العتبات ثابتة عبر الفئات. أي أنّ الصعوبة النسبية بين العتبات هي نفسها عبر الفقرات، باختلاف معلمة الصعوبة لكل فقرة من فقرات الاختبار.

3. *نماذج ليست من نماذج راش لبيانات ترتيبية متعددة التدريج:*

هي تعاميم رياضية للنموذج اللوجستي ثنائي المعلمة (2PL) أي أنّها تتعامل مع معلمة الصعوبة (b) والتمييز (a)، ويتم وصف كل فقرة بمعامل تمييز واحد (Item Slope) وعدد من العتبات (Thresholds) يساوي عدد الفئات ناقص واحد (mj-1).

a. نموذج الاستجابة المتدرجة Graded Response Model (Samejima, 1969): يستخدم GRM لتحليل البيانات الترتيبية كما يحدد احتمال استجابة الشخص بدرجة فئة معينة أو أعلى مقابل الاستجابة بدرجة فئة أدنى.

b. نموذج الاستجابة الاسمية Nominal Response Model (Bock, 1972): يستخدم NRM لتحليل البيانات الفئوية التي لا يوجد علاقة طردية أو عكسية بين الاستجابات المشاهدة وحجم القدرة.

c. نموذج التقدير الجزئي المعمم Generalized Partial Credit Model (Muraki, 1992): يُعد GPCM امتدادًا لنموذج التقدير الجزئي (PCM) الذي طوره Masters عام 1982 إلّا أنّ نموذج GPCM يتضمن معلمة التمييز لتقدير جودة قياس الفقرة. .

والجدير بالذكر أنّ نظرية الاستجابة للفقرة (IRT) تقوم على عدد من الافتراضات التي ذكرها Hambleton & Swaminathan (1985) من أجل تحقيق أدق نتائج للقياس ومنها:

- أحادية البعد (Unidimensional): أي أنّ الاستجابات على المقياس تعزى إلى سمة فقرة (جميع الفقرات تقيس سمة واحدة).
- الاستقلال الموضعي (Local Independence): ويعني الاستقلال الإحصائي لاستجابات الفرد للفقرات المختلفة في المقياس.
- منحنى خصائص الفقرة (Item Characteristic Curve): وجود دالة مميزة لكل فقرة من فقرات المقياس تربط بين احتمال الإجابة الصحيحة عن فقرات المقياس والقدرة التي تقيسها فقرات المقياس.
- التحرر من السرعة (Speedness): يعني أنّ عامل السرعة لا يلعب دورًا في الإجابة عن الفقرة.

نموذج دلتا لتقدير الدرجات D-scoring Method of Measurement (DSM):

استكمالًا للجهود الحثيثة لعلماء القياس في تحسين وتجويد دقة القياس قدم ديميتر ديميتروف نموذج دلتا لتقدير الدرجات D-scoring Method of Measurement (DSM)؛ وهو نموذج

ثنائي حديث نُشرت أوّل ورقة علمية له عام 2016 (Dimitrov, 2016)؛ وهو امتدادٌ للنظرية الكلاسيكية في القياس، إلّا إنّه يتضمن بعض مزايا النّظرية الحديثة مثل: (أ) اعتماد درجة الاختبار على نمط استجابة الممتحن، وتأخذ في الاعتبار الصعوبة المتوقعة لعناصر الاختبار، (ب) إمكانية تدريج درجات المفحوصين وصعوبة الفقرات على مقياس واحد يعبر عن السمة المقاسة، (ج) تتوفر معادلات تحليلية لدالة الاستجابة للفقرة item response function (IRF) لتقدير الدرجات الحقيقية، والخطأ المعياري المشروط في القياس(Conditional Standard Error of measurement(CSEM، وغيرها من القياسات النفسية المفيدة (Dimitrov et al., 2020). يتراوح تدريج دلتا في الإطار الكلاسيكي (Classical Framework DSM-C) بين (0) إلى (1) ويستخدم لتعيين قدرات الأفراد (D) والصعوبات المتوقعة لمفردات الاختبار (δi) المقدرة وفق أسلوب دلتا لتقدير الدرجات (DSM) وتحديد مواقعهم على متصل واحد للسمة. كما يمكن تمثيل درجات الأفراد (D) والصعوبات المتوقعة للفقرات (δi) باستخدام التوزيعات التكرارية لكل منهما وهو ما يعرف بالخريطة للفقرات والفرد Item Person Map (IPM) التي تساعد على المقارنة بين درجات الأفراد بخصائص الفقرات. (Dimitrov, 2020).

لاحقًا، تمت معالجة نموذج دلتا لتقدير الدرجات لكي يعالج بيانات متعددة التدريج تحت ما يسمى بالإطار الكامن Latent Framework DSM-L لأسلوب دلتا لتقدير الدرجات (Dimitrov & Luo, 2019). تُقدر درجة (D-score) للفرد على نمط استجابته التي يتم ترجيحها حسب الصعوبات المتوقعة للفقرات بالنسبة لمجموعة المتقدمين للاختبار(Dimitrov, 2016). طُور ديميتر ديميتروف نموذج دلتا لتقدير درجة المفحوص التي تعبر عن قدرته في السمة المقاسة بالاختبار التي يرمز لها بالرمز (D) وهي تقابل القدرة (θ) في نظرية القياس الحديثة (IRT). تُقدر درجة (D-score) للفرد على متجه (نمط) استجابته، الذي يتم ترجيحها حسب الصعوبات المتوقعة للفقرات بالنسبة لمجموعة المتقدمين للاختبار (Dimitrov, 2016). ويمكن تقدير الصعوبة المتوقعة لفقرات الاختبار كدالة تحليلية لمعلمات IRT الخاصة بها (a,b,c). يتشارك النموذجان في عدد من المفاهيم والتفسيرات السيكومترية الرئيسية؛ إذ نجد أنّ كلا الإطارين يتشاركان التدريج D-scale نفسه، يتراوح تدريج D في كلا الإطارين بين ($0 \leq D \leq 1$) إذ تشير ($D = 0$) إلى أنّه لم يتم الإجابة عن أي من عناصر الاختبار بشكل صحيح و($D = 1$) عندما تتم الإجابة عن جميع العناصر بشكل صحيح. تشير درجة Dw للفرد إلى نمط استجابته الذي يتم ترجيحه حسب الصعوبات المتوقعة للعناصر بالنسبة لمجموعة المتقدمين للاختبار وهي مقدار القدرة المطلوبة للنجاح الكامل في الاختبار (Dimitrov, 2016). إلّا أنّ الإطارين يختلفان في طرقهما

لتقدير معلمات الفقرات والفرد؛ إذ أظهرت دراسة (Dimitrov & Atanasov, 2021) أنّ تقديرات D-scale ضمن DSM-L أكثر دقة بشكل عام من نظيرتها DSM-C، إلّا أنّ معامل الارتباط بين قيمها مرتفع جدًا ($r \approx 0.99$).

تُعد النماذج المعلمية للنظرية الحديثة في القياس من الأدوات الأساسية في تقييم جودة الفقرات واستخلاص معلومات دقيقة حول قدرات الأفراد، وقد حظيت باهتمام متزايد في الأوساط التربوية والنفسية. كما أنّ لنوع النموذج المستخدم في تقدير معالم الفقرات والأفراد دور رئيسي في تحديد دقة التقدير؛ ومن بين هذه النماذج، يبرز كل من نموذج التقدير الجزئي المعمم (Generalized Partial Credit Model - GPCM) ونموذج دلتا لتقدير الدرجات (Delta Scoring Model - DSM) كأداتين فعّالتين في تحليل البيانات الاسمية والمتعددة المستويات. تعتمد فعالية هذه النماذج على مدى دقة تقدير معالم الفقرات، التي تتأثر بعدد من العوامل التصميمية منها: سؤال البحث، مجال الدراسة، عدد معلمات الفقرات والأفراد المراد تقديرهم، طول الاختبار، وحجم العينة؛ إلا أنه لا تزال هناك حاجة إلى مزيد من البحث لفهم كيف تتفاعل هذه العوامل مع نماذج محددة مثل GPCM وDSM، خصوصًا في ظل تفاوت مستويات استخدامها وتطورها النظري والتطبيقي. وانطلاقًا من هذا السياق، تهدف الدراسة الحالية إلى تحليل أثر كل من حجم العينة وطول الاختبار على دقة تقدير معالم الفقرات، وذلك من خلال مقارنة أداء كل من نموذج التقدير الجزئي المعمم ونموذج دلتا لتقدير الدرجات. وتكمن أهمية هذه الدراسة في تقديم أدلة كمية تساعد الباحثين والممارسين في اختيار التصميم الأمثل للاختبارات والنموذج الأكثر ملاءمة لأهداف القياس المختلفة.

ركزت الأبحاث السابقة - بشكل أساسي - على نماذج تقدير المعالم، موضحة نقاط قوتها وحدودها في سياقات مختلفة. كما أظهرت الدراسات أنّ حجم العينة يلعب دورًا حاسمًا في دقة التقديرات، وتميل أحجام العينات الأكبر إلى توفير تقديرات معلمات أكثر استقرارًا، وجديرة بالثّقة أكثر من تلك المرتبطة بالعينات الأصغر. وبالمثل، يمكن أن يؤثر طول الاختبار بشكل كبير على صحة النتائج. ومنها دراسة بني عطا (2017) التي هدفت إلى تقصي أثر طول الاختبار وحجم العينة على دقة طرق تقدير معالم الفقرات وقدرات الأفراد في برنامج بايلوج. فقد تم توليد بيانات ثنائية التدريج وفق نموذج (3PL) وباستخدام طرق التقدير (الأرجحية العظمى، توقع الاقتران، تعظيم الاقتران) المستخدمة في برنامج (Bilog- Mg3). وأظهرت النتائج وجود أثر ذي دلالة إحصائية لكل من طول الاختبار وحجم العينة والتفاعل بينهما في دقة تقديرات معالم الفقرات (التمييز، الصعوبة،

التخمين) عند استخدام طريقة الأرجحية العظمى في معايرة الفقرات، وأنّ دقة تقديرات معالم الفقرات تزداد بزيادة طول الاختبار وحجم العينة. كذلك هدفت دراسة المحاكاة لـ Djidu, Retnawati, & Haryanto (2023) لتقييم دقة تقدير معلمات الفقرات باستخدام نموذج التقدير ثلاثي المعلم (3PL)، من خلال دراسة تأثيرات حجم العينة وطول الاختبار. استخدمت الدراسة ست مجموعات تم توليدها، كما تمت إجراء عمليات المحاكاة عن طريق إعادة تحليل البيانات (15) مرة، تم استخدام برنامج (RStudio) لتحليل البيانات من خلال الحزمة الإحصائية "mirt"، وتم تقدير معلمات العناصر (الصعوبة b، التمييزa، التخمينc)، وتم تقييم دقة تقدير المعلمات باستخدام (RMSD)، كما تم استخدام تحليل التباين لتقييم تأثير حجم العينة وطول الاختبار على دقة المعلمات. كشفت النتائج أن حجم العينة وطول الاختبار يؤثران بشكل كبير على دقة تقدير معلمات العنصر، وخلصت الدراسة إلى أنّ الحد الأدنى من (25 أو 40) عنصر اختبار وحجم عينة لا يقل عن (3000) ضروريان لتقدير المعلمات بدقة باستخدام نموذج التقدير ثلاثي المعلم(3PL).

أجرى Jiang, Wang, and Weiss (2016) دراسة محاكاة للتوصل إلى العوامل التي تؤثر على تقدير معلمات الفقرات في نموذج الاستجابة التدريجية متعدد الأبعاد (MGRM) من خلال أحجام عينات وأطوال اختبار مختلفة، والارتباطات المتبادلة للمقياس. قام الباحثون بتنويع أحجام العينات، وأطوال الاختبار (9 و21 و42 فقرة)، كما تم استخدام برنامج (flexMIRT) للحصول على تقديرات المعلمات، وتم تقييم جودة هذه التقديرات من خلال مقاييس، مثل الارتباط بين المعلمات الحقيقية والمقدرة والتحيز وخطأ الجذر التربيعي المتوسط (RMSE) تضمنت المحاكاة (324) شرطًا، كشفت نتائج دراسة المحاكاة أنّه بالنسبة لمعظم السيناريوهات كان حجم العينة المكون من (500) شخص كافيًا لتحقيق تقديرات دقيقة للمعلمات. وبالنسبة للاختبارات التي تحتوي على (240) عنصرًا كان حجم العينة الأكبر من (1000) ضروريًا لضمان تقديرات موثوقة. وجدت الدراسة أيضًا أنّ زيادة حجم العينة إلى ما يزيد عن (1000) لم يعزز بشكل كبير من دقّة تقديرات معلمات (MGRM). كما تسلط دراسة داي وآخرين (2021) الضوء على نظرية الاستجابة للفقرة متعددة التدريج (IRT) ، وخاصة نموذج الاستجابة المتدرجة (GRM) ونموذج التقدير الجزئي المعمم(GPCM)، في ظل أطوال القصيرة للاختبارات، والأحجام العينات الصغيرة، والبيانات المفقودة. وتشير النتائج إلى أنّ وجود حجم عينة لا يقل عن (300) فرد وطول أداة لا يقل عن خمس فقرات يتم تقدير قدرة الأفراد ومعالم الفقرات بدقة، بالإضافة إلى ذلك يلاحظ أنّه في حين يعمل(GPCM) بشكل متسق عبر أطوال

اختبارات مختلفة، فإنّ أداء (GRM) يتحسن مع الاختبارات الأطول، ويتعامل النموذجان بشكل مختلف في وجود بيانات مفقودة إذ يقدم نموذج الاستجابة المتدرجة (GRM) تقديرات أكثر دقة بنسب أصغر في حال وجود البيانات المفقودة، بينما كان نموذج التقدير الجزئي المعمم (GPCM) أكثر موثوقية في السيناريوهات ذات البيانات المفقودة.

ونظرًا لحداثة أسلوب دلتا لتقدير الدرجات (DSM) فإنّ عدد الدراسات التي أجريت لبحثه وتقييمه محدودة جدًا، ومعظم الدراسات التي أُجريت في هذا الصدد تناولت إطار DSM-C؛ فنجد دراسة Dimitrov and Luo (2019) التي اهتمت بمقارنة النتائج التي تم الحصول عليها عن نموذج دلتا لتقدير الدرجات (DSM) مع النتائج المستمدة من نموذج الاستجابة المتدرجة (GRM)، يصف المؤلفون محاكاة البيانات بموجب (GRM) لاختبار يتكون من (15) عنصرًا متعدد الأجزاء، كل منها بأربع فئات مرتبة (0، 1، 2، 3). تم تحديد معلمات العنصر، وتم استخلاص درجات القدرة من عينة مكونة من (1000) مفحوص تتبع قدرتهم التوزيع الطبيعي؛ أشارت النتائج إلى أن المخرجات التي تم الحصول عليها من نموذج دلتا لتقدير الدرجات (DSM) كانت مرتبطة ارتباطًا وثيقًا بدرجات القدرة المستمدة من نموذج الاستجابة المتدرجة (GRM)، ممّا يشير إلى أنّ طريقة نموذج دلتا لتقدير الدرجات (DSM) تُقدر بشكل فعال القدرة الأساسية للأفراد. وجدت الدراسة أيضًا؛أنّ الاستجابة التراكمية للفئة التي تم الحصول عليها بموجب نموذج دلتا لتقدير الدرجات (DSM) أظهرت منحدرات مختلفة، ممّا يعكس قوة التمييز المتغيرة لفئات استجابة العنصر، على عكس نموذج الاستجابة المتدرجة (GRM)، إذ ظلت معلمات التمييز ثابتة عبر الفئات. وفي دراسة محاكاة قدمها Dimitrov, D. M., & Atanasov, D. V. (2021) لفحص نموذج دلتا لتقدير الدرجات بالإطار الكامن (DSM-L) هدفت لتقييم تقدير معلمات الفقرات وقدرات الأفراد من خلال نموذج دلتا لتقدير الدرجات بالإطار الكامن (DSM-L) باستخدام نموذج دالة استجابة العنصر (RFM2) كما هدفت الدراسة الى المقارنة بين المعلمات المستخرجة عن طريق نظرية استجابة للفقرة (IRT) والمعلمات المستخرجة من خلال نموذج دلتا لتقدير الدرجات بالإطار الكامن (DSM-L). تم تقدير معلمات الفقرات وقدرات الأفراد تبعا لطول وحجم العينة، وقد كشفت النتائج أن مواقع الفقرة (b) تم تقديرها بدقة في جميع الظروف؛ مع عدم وجود تحيز ذي دلالة إحصائية؛ وعلى النّقيض من ذلك كان تقدير معلمة شكل الفقرة (s) أقل دقة إلى حد ما، مع إظهار بعض الظروف تحيزًا ذا دلالة إحصائية. كما أظهر تقدير قدرة الفرد (D) قيمًا ذا دلالة إحصائية، بالإضافة إلى معامل الارتباط بين التقديرات عن طريق

نموذج دلتا لتقدير الدرجات بالإطار الكامن (DSM-L) والطرق الكلاسيكية مرتفعًا بشكل ملحوظ جدًا عبر جميع ظروف المحاكاة (r>.992).

مشكلة الدراسة:

تتحدد إشكالية الدّراسة الحالية في الوقوف على فاعلية الإطار الكامن لنموذج دلتا لتقدير الدرجات (DSM-L) ونموذج التقدير الجزئي المعمم (GPCM)، في دقة تقدير المعلمات والمقارنة بينهما؛ وبالنّظر لحداثة نموذج دلتا فهو بحاجة لهذا النوع من الدراسات الدّاعمة والمعززة لدوره في ميدان القياس النّفسي والتّربوي؛ ونظرًا لقلة الأبحاث والدّراسات التي تناولت تقدير المعالم ضمن النّماذج متعددة التّدريج مقارنة بالنماذج ثنائية التدريج؛ بالإضافة إلى عدم وجود مقارنات بين نموذج دلتا لتقدير الدرجات بالاطار الكامن (DSM-L) ونموذج التقدير الجزئي المعمم (GPCM)، وفي دقة تقدير المعلمات؛ جاءت الحاجة لإجراء مثل هذه الدراسة التي سوف تتناول متغيري "حجم العيّنة" و"طول الاختبار" ودورهما في دقّة تقدير المعالم وفقًا لنموذجين، الأول: "الإطار الكامن لنموذج دلتا لتقدير الدرجات (DSM-L)" والآخر: "نموذج التقدير الجزئي المعمم (GPCM)" وتتلخص مشكلة الدراسة في الأسئلة الآتية.

أسئلة الدراسة:

1) هل هناك أثر لحجم عيّنة مكوّن من (500، 1000، 5000) فرد، واختبار مكوّن من (5، 10، 15) فقرة على دقة تقدير معلم صعوبة فئات الاستجابة وِفقًا لنموذج التقدير الجزئي المعمم (GPCM) ونموذج دلتا لتقدير الدرجات (DSM)؟
2) هل هناك أثر للتفاعل بين حجم عيّنة مكوّن من (500، 1000، 5000) فرد، واختبار مكوّن من (5، 10، 15) فقرة على تقدير معالم صعوبة فئات الاستجابة وِفقًا لنموذج التقدير الجزئي المعمم (GPCM) ونموذج دلتا لتقدير الدرجات (DSM)؟
3) هل هناك أثر لحجم عيّنة مكوّن من (500، 1000، 5000) فرد، واختبار مكون من (5، 10، 15) فقرة على دقة تقدير معلم التّمييز وِفقًا لنموذج التقدير الجزئي المعمم (GPCM) ونموذج دلتا لتقدير الدرجات (DSM)؟
4) هل هناك أثر للتفاعل بين حجم عيّنة مكوّن من (500، 1000، 5000) فرد، واختبار مكون من (5، 10، 15) فقرة على تقدير معلم التّمييز وِفقًا لنموذج التقدير الجزئي المعمم (GPCM) ونموذج دلتا لتقدير الدرجات (DSM)؟

أهداف الدراسة: تهدف الدراسة إلى تحقيق الأهداف الآتية:

1) التّعرف على أثر حجم عيّنة مكوّن من (500، 1000، 5000) فرد وطول اختبار مكوّن من (5، 10، 15) فقرة على تقدير معلم الصعوبة وِفقًا لنموذج التقدير الجزئي المعمم (GPCM) ونموذج دلتا لتقدير الدرجات (DSM).
2) التّعرف على أثر التفاعل بين حجم عيّنة مكوّن من (500، 1000، 5000) فرد، واختبار مكوّن من (5، 10، 15) فقرة على تقدير معلم الصعوبة وِفقًا لنموذج التقدير الجزئي المعمم (GPCM) ونموذج دلتا لتقدير الدرجات (DSM).
3) التّعرف على أثر حجم عيّنة مكوّن من (500، 1000، 5000) فرد وطول اختبار مكون من (5، 10، 15) فقرة على تقدير معلم التمييز وِفقًا لنموذج التقدير الجزئي المعمم (GPCM) ونموذج دلتا لتقدير الدرجات (DSM).
4) التّعرف على أثر التفاعل بين حجم عيّنة مكوّن من (500، 1000، 5000) فرد، وطول اختبار مكون من (5، 10، 15) فقرة على تقدير معلم التمييز وِفقًا لنموذج التقدير الجزئي المعمم (GPCM) ونموذج دلتا لتقدير الدرجات (DSM).

أهمية الدراسة:

الأهمية النظرية للدراسة:

1. تسليط الضوء على فعالية نموذج دلتا لتقدير الدرجات بالإطار الكامن (DSM-L) في معالجة البيانات متعددة التّدريج.
2. الكشف عن فاعلية نموذج دلتا لتقدير الدرجات بالإطار الكامن (DSM-L) ونموذج التقدير الجزئي المعمم (GPCM) في إعطاء تقديرات دقيقة لمعالم الفقرات والأفراد باختلاف طول الاختبارات لاسيما في الاختبارات القصيرة.
3. الكشف عن فاعلية نموذج دلتا لتقدير الدرجات بالإطار الكامن (DSM-L) ونموذج التقدير الجزئي المعمم (GPCM) في إعطاء تقديرات دقيقة لمعالم الفقرات والأفراد باختلاف حجم العيّنة.

الأهمية التّطبيقية للدّراسة:

1. يتوقع أن تُفيد نتائج الدراسة في استخدام نموذج دلتا لتقدير الدرجات بالإطار الكامن (DSM-L) في معالجة البيانات متعددة التدريج.

2. سوف توفر نتائج الدراسة معلومات حول فاعلية نموذج دلتا لتقدير الدرجات ونموذج التقدير الجزئي المعمم في إعطاء تقديرات دقيقة لمعالم الفقرات والأفراد.
3. يتوقع من نتائج هذه الدراسة توفير معلومات تساعد الباحثين في توظيفها لتقدير السمات الكامنة للمفحوصين بأعلى دقة ممكنة باستخدام نموذجي دلتا لتقدير الدرجات التقدير الجزئي المعمم.

حدود الدراسة:

1. تقتصر الدراسة على استخدام أسلوب دلتا لتقدير الدرجات لديمتروف (2019) حسب الإطار الكامن (DSM-L).
2. تقتصر الدراسة على استخدام نموذج التقدير الجزئي المعمم (GPCM) لموراكي (1992).
3. تقتصر الدّراسة على بيانات متعددة التدريج بخمس فئات استجابة.
4. تقتصر الدراسة على استخدام بيانات مولدة بأسلوب المحاكاة (Simulation Study) لـ(5، 10، 15) فقرة متعددة التدريج، وفقًا لأحجام عينات مختلفة تتراوح بين (500، 1000، 5000).

مصطلحات الدراسة:

- نظرية الاستجابة للفقرة (Item Response Theory): من أبرز النماذج المعاصرة في القياس التربوي والنفسي، وقد جاءت امتدادًا وتطويرًا لنماذج القياس الكلاسيكية. تهدف هذه النظرية إلى فهم العلاقة بين استجابات الأفراد على الفقرات وسماتهم الكامنة، وذلك من خلال نماذج رياضية تصف احتمال استجابة الفرد لفقرة معينة على أساس مستوى سمة معينة لديه (دي أيالا، 2017).
- معالم الفقرات (Items Parameters): هي معلمتا الصعوبة والتمييز، إذ يُعرف معلم الصعوبة بـ (b) وهو نقطة الانعطاف في دالة الاستجابة للفقرة، التي يتغير عندها اتجاه ميل المنحنى. كما يعرف معلم التمييز بأنّه قدرة الفقرة على التمييز بين أداء الأفراد على متصل السمة وبازدياد قيمة (a) يصبح ميل الدالة أكثر انحدارًا، بالتالي أكثر قدرة على التمييز بين الأفراد (دي أيالا، 2017) ؛ وتُعرف إجرائيًا بأنها: معالم الصعوبة والتمييز للمفردات وفئات الاستجابة المقدرة باستخدام نموذج التقدير الجزئي المعمم ونموذج دلتا لتقدير الدرجات.
- نموذج التقدير الجزئي المعمم (Generalized Partial Credit Model): أحد نماذج نظرية الاستجابة للفقرة طوره (Masters) عام (1982) لمعالجة بيانات متعددة التدريج، ويُعرف إجرائيًا

بأنّه: النّموذج اللوغاريتمي الذي يُعدّ امتدادًا لنموذج التقدير الجزئي (PCM) إلّا أنّ نموذج (GPCM) يتضمن معلمة التّمييز لتقدير جودة قياس الفقرة (دي أيالا، 2017).

- نموذج دلتا لتقدير الدرجات (Delta-Scoring Model): هو نموذج ثنائي يُعدّ امتدادًا للنظرية الكلاسيكية في القياس، ويتضمن بعض مزايا النّظرية الحديثة في القياس، يتعامل مع بيانات ثنائية ومتعددة التّدريج. ويُعرف إجرائيًا بأنّه: نموذج دالة نسبية ثنائي المعلم، يقدر المعالم متعددة التدريج وفقًا للإطار الكامن لأسلوب دلتا ويرمز لها بالرمز (D) وهي تقابل القدرة (θ) في نظرية القياس الحديثة، أو نظرية الاستجابة للفقرة (Dimitrov & Atanasov, 2021)
- طرق التقدير (Estimation Methods): هي مجموعة من الأساليب الرياضية الاحتمالية بغية تقدير معالم الفقرة والأفراد في ضوء افتراضات نظرية الاستجابة للفقرة مثل (EAP، MAP، MLP) (دي أيالا، 2017) .
- الدّقة (Accuracy): تشير إلى جودة التقدير لمعالم الفقرة والأفراد، وتتميز باحتمال كبير في أن يكون التقدير قريبًا من القيمة الحقيقية للمعلم، ويُعرف إجرائيًا بأنّه التّقدير غير المتحيّز، صاحب أقل تباين بين التقديرات الأخرى غير المتحيّزة، ويُحدد باستخدام الخطأ المعياري في التّقدير (دي أيالا، 2017).

فرضيات الدراسة:

1. لا توجد فروق ذات دلالة إحصائية عند مستوى الدلالة (.005) بين متوسطات تقديرات معالم صعوبة كل من الفقرات وفئات الاستجابة تبعاً لمتغيري حجم العينة (500، 1000، 5000) وعدد فقرات الاختبار (5، 10، 15) ونموذجي التقدير الجزئي المعمم ودلتا لتقدير الدرجات والتفاعل بينهما.
2. لا توجد فروق ذات دلالة إحصائية عند مستوى الدلالة (.005) بين متوسطات تقديرات معالم تمييز كل من الفقرات وفئات الاستجابة تبعاً لمتغيري حجم العينة (500، 1000، 5000) وعدد فقرات الاختبار (5، 10، 15) ونموذجي التقدير الجزئي المعمم ودلتا لتقدير الدرجات والتفاعل بينهما.

منهج الدراسة:

تتبنى الدراسة الحالية المنهج التجريبي وهي منهجية تستخدم في تحديد العلاقة السببية بين متغيرين أو أكثر، ويتميز هذا المنهج بالتحكم الدقيق في الظروف والعوامل المختلفة وتطبيق المعالجة التجريبية على إحدى المجموعات التجريبية، ومقارنتها بمجموعات أخرى.

بيانات الدراسة: الدراسة الحالية دراسة محاكاة، ويقصد بالمحاكاة توليد بيانات حاسوبية من خلال إنتاج عينات عشوائية تسمح للباحثين باستكشاف خصائص الأساليب والأنظمة الإحصائية المراد دراستها من خلال محاكاة سيناريوهات مختلفة؛ تساعد دراسات المحاكاة في تقييم أداء النماذج الإحصائية في ظروف مختلفة، ممّا يوفر دراية أوسع بكفاءتها. وفي الدراسة الحالية تم توليد ثلاث عينات عشوائية يقابلها ثلاثة أطوال للاختبار من خلال برنامج (WinGen v1.4) (Mills, 2002).

التصميم:

تتألف الدراسة الحالية من ثلاثة متغيرات مستقلة؛ وهي: حجم العينة، وطول الاختبار، ونوع النموذج، وكل نوع من هذه المتغيرات يتضمن عددًا من المستويات على النحو الآتي:

- حجم العينة: ثلاث أحجام تتمثل في (500، 1000، 5000) فرد.
- عدد الفقرات: ثلاثة أطوال وهي الآتي: (5، 10، 15).
- النماذج: تقدير معالم الفقرات والأفراد وِفقًا لنموذجين هما: (DSM، GPCM).

بالتالي أصبح تصميم الدراسة 2x3x3 وسوف يتم تناولها حسب الترتيب الآتي:

الجدول (1) تصميم 2x3x3 يتضمن نوع النموذج وعدد الأفراد وطول فقرات الاختبار.

النموذج		GPCM			DSM		
عدد فقرات الاختبار وعدد الأفراد في العينة		5 فقرات	10 فقرات	15 فقرة	5 فقرات	10 فقرات	15 فقرة
		(500)	(500)	(500)	(500)	(500)	(500)
		مفحوص	مفحوص	مفحوص	مفحوص	مفحوص	مفحوص
		5 فقرات	10 فقرات	15 فقرة	5 فقرات	10فقرات	15 فقرة
		(1000)	(1000)	(1000)	(1000)	(1000)	(1000)
		مفحوص	مفحوص	مفحوص	مفحوص	مفحوص	مفحوص
		5 فقرات	10فقرات	15 فقرة	5 فقرات	10فقرات	15 فقرة
		(5000)	(5000)	(5000)	(5000)	(5000)	(5000)
		مفحوص	مفحوص	مفحوص	مفحوص	مفحوص	مفحوص

تم توليد البيانات حسب معطيات الدراسة باستخدام برنامج (WinGen v1.4). وفقا للخطوات الآتية:

الخطوة الأولى: توليد بيانات حجم العينة، بما أن الدراسة الحالية تتبع النماذج المعلمية لنظرية الاستجابة للفقرة (Parametric Item Response Theory Models) وهي نماذج إحصائية تتبع إجراءات وافتراضات محددة حول معلمات توزيع العينة؛ والافتراض الأكثر شيوعًا في النماذج المعلمية هو أن البيانات تتبع التوزيع الطبيعي (Normal Distribution)؛ لذا تم إنشاء استجابات لثلاثة أحجام من العينات، مكونة من (500، 1000، 5000) فرد؛ بافتراض أنّ قدراتهم (θ) تتبع التوزيع الطبيعي بمتوسط يساوي صفر *(M = 0)* وانحراف معياري يساوي واحد *(SD = 1)*.

الخطوة الثانية: توليد ثلاثة اختبارات مكونة من (5، 10، 15) فقرة متعددة التدريج بخمس فئات استجابة (mj=5) وِفقًا للنموذج التقدير الجزئي المعمم (GPCM)، لكل فئة استجابة معلمات صعوبة مختلفة، تساوي عدد فئات الاستجابة ناقص واحد (mj-1). وتُعرف الصعوبة بأنّها: نقطة الانعطاف في دالة الاستجابة للفقرة التي يتغير عندها اتجاه ميل المنحنى ويكون على شكل حرف S وتصبح عندها الدالة أكثر انحدارًا، كما يجب أن يتمتع الفرد بقدرة مرتفعة حتى يجيب بشكل صحيح على الفقرات الصعبة. كما أنّ الفقرة التي معامل صعوبتها (b) أكبر من (1) تشير إلى أنّها فقرة صعبة وبالمثل، فإنّ الفقرة التي تحتوي على معامل صعوبة (b) منخفضة (أقل من -1) تشير إلى أنها فقرة سهلة، وسيكون لدى معظم الأفراد ذوي مستوى القدرة المنخفض فرصة للإجابة عنها بشكل صحيح. تتضمن كل فقرة في هذه الاختبارات خمس فئات استجابة (1، 2، 3، 4، 5)، بالتالي أربع معلمات صعوبة (b1، b2، b3، b4) تمثل الحدود بين هذه الفئات. كما تشير نتائج دراسة (Auné, Abal, & Attorresi, 2020) إلى أن معلمات الصعوبة (b) تتراوح بين (-1.43 إلى 3.15)، مما يشير إلى أنّ مستويات الصّعوبة تمتد على نطاق واسع نسبيًا من السمة الكامنة. لذلك سيتم توليد بيانات بمعامل صعوبة يتبع التوزيع الطبيعي بمتوسط يساوي صفر (M = 0) وانحراف معياري يساوي واحد (SD = 1).

أمّا فيما يتعلق بمعامل التمييز(a) الذي يقيس مدى قدرة الفقرة على التمييز بين الأفراد الذين لديهم مستويات مختلفة من السمة الكامنة موضع القياس، ويوفر ملاءمة أفضل للبيانات من النماذج التي تفرض تمييزًا موحدًا عبر جميع الفقرات، كما يفترض أن لكل فقرة معلمة تمييز (a) تختلف من فقرة إلى أخرى، ممّا يُمكّن نموذج الاختبار من مراعاة الاختلافات في القدرات لدى

الأفراد. استخدم Hambleton & Swaminathan (1985) قيم تميز تتراوح بين (0، 2) لوجيت. ويشير معامل التمييز المرتفع إلى أنّ الفقرة تُميّز جيدًا بين الأفراد المختبرين، والعكس عندما يقترب معامل التمييز من (0). ويتضمن الجدول الآتي الإحصاء الوصفي لعينة البيانات التي تم توليدها لأغراض الدراسة.

الجدول (2) الإحصاء الوصفي لعينة الدراسة:

	العينة	المتوسط	الانحراف المعياري
500	5 فقرات	9.294	4.9494
	10 فقرات	19.06	9.115183
	15 فقرة	30.148	15.08221
1000	5 فقرات	9.306	5.166537
	10 فقرات	19.36	10.17468
	15 فقرة	32.47	15.53133
5000	5 فقرات	10.9138	4.977102
	10 فقرات	20.3452	11.51176
	15 فقرة	33.1564	16.2069

الخطوة الثالثة: التحقق من مناسبة البيانات المولدة في كل ظرف من ظروف الدراسة لافتراضات نموذج التقدير الجزئي المعمم (GPCM) ونموذج دلتا لتقدير الدرجات بالإطار الكامن (DSM-L).

تم إجراء التحليل العاملي الاستكشافي(EFA) (Exploratory Factor Analysis) للبيانات المولدة باستخدام برنامج (SPSS) وذلك من أجل التحقق من افتراض أحادية البعد الذي يُعنى بأن جميع الفقرات في الاختبار مصممة لتقيس السمة الأساسية نفسها. ويتم الاستدلال على هذا الافتراض من خلال التحقق من عدد من الاختبارات الإحصائية:

1) محك كيرز: قاعدة الجذر الكامن 1< (Kaiser rule: Eigenvalue >1)، وفيه يتم جمع مربعات تشبعات المتغيرات على العامل الأوّل، والعامل الثاني إنْ وجد؛ لإيجاد الجذر الكامن لكل منهما. وهو إذًا كما أورده دودين (2009) مقدار التباين الكلي الذي يُفسره العامل، وقد تم تعيين المقدار (1) لاعتبار العامل الذي قيمة جذره الكامن واحد صحيح وما فوق فقط، وأمّا

إذا كانت قيمة الجذر الكامن لعامل ما أقل من (1) فإنّ هذا يعني أنّ هذا العامل لا يختلف فعليًا عن متغير مستقل وحيد من متغيرات الدراسة، وبالتالي لا يمكن اعتباره عاملا (دودين، 2009)

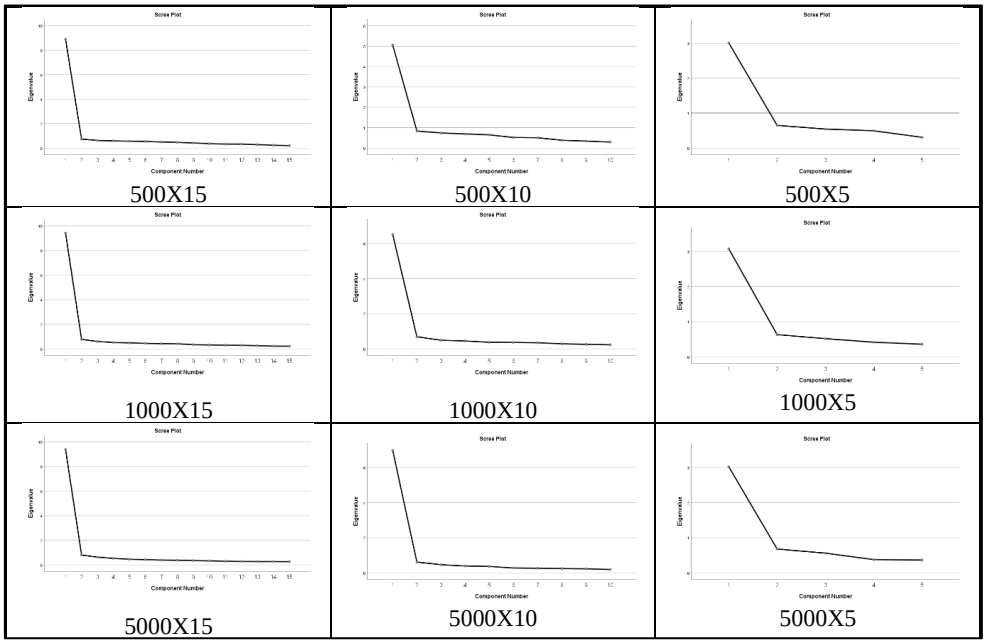
وعند فحص نتائج البيانات المولدة للدراسة الحالية يتبين أن هناك عاملًا وحيدًا في كل ظرف من ظروف الدراسة قيمة جذره الكامن أعلى من الواحد الصحيح؛ وهذا يعني أن بيانات الدراسة أحادية البعُد. وبمراجعة بيانات الجدول (3) أدناه نجد قيم الجذور الكامنة لعينات الدراسة الأربع في كل طول من أطوال الاختبارات الثلاثة، نلاحظ أن قيمة الجذر الكامن في الظرف الأول مقداره (3.408)، وكذلك نلاحظ أن هذا العامل وحده استطاع أن يُفسر (68.167%) من التباين الكلي للاختبار وهي نسبة تشير إلى الاستفادة من التحليل العاملي في تفسير معظم التباين في الظاهرة المدروسة بعدد أقل من المتغيرات. وهكذا يتم فحص جميع الجذور الكامنة ومقدار التباين المفسر لها (دودين، 2009).

الجدول (3) قيم الجذور الكامنة والتباين المفسر لها:

	العينة	قيمة الجذر الكامن	مقدار التباين المفسر
5	500	3.010	60.20
	1000	3.076	61.51
	5000	3.020	60.41
10	500	5.053	50.35
	1000	6.511	65.10
	5000	6.974	69.74
15	500	8.898	59.32
	1000	9.417	62.77
	5000	9.380	62.53

2) محك اختبار المنحدر لكاتيل (Kattell's Scree test): وهي طريقة تقوم أيضا على الجذور الكامنة، ولكن تعتمد على الرسم البياني لقيم الشكل قبل أن يصبح مستويا نوعا ما؛ ويمثل المحور السيني الأفقي العوامل، فيما يمثل المحور الصادي قيم الجذور الكامنة، ويتم رسم منحنى ينحدر من أعلى قيمة للجذر الكامن عند العامل الأوّل ثم يأخذ في التناقص إلى أن يصل نقطة ما تقابل عاملًا معينا تتباطأ عنده درجة انحدار المنحنى. ولقد اقترح هذه الطريقة كاتيل (1966) وأسماها بمنحنى المنحدر (Scree test)(دودين، 2009)

الشكل (1) *رسم سكري (Scree) لجميع ظروف العينة:*



وبذلك، تم التحقق من فرضية أحادية البعد (Unidimensional) ونجد أن جميع المحكات تشير إلى أحادية البعُد أي أنّ جميع الفقرات في بيانات الدراسة تقيس سمة واحدة.

أمّا الافتراض الثاني فيتمثل بالاستقلال الموضعي (Local Independence) وهو كما عرفه Lord and Novick (1968) بأنّ أي مستوى من مستويات القدرة للأفراد تكون التوزيعات الشرطية لدرجات الاختبار جميعها مستقلة عن بعضها بعضًا، في حين أشار Hambleton & Swaminathan (1985) إلى أنّ الاستقلال الموضعي يهتم بأن تكون استجابات الأفراد على فقرات الاختبار من القدرة نفسها، وأنّ تكون مستقلةً إحصائيًا، وبمعنى آخر، يجب أن تكون استجابة الفرد لفقرة ما لا تتأثر سلبًا أو إيجابًا باستجابته لأي فقرة أخرى، وللكشف عن الاستقلال الموضعي لفقرات الدراسة تم استخدام برنامج (IRTPRO 6.0) من خلال المؤشر الإحصائي (Standardized LD $\chi 2$) وهو عبارة عن معامل الارتباط بين البواقي (Residuals) لزوج من الفقرات بعد ضبط قدرة الفرد (θ).

تقوم برمجيات برنامج (IRTPRO 6.0) بحساب معامل الارتباط بين البواقي لأزواج من الفقرات بعد ضبط القدرة وتحويلها إلى درجات معيارية (زائية) وقد تشكلت (10) معاملات ارتباط بين بواقي أزواج فقرات الاختبار ذي (5) فقرات و (45) معاملات ارتباط بين بواقي أزواج فقرات الاختبار ذي (10) فقرات و (105) معامل ارتباط بين بواقي أزواج فقرات الاختيار ذي (15) فقرة لكل عينة من عينات

الدراسة الثلاث. وقد تراوحت معاملات الارتباط بين (0.1) في معظم فقرات الاختبار حتى (1.2) وفي خمس حالات بلغ معامل الارتباط (3.5) تقريباً. والجدير بالذكر أنّه لا يوجد معيار محدد يمكن من خلاله تحديد درجة الاعتمادية لكل المواقف، وأن أحد المشاكل الرئيسة لهذا المؤشر هو عدم وجود معيار محدد يمكن الاستناد إليه في تحديد درجة انتهاك الاستقلال الموضعي، إذ يبقى تحديد درجة الاعتمادية اعتباطيًا؛ كما اقترح Shen (1997) أنّ قيم إحصائي (LD X^2) تصبح كبيرة (أكبر من درجة 10 زائية) إذا تم انتهاك افتراض الاستقلال الموضعي، دون تلك القيمة لا يوجد انتهاك لهذا الافتراض. وعند تفحص قيم إحصائي (LD X^2) التي تم الحصول عليها نلاحظ أن جميع القيم المعيارية أقل من (3.5) إذن، لا يوجد انتهاك لفرضية الاستقلال الموضعي للبيانات.

تشير نتائج التحليل التي تم إجراءها إلى تحقق شرط الاضطرادية الذي يؤكد أنه بزيادة مستوى السمة الكامنة لدى الفرد؛ تزداد احتمالية تقديم استجابة صحيحة، وهو ما يمكن تمثيله بيانيًا بمنحنى منحدر لأعلى يُعرف باسم منحنى خصائص الفقرة (ICC)، ويمكن تشبيه المنحنى بشكل حرف "S"، مما يشير إلى أن مستويات القدرة العالية تظهر احتمالات أعلى للاستجابات الصحيحة .أمّا فيما يتعلق بالافتراض الأخير وهو التحرر من السرعة؛ فإنّ بيانات الدّراسة الحالية؛ بيانات محاكاة وليست بيانات حقيقية، وبالتالي هي متحررة من هذا الافتراض بالأصل (Hambleton & Swaminathan, 1985).

الخطوة الرابعة: فحص جودة مطابقة البيانات المولدة للنماذج المستخدمة في الدراسة، ويعني تحديد مدى ملاءمة الفقرات للنموذج المستخدم في الدراسة، وبالتالي ضمان صدق وثبات الاختبارات. تم استخدام برنامج (R) بحزم إحصائية مختلفة لكل حجم من أحجام العينات الواردة في الدراسة للحكم على جودة مطابقة الفقرات لنموذج التقدير الجزئي المعمم؛ وبرنامج دلتا (DELTA_3) للحكم على جودة مطابقة الفقرات ثم الأفراد لنموذج دلتا لتقدير الدرجات.

أوّلًا: فحص جودة مطابقة الفقرات لنموذج التقدير الجزئي المعمم (GPCM) من خلال الحزمة الإحصائية (mirt)، وكان الحكم من خلال المؤشر الإحصائي (S-X^2) المقترح من قبل Orlando and Thissen (2000) وهو نسخة معدلة من المؤشر الإحصائي مربع كاي التقليدي (χ^2) إذ يتم استخدامه لتقييم مدى ملاءمة الفقرات ضمن نظرية الاستجابة للفقرة (IRT) من خلال مقارنة أنماط الاستجابة الملحوظة بالاستجابة المتوقعة بموجب النموذج المستخدم. في ضوء هذا المؤشر تُعد الفقرة غير مطابقة للنموذج إذا قلّت قيمة الدلالة الإحصائية (p.S-X^2) عن (0.05)، إذ أظهرت

دراسات المحاكاة Han, Sinharay, Johnson, and Liu (2023) أنه مع أحجام العينات (500 و1000 و4000) قد تقل معدلات الخطأ من النوع الأوّل مستوى الدلالة (0.05)، وهذا يشير إلى أنّ أحجام العينات الكبيرة قد يؤدي إلى ظهور مؤشرات غير ملائمة للفقرات؛ وبالتالي قد تظهر مؤشرات غير ملائمة للفقرات والنموذج المستخدم خصوصاً في العينات الكبيرة، وتكون قيمة المؤشر غير دالة عند مستوى الدلالة (0.05). ويعرض جدول (الملحق) قيم الدلالة الإحصائية ($p.S\text{-}X^2$) لكل فقرة من فقرات الدراسة ومن خلالها تم الحكم على جودة المطابقة.

أظهرت نتائج التحليل تفاوتًا في مؤشر المطابقة المستخدم؛ فنجد أن جميع الفقرات في الاختبارات (5، 10، 15) في عينة (500، 1000) كانت ملائمة لنموذج التقدير الجزئي المعمم حسب إحصائي الملاءمة ($S\text{-}X^2$) إذ أظهرت قيمة الدلالة الإحصائية لها قيم أكبر من (0.05) مما يحقق افتراض مؤشر حسن المطابقة للإحصائي($S\text{-}X^2$). إلّا أنّ قيمة الدلالة الإحصائية لمؤشر ($p.S\text{-}X^2$) للفقرات في عينة (5000) بدأت بالنزول عن مستوى الدلالة (0.05). فنجد أن مستوى الدلالة يُظهر أقل مستوى له في العينات الكبيرة والاختبارات الطويلة والعكس صحيح. فنجد أن مؤشر المطابقة يتحسن في عينات (500، 1000) في الاختبار المتضمن (5) فقرات. وكما تمت الإشارة أعلاه لحساسية هذا المؤشر للعينات الكبيرة، إذ نجد أنّ قيم المؤشر الإحصائي ($S\text{-}X^2$) تميل إلى أن يكون لها معدلات خطأ من النوع الأوّل مبالغ فيها قليلًا، خصوصًا في العينات الكبيرة وهذا يعني أنّها قد تظهر الفقرات - بشكل غير صحيح - على أنّها غير ملائمة أكثر من المتوقع.

جدول (4) مؤشر حسن الملاءمة للإحصائي ($p.S\text{-}X^2$) حسب نموذج التقدير الجزئي المعمم:

الفقرات	العينة	$p.S\text{-}X^2$
5	500	0.05>
	1000	0.05>
	5000	0.05≥
10	500	0.05>
	1000	0.05>
	5000	0.05≈
15	500	0.05>
	1000	0.05≈
	5000	0.05≤

ثانيا: فحص جودة ملاءمة الفقرات لنموذج دلتا لتقدير الدرجات (DSM).

يعتمد فحص جودة ملاءمة الفقرات لنموذج دلتا لتقدير الدرجات (DSM) حول مدى توافق الاستجابات الملاحظة مع قيمها المتوقعة بموجب نموذج دالة استجابة العنصر (IRF) عبر حساب متوسط الفرق المطلق Mean Absolute Difference (MAD) بين الدرجات المرصودة للفقرات التي تتراوح بين (0/1) واحتمال الاستجابة الصحيحة للأفراد على فقرات الاختبار التي تم الحصول عليها بموجب نموذج الدالة النسبية المستخدم (Dimitrov & Atanasov, 2021).

يمكن الحكم على جودة الفقرة من خلال عدد من المحكات، وضعها كل (Atanasov ،2021 Dimitrov &) ، فإذا كانت قيمة الإحصائي (MAD ≤ 0.07) فهذا يعني أن المطابقة جيدة، أمّا إذا كانت قيمة المؤشر تتراوح ما بين (0.07 < MAD < 0.10) فهذا يدل على مطابقة مقبولة، وأخيرًا إذا كانت قيمة (MAD ≥ 0.10) فتدل على ملاءمة ضعيفة. يوضح الجدول (5) نتائج تحليل المطابقة المستخرجة من بيانات الدراسة الحالية التي تدل على وجود توافق إلى حد كبير بين الاستجابات الملاحظة وقيمها المتوقعة بموجب نموذج دالة استجابة للفقرة (IRF)، إذ أظهرت قيم مؤشر متوسط الفرق المطلق (MAD) لـ (77) فقرة من أصل (90) قيمًا أقل من (0.07) ممّا يدل على مطابقة جيدة. بينما أظهرت (13) فقرة قيمًا تتراوح بين (0.07 < MAD < 0.10) وتشير إلى مطابقة مقبولة؛ بينما لم تظهر أي فقرة مطابقة ضعيفة.

الجدول (5): مؤشر حسن الملاءمة للإحصائي (MAD) حسب نموذج دلتا لتقدير الدرجات:

الفقرات	العينة	(MAD ≤ 0.07)	(0.07 < MAD < 0.10)	(MAD ≥ 0.10)
5	500	4	1	0
	1000	4	1	0
	5000	3	2	0
10	500	8	2	0
	1000	8	2	0
	5000	9	1	0
15	500	13	2	0
	1000	14	1	0
	5000	14	1	0

نتائج الدراسة وتفسيرها:

صممت هذه الدراسة بهدف المقارنة بين نموذجين من النماذج المستخدمة لتقدير معالم الفقرات والأفراد، وهما نموذج التقدير الجزئي المعمم، ونموذج دلتا لتقدير الدرجات. وذلك تحت ظروف تجريبية مختلفة تمثلت في ثلاثة أحجام مختلفة للعينات وهي كالآتي: (500، 1000، 5000) وثلاثة أطوال مختلفة للاختبارات وهي كالآتي (5، 10، 15) فقرة؛ تم تقدير معالم فقرات الاختبار المكونة من (5، 10، 15) فقرة، وتقدير استجابة (500، 1000، 5000) فرد عليها تبعا لنموذج (GPCM) باستخدام برنامج (R) وتقدير المعالم نفسها مرة أخرى وفقاً لنموذج (DSM-L) باستخدام برنامج (DELTA)؛ من خلال المتوسطات الحسابية والانحرافات المعيارية لكل ظرف من ظروف الدراسة (2x3x3). ثم تم حساب الخطأ المعياري في دقة تقدير معالم الأفراد والفقرات تبعا لنوع النموذج المستخدم والمقارنة بينهما باستخدام تحليل التباين المختلط (Mix design ANOVA) باستخدام برنامج (SPSS)، لدلالة الفروق في دقة التقدير لكل ظرف من ظروف الدراسة والتفاعل بينهما.

النتائج المتعلقة بالسؤال الأول: "هل هناك أثر لحجم عيّنة مكوّن من (500، 1000، 5000) فرد، واختبار مكوّن من (5، 10، 15) فقرة على دقة تقدير معلم صعوبة فئات الاستجابة وِفقًا لنموذج التقدير الجزئي المعمم (GPCM) ونموذج دلتا لتقدير الدرجات (DSM)؟"

للإجابة عن هذا السؤال تم تقدير صعوبة فئات الاستجابة الأربع (δ1، δ2 ، δ3، δ4)؛ من خلال برنامج (R) عبر طريقة التقدير بالأرجحية القصوى الهامشية (MMLE). تجدر الإشارة إلى أنه عند توليد البيانات تم إنشاء خمس فئات استجابة (0، 1، 2، 3، 4)، لكل فقرة من فقرات الدراسة (5، 10، 15) مما يعني وجود أربعة انتقالات محتملة بين الفئات. وبما أن نموذج التقدير الجزئي المعمم (GPCM) يقدّر معلمات الصعوبة بين فئات الاستجابة وليس للفقرة نفسها، فإن النموذج يقوم فعليًا بتقدير أربع معلمات فقط (عدد الفئات - 1). كما تم إجراء تحليل آخر من خلال برنامج (DELTA) لاستخراج تقديرات معالم صعوبة فئات الاستجابة وفقا لنموذج دلتا لتقدير الدرجات (DSM). وتوضح النتائج التي يعرضها الجدول (6) حجم العينة، طول الاختبار، ومعالم صعوبة فئات الاستجابة (δ)، بالإضافة إلى المتوسط الحسابي (*M*) والانحراف المعياري (*SD*) لكل عينة اختبار وفقا للنموذجين المتبعين في الدراسة.

جدول (6): الإحصاء الوصفي لخصائص معلمات صعوبة فئات الاستجابة المقدّرة.

		5			10			15		
		δ	M	SD	δ	M	SD	δ	M	SD
GPCM	500	1.65 ،-1.44	0.166	0.01	2.26 ،-2.29	0.106	1.08	2.91 ،-2.39	0.09	1.02
	1000	1.37 ،-1.73	0.133	0.94	2.67 ،-2.26	0.135	1.11	2.31 ،-1.96	-0.18	1.00
	5000	1.53 ،-2.08	-0.24	0.94	2.43 ،-1.96	0.04	0.96	1.63 ،-2.46	-0.24	0.91
DSM	500	0.86 ،0.4	0.36	0.27	0.94 ،0.01	0.35	0.28	0.90 ،0.01	0.35	0.27
	1000	0.91 ،0.4	0.40	0.28	0.93 ،0.02	0.36	0.27	0.89 ،0.01	0.35	0.27
	5000	0.89 ،0.1	0.37	0.27	0.90 ،0.01	0.36	0.27	0.89 ،0.01	0.35	0.27

تشير النتائج المتعلقة بنموذج التقدير الجزئي المعمم (GPCM) إلى أن المواقع الانتقالية لفئات الاستجابة تأخذ مكانها بين (2.46-، 2.91) وهي قيم ضمن نطاق الصعوبة المقبول إذ يتراوح المدى الذي اقترحه Hambleton & Swaminathan (1985) بين (3.0- ≤ δ ≤ 3.0+)؛ ويُعتمد هذا المدى لتحديد معاملات صعوبة فئات الاستجابة لأنه يغطي معظم انتشار مستويات القدرة المفترضة (الموزعة طبيعيًا). إذ أشار Embretson & Reise (2000) إلى أن معلمات صعوبة فئات الاستجابة غالبًا ما تقع في النطاق ما بين (-3 و +3) وذلك لأن هذا النطاق يغطي غالبية التوزيع الطبيعي للسمات الكامنة لدى المفحوصين كما تشير الدراسة إلى أنّ تجاوز هذا النطاق قد يشير إلى وجود مشكلات في الفقرة.

أما النتائج المتعلقة بنموذج دلتا لتقدير الدرجات (DSM) فتشير إلى أن قيم معاملات الصعوبة لفئات الاستجابة تأخذ مكانها بين (0.01، 0.91) وتعد هذه القيم ضمن نطاق مدى الصعوبة المقترح في الاطار الكامن لنموذج دلتا لتقدير الدرجات؛ إذ يتراوح تدريج دلتا بين (1 ≤ δ ≤ 0)؛ كما هو الحال في نموذج التقدير الجزئي المعمم (GPCM) فإن قيم معامل الصعوبة للخطوات في نموذج دلتا تأخذ ترتيباً تصاعدياً وتعكس تدرجًا منطقيًا في الفئة؛ كما يظهر في الجدول (6) في الجزء المتعلق بنموذج دلتا لتقدير الدرجات، إذ تشير قيمة معامل الصعوبة للموقع الانتقالي الأول في الفقرة الأولى (11δ = 0.4)وتأخذ بقية معاملات الصعوبة في الصعود إلى أن تبلغ قيمة معامل صعوبة الموقع الانتقالي الرابع (14δ = 0.86) بالتالي تزداد احتمالية الانتقال من فئة استجابة إلى أخرى كلما تجاوزت القدرة الكامنة عتبة معينة وهو ما يتماشى مع الانتقال من (0) إلى (1).

تم تقييم دقة التقديرات المستخرجة في كلا النموذجين (GPCM و DSM) فقد تم مقارنة القيم الحقيقية التي تم الحصول عليها عند توليد البيانات بالقيم المقدّرة بالنماذج المعمول بها في الدراسة وذلك بحساب الفرق المتوسط بين القيمة التقديرية والقيمة الحقيقية وهو ما يعرف بالتحيّز (*Bias*)، ومن خلال حساب متوسط مربع الفرق (*MSE*) بين القيم نفسها، بالإضافة الى معامل ارتباط بيرسون الذي يقيس مدى قوة العلاقة بين القيم الحقيقية والقيم المقدّرة؛ تم حساب هذه المؤشرات لكل حالة من الحالات التجريبية (18)، كما يظهر الجدول (7) النتائج:

جدول (7): متوسطات مؤشرات دقة تقدير معلمات صعوبة فئات الاستجابة.

		15			10			5		
		Bias	*MSE*	*COR*	*Bias*	*MSE*	*COR*	*Bias*	*MSE*	*COR*
GPCM	*500*	0.024	0.035	0.97	-0.032	0.120	0.94	0.011	0.011	0.99
	1000	-0.026	0.021	0.99	0.083	0.017	0.99	0.066	0.022	0.99
	5000	-0.021	0.011	0.99	-0.043	0.009	0.99	0.013	0.001	0.99
DSM	*500*	0.362	0.833	0.65	0.230	0.567	0.93	0.196	0.590	0.93
	1000	0.506	0.654	0.96	0.310	0.844	0.75	0.304	0.498	0.92
	5000	0.577	0.668	0.91	0.363	0.516	0.96	0.603	0.856	0.89

يتبين من الجدول (7) أن نموذج التقدير الجزئي المعمم (GPCM) يتفوق في تقدير معلمات صعوبة فئات الاستجابة "عبر جميع مستويات الدراسة" على نموذج دلتا لتقدير الدرجات (DSM)؛ فقد سجل نموذج التقدير الجزئي المعمم (GPCM) قيماً منخفضة لـ (Bias) في جميع الحالات مما يعني عدم وجود تحيز في دقة تقدير المعلمات، وتشير الأبحاث إلى أن انخفاض هذا المؤشر يدل على قدرة النموذج على إعطاء تقديرات دقيقة غير متحيزة للمعلمات الحقيقية (Embretson & Reise, 2000) . كما هو الحال في مؤشر (MSE) الذي يظهر متوسط مربع الفروقات بين التقديرات والقيم الحقيقية؛ وكما يظهر في الجدول أعلاه فإن جميع القيم تظهر قيماً قريبة للصفر بالتالي تقديرات أدق؛ كما أن معاملات الارتباط مرتفعة ومستقرة، إذ بلغت في معظم الحالات (0.99) مما يدل على وجود ارتباط قوي جدًا بين التقديرات والقيم الحقيقية. كما أظهر نموذج دلتا لتقدير الدرجات (DSM) انحيازًا أكبر ومعدلات خطأ أعلى في مؤشري (Bias، MSE) على الترتيب، خصوصا عند زيادة عدد الفقرات، في

حين انخفض معامل الارتباط إلى (0.65) عند حجم العينة (500) وعدد فقرات (5)، مما يدل على ارتباط ضعيف بين القيم التقديرية والقيم الحقيقية.

تشير النتائج إلى تفوق نموذج التقدير الجزئي المعمم (GPCM) على نموذج دلتا لتقدير الدرجات (DSM) في دقة تقدير معلمات صعوبة الفقرات عبر مختلف مستويات العينة وعدد الفقرات. فقد سجّل GPCM قيمًا منخفضة لمؤشري التحيز (Bias) ومتوسط مربع الخطأ (MSE)، مما يعكس دقة تقدير عالية وغير متحيزة، إضافة إلى معاملات ارتباط مرتفعة وصلت إلى (0.99) مما يشير إلى توافق كبير بين القيم التقديرية والحقيقية. بالمقابل، أظهر نموذج DSM انحيازًا أكبر وأخطاء تقديرية أعلى، خصوصًا مع ازدياد عدد الفقرات، كما انخفض معامل الارتباط إلى (0.65) عند حجم عينة (500) وعدد فقرات (5)، مما يدل على ضعف في الدقة التقديرية في هذه الحالة. أما فيما يتعلق في معاملات صعوبة فئات الاستجابة فقد ظهرت في كلا النموذجين ضمن النطاق المقبول وهو ما يتماشى مع النطاق المقترح من قبل Hambleton & Swaminathan (1985) و Embretson & Reise (2000). كما أظهرت النتائج ترتيبًا تصاعديًا منطقيًا في معاملات الصعوبة داخل الفقرات، ما يعكس سلوكًا تدريجيًا مناسبًا في الانتقال بين فئات الاستجابة، مما يعزز من صحة النموذج من حيث التدرج والاستجابة التراكمية للقدرة الكامنة.

النتائج المتعلقة بالسؤال الثاني: هل هناك أثر للتفاعل بين حجم عيّنة مكوّن من (500، 1000، 5000) فرد، واختبار مكوّن من (5، 10، 15) فقرة على تقدير معلم الصعوبة وِفقًا لنموذج التقدير الجزئي المعمم (GPCM) ونموذج دلتا لتقدير الدرجات (DSM)؟

وللتحقق من نتائج السؤال الثاني تم استخدام الاختبار الإحصائي تحليل التباين الثلاثي (Three Way ANOVA)، وذلك بعد فحص الافتراضات الأساسية لهذا الاختبار المتمثلة في اختبار (Levene's Test) لتجانس التباين في الخلايا، وتظهر قيمة (*p*) فيه أقل من (.05) لذا فإننا نرفض الفرضية الصفرية التي تقول إن التباين متساو؛ أي أن هذا الفرض لم يتحقق، ولكن تشير المراجع الإحصائية إلى أنه إذا كانت اعداد الأفراد بين الخلايا متساوية - كما هو الحال في البيانات الحالية - فإن عدم تحقق هذا الفرض لا يؤثر كثيرا على النتائج (دودين، 2009).

يعرض الجدول (8) نتائج تحليل التباين الثلاثي لتقديرات معالم الصعوبة وفقا لنموذجي (التقدير الجزئي المعمم، دلتا لتقدير الدرجات) وطول الاختبار (5، 10، 15)، واخيرً حجم العينة (500، 1000، 5000، 10000) بالإضافة إلى التفاعل الثنائي والثلاثي بين المتغيرات.

جدول (8): نتائج تحليل التباين الثلاثي لمتغيرات الدراسة والتفاعل بينهما.

المتغير	df	قيمة (F)	Sig.	Partial η^2
طول الاختبار	2	1.596	0.203	0.003
حجم العينة	3	1.262	0.286	0.004
النموذج	1	53.45	0.000	0.054
طول الاختبار*حجم العينة	6	0.279	0.947	0.002
طول الاختبار*النموذج	2	1.440	0.237	0.003
حجم العينة* النموذج	3	1.208	0.306	0.004
طول الاختبار*حجم العينة* النموذج	6	0.199	0.977	0.001

يتبين من الجدول (8) أنه لا يوجد تأثير رئيس دال إحصائيًا عند مستوى الدلالة (.05) لمتغير طول الاختبار على متوسطات تقديرات معالم صعوبة فئات الاستجابة، فقد بلغت قيمة اختبار(F) المحسوبة (F (1.936) = 1.596, p =. 203) مع حجم أثر متوسط يساوي (Partial η^2 =.003) وهو نسبة التباين في المتغير التابع التي يمكن تفسيرها بواسطة طول الاختبار وهي نسبة صغيرة حسب كوهن (1988) بالتالي لا يوجد فروق ذات دلالة إحصائية لمتغير طول الاختبار (5، 10، 15) فقرة على دقة تقدير معالم صعوبة فئات الاستجابة. وبالمثل لا يوجد هناك تأثير رئيس دال إحصائيًا عند مستوى الدلالة (.005) لمتغير حجم العينة على متوسطات تقديرات معالم صعوبة فئات الاستجابة المُقدرة بنموذجي التقدير الجزئي المعمم ودلتا لتقدير الدرجات فقد بلغت قيمة اختبار(F) المحسوبة (F (2.936) = 1.262, p = .286)مع حجم أثر متوسط (Partial η^2 = 0.004) وهو قيمة التباين في المتغير التابع التي يمكن تفسيرها بواسطة حجم العينة وهي نسبة صغيرة حسب Cohen (1988) إذن؛ لا توجد فروق ذات دلالة إحصائية لمتغير حجم العينة (500، 1000، 5000) على دقة تقدير معالم صعوبة فئات الاستجابة؛ مما يشير إلى أن هذه العوامل لم تُحدث فرقًا جوهريًا على دقة تقدير قيم معالم صعوبة فئات الاستجابة.

بينما أظهرت النتائج وجود أثر دال إحصائيًا لمتغير (النموذج) نموذجي التقدير الجزئي المعمم ودلتا لتقدير الدرجات على متوسطات تقديرات معالم صعوبة فئات الاستجابة، فقد بلغت قيمة اختبار(F) المحسوبة (F(1,936) = 53.45, p < 0.000)، مع حجم أثر متوسط (Partial η^2 = 0.054) وتشير إلى نسبة التباين في المتغير التابع التي يمكن تفسيرها بواسطة نوع النموذج، وهي نسبة كبيرة حسب

مؤشر Cohen (1988) مما يدل إلى أن نماذج التقدير المعمول بها في الدراسة الحالية تختلف من حيث قدرتها على تقدير معالم صعوبة فئات الاستجابة.

كما أظهر الجدول عدم وجود دلالة إحصائية لجميع التفاعلات الثنائية والثلاثية فقد كانت جميع القيم الاحتمالية أكبر من (0.05). فعند النظر إلى متغيري طول الاختبار وحجم العينة؛ نجد أن قيمة اختبار (F) المحسوبة تشير إلى ($F(6, 936) = 0.279, p = .947$)؛ بالمثل تبين قيمة (F) المحسوبة للتفاعل بين طول الاختبار والنموذج إلى ($F(2, 936) = 1.440, p = .237$)؛ كما تسجل قيمة (F) المحسوبة للتفاعل الأخير بين حجم العينة والنموذج إلى ($F(3, 936) = 1.208, p = .306$) وهذا ما يدل على عدم وجود تأثير مشترك بين المتغيرات الثنائية، وفي نفس السياق نجد أيضا أن التفاعلات الثلاثية لم تؤثر بشكل دال إحصائيًا على معالم الصعوبة المقدَّر، حيث لم تكن هناك فروق ذات دلالة إحصائية بين المجموعات المختلفة الناتجة عن هذه التفاعلات كما أظهرت قيمة (F) المحسوبة ($F(6, 936) = 0.196, p$.(= .977

للإجابة عن السؤال الثالث هل هناك أثر لحجم عيّنة مكوّن من (500، 1000، 5000) فرد، واختبار مكون من (5، 10، 15) فقرة على دقة تقدير معلم التّمييز وِفقًا لنموذج التقدير الجزئي المعمم (GPCM) ونموذج دلتا لتقدير الدرجات (DSM)؟

تم تقدير معلم التّمييز وِفقًا لنموذج التقدير الجزئي المعمم (GPCM) من خلال برنامج (R) واستخرج معالم تمييز لكل فقرة من فقرات الاختبار؛ بينما تم تقدير معلم التّمييز وِفقًا لنموذج دلتا لتقدير الدرجات (DSM) من خلال برنامج (DELTA) وهو بطبيعة الحال يستخرج معالم تمييز لفئات الاستجابة ؛ ولكي تتم المقارنة بين معاملات التمييز في كلا النموذجين بشكل سليم تم تقدير معلم التمييز لكل فقرة في نموذج دلتا لتقدير الدرجات من خلال حساب متوسطات تمييز كل فئة استجابة، وبذلك أصبح كل معامل تمييز في نموذج التقدير الجزئي المعمم يقابله معامل تمييز في نموذج دلتا لتقدير الدرجات. يعرض الجدول (9) المتوسط الحسابي لقيم معامل التمييز (M) وانحرافه المعياري (SD) لكل حجم عينة، وطول الاختبار وفقا للنموذجين المتبعين في الدراسة.

جدول (9): الإحصاء الوصفي لخصائص معلمات التمييز المقدّرة.

		5		10		15	
		M	SD	M	SD	M	SD
GPC M	500	1.62	1.06	1.16	0.63	1.50	0.64
	1000	1.29	0.47	1.82	0.61	1.58	-0.56

		15		10		5	
		SD	*M*	*SD*	*M*	*SD*	*M*
DSM	*5000*	0.49	1.47	0.46	1.94	0.81	1.39
	500	0.74	4.05	1.00	3.14	1.76	4.34
	1000	0.74	4.05	0.88	4.55	0.89	3.71
	5000	1.05	3.59	0.86	4.54	1.49	3.65

يتبين من الجدول (9) أن متوسط قيم معالم التمييز (α) المتعلقة بنموذج التقدير الجزئي المعمم (GPCM) تنحصر بين (1.29، 1.82) وهي قيم خارجة بقليل عن نطاق قيم معامل التمييز المقبول في نموذج التقدير الجزئي المعمم (GPCM) إذ يعد Hambleton, Swaminathan, and Rogers (1991) أن قيم معامل التمييز الواقعة بين (0.5 و 2.0) مقبولة في معظم مواقف الاختبار؛ إلا أنه لم تُظهر النتائج معامل تمييز سالبة أو أقل من (0.4) إذ تشير الفقرات ذات معامل التمييز أقل من (0.4) إلى أن الفقرة لا تمييز جيدًا بين الأفراد ذوي القدرات المختلفة وقد تحتاج إلى حذف أو مراجعة؛ بالتالي تُعد جميع معاملات التمييز المستخرجة عبر نموذج التقدير الجزئي المعمم مقبولة ومناسبة.

كما يتبين أيضا من الجدول (9) أن متوسط قيم معالم التمييز (α) المتعلقة بنموذج دلتا لتقدير الدرجات (DSM) تراوحت بين (3.14، 4.55) وهي قيم أكبر بكثير من مدى التمييز المقبول في نموذج دلتا لتقدير الدرجات إذ يتبع النموذج مبادئ مشابهة لما هو متبع في نماذج نظرية الاستجابة للفقرة (IRT) بينما تشير الفقرات ذات معامل التمييز أعلى من (3.00) إلى مشكلة في الفقرة تتمثل في صعوبة مفرطة، بالتالي عدم قدرة الفقرة على التمييز بين الأفراد.

ولتقييم دقة التقديرات المستخرجة في كلا النموذجين (GPCM و DSM) فقد تم مقارنة القيم الحقيقية التي تم الحصول عليها عند توليد البيانات بالقيم المقدّرة بالنموذجين السابقين، وذلك من خلال حساب الفرق المتوسط بين القيمة التقديرية والقيمة الحقيقية وهو ما يعرف بالتحيّز (*Bias*)، ومن خلال حساب متوسط مربع الفرق (*MSE*) بين القيم نفسها، بالإضافة إلى معامل ارتباط بيرسون الذي يقيس مدى قوة العلاقة بين القيم الحقيقية والقيم المقدّرة؛ وقد تم حساب هذه المؤشرات لكل حالة من الحالات التجريبية (18)، كما يظهر الجدول (10) النتائج:

جدول (10): متوسطات مؤشرات دقة تقدير معلم التمييز.

		5			10			15		
		Bias	MSE	COR	Bias	MSE	COR	Bias	MSE	COR
GPCM	500	0.17	0.15	0.97	-0.07	0.01	0.98	-0.14	0.03	0.99
	1000	-0.07	0.00	0.99	-0.12	0.04	0.97	-0.18	0.04	0.99
	5000	-0.02	0.00	0.99	-0.3	0.10	0.99	-0.2	0.05	0.99
DSM	500	2.9	11.0	0.12	1.96	4.44	0.59	2.40	6.78	0.05
	1000	2.4	6.17	0.69	2.60	8.22	-0.23	2.29	6.31	-0.21
	5000	2.2	5.8	0.76	2.29	6.05	0.25	1.91	4.75	0.21

يتبين من الجدول (10) أنه في نموذج التقدير الجزئي المعمم (GPCM) كانت قيم التحيّز (*Bias*) عبر جميع الفقرات سلبيَّة؛ مما يشير إلى أن تقديرات معلم التمييز تميل إلى أن تكون أقل من القيم الحقيقية. أما بالنسبة لمؤشر *(MSE)*، فقد أظهر قيماً منخفضة نسبيًا، مما يدل على أن نسبة الأخطاء في تقدير التمييز كانت محدودة. وأما فيما يتعلق بمعاملات الارتباط؛ فقد كانت مرتفعة، مما يدل على دقة تقدير معقولة إلى حد كبير. بالتالي يمكن القول أن عملية تقدير معلم التمييز كانت دقيقة وفقًا لنموذج التقدير الجزئي المعمم (GPCM) (Dai et al., 2021).

كما يظهر في الجدول (10) أن نتائج التقدير بنموذج دلتا لتقدير الدرجات سجلت متوسطات تحيز (*Bias*) مرتفعة في عدد كبير من الحالات؛ وهو ما يشير إلى أن النموذج لم يعط نتائج دقيقة. أما بالنسبة لمؤشر *(MSE)* الذي يُفضل أن يكون منخفضًا؛ فقد أعطى نتائج مرتفعة مما يشير الى تشتت ملحوظ في مؤشر تباين الخطأ. وأما فيما يتعلق بمعاملات الارتباط فقد كانت متذبذبة في كثير من الحالات. وبالتالي يمكن القول إن عملية تقدير معلم التمييز كانت غير دقيقة وفقًا لنموذج دلتا لتقدير الدرجات (DSM).

للإجابة عن السؤال الرابع المتعلق هل هناك أثر للتفاعل بين حجم عيّنة مكوّن من (500، 1000، 5000) فرد، واختبار مكون من (5، 10، 15) فقرة على تقدير معلم التّمييز وِفقًا لنموذج التقدير الجزئي المعمم (GPCM) ونموذج دلتا لتقدير الدرجات (DSM)؟

للتحقق من نتائج السؤال الرابع تم استخدام الاختبار الإحصائي تحليل التباين الثلاثي (Three Way ANOVA)، وذلك بعد فحص الافتراض الأساسي لهذا الاختبار المتمثل في اختبار (Levene's Test) لتجانس التباين في الخلايا، الذي أشار إلى تحقق هذا الافتراض عند مستوى الدلالة (0.05) إذ

سجلت قيمة اختبار (Levene) لجميع مستويات الدراسة قيمًا أكبر من (0.05)، ما يعني أننا نفشل في رفض الفرضية الصفرية، وبناءً عليه لا توجد فروق ذات دلالة إحصائية في تباين الخطأ بين المجموعات المختلفة. يعرض الجدول (11) نتائج تحليل التباين الثلاثي لتقديرات معالم التمييز وفقا لنموذجي (التقدير الجزئي المعمم، ودلتا لتقدير الدرجات) وطول الاختبار (5، 10، 15) وأخيرًا حجم العينة (500، 1000، 5000) بالإضافة إلى التفاعل الثنائي والثلاثي بين المتغيرات.

أُجري تحليل التباين الثلاثي لاختبار أثر نموذج التقدير، وعدد الفقرات وحجم العينة، بالإضافة إلى التفاعلات فيما بينها على دقة تقدير معلم التمييز والجدول (11) يوضح ذلك:

جدول (11): نتائج تحليل التباين الثلاثي لمتغيرات الدراسة والتفاعل بينهما.

المتغير	df	قيمة (F)	Sig.	Partial η^2
طول الاختبار	2	3.942	0.021	0.035
حجم العينة	3	1.741	0.160	0.024
النموذج	1	436.960	0.000	0.669
طول الاختبار* حجم العينة	6	6.402	0.000	0.151
طول الاختبار* النموذج	2	0.966	0.382	0.009
حجم العينة * النموذج	3	0.194	0.901	0.003
عدد الأسئلة * حجم العينة * النموذج	6	1.140	0.340	0.031

يبين الجدول (11) وجود تأثير رئيس دال إحصائيًا عند مستوى الدلالة (.05) لمتغير طول الاختبار على تقدير معلم التمييز فقد بلغت قيمة اختبار(F) المحسوبة ($F(2.219) = 3.942, p < .021$) مع حجم أثر متوسط يساوي (Partial $\eta^2 = 0.035$) أي أن التباين في المتغير التابع الذي يمكن تفسيره بواسطة طول الاختبار تبلغ نسبته (3.5%) وهي نسبة مقبولة نسبيا حسب Cohen(1988) بالتالي توجد فروق ذات دلالة إحصائية لمتغير طول الاختبار على دقة تقدير معلم التمييز. بينما أظهرت النتائج عدم وجود تأثير رئيس دال إحصائيًا عند مستوى الدلالة ($p < 0.005$) لمتغير حجم العينة على تقدير معلم التمييز المُقدر بنموذجي التقدير الجزئي المعمم ودلتا لتقدير الدرجات، فقد بلغت قيمة اختبار(F) المحسوبة ($F(3.219) = 1.741, p < 0.160$) مع حجم أثر متوسط (Partial $\eta^2 = 0.024$) وهو قيمة التباين في المتغير التابع التي يمكن تفسيرها بواسطة حجم العينة، وتشير إلى نسبة صغيرة حسب Cohen(1988) إذن؛ لا توجد فروق ذات دلالة إحصائية لمتغير حجم العينة على تقدير معلم التمييز؛ مما يشير إلى أن هذا العامل لم يحدث فرقًا جوهريًا على تقدير قيم معالم التمييز.

بينما أظهرت النتائج أن النموذج المستخدم لتقدير المعالم كان له أثر دال إحصائيًا على تقدير معامل التمييز، إذ سجلت قيمة (F = 436.960) عند مستوى دلالة (p < 0.001)، مع قيمة (Partial η^2 == 0.669)مما يعني أن النموذج المستخدم يفسر بنسبة (66.9%) من التباين في معامل التمييز، وهو تأثير كبير يشير إلى فروق ذات دلالة إحصائية بين نموذجي التقدير الجزئي المعمم ودلتا لتقدير الدرجات ($F(1,216) = 436.960, p < .001$)، من حيث قدرتها على تقدير معالم التمييز.

أما فيما يتعلق بالتفاعل بين المتغيرات المستقلة؛ فقد أظهرت النتائج وجود تفاعل دال احصائيًا عند مستوى ($p < 0.005$) بين متغيري حجم العينة وطول الاختبار، فقد سجلت قيمة اختبار (F) المحسوبة ($F (6,216) = 6.402, p < .001$) مع قيمة تأثير متوسطة (Partial $\eta^2 = 0.151$)، مما يدل على أن تأثير أحد هذين العاملين يعتمد على مستوى العامل الآخر.

أظهرت نتائج اختبار المقارنات البعدية (Tukey HSD) عدم وجود فروق دالة إحصائيًا بين مستويات متغير حجم العينة (500، 1000، 5000)، من حيث تأثيره على تقدير معلم التمييز، إذ كانت جميع القيم الاحتمالية أكبر من مستوى الدلالة($p = 0.05$) ؛ وهذه النتائج تشير إلى أن زيادة حجم العينة لم يسفر عن تحسن دقة تقدير معلم التمييز.

كما أوضحت النتائج أن طول الاختبار يؤثر بشكل دال إحصائيًا على متوسطات تقدير معامل التمييز إذ وُجدت فروق ذات دلالة إحصائية بين الاختبار المكوّن من 10 فقرات والاختبار المكوّن من 15 فقرة، إذ بلغ متوسط الفرق ($M = 0.302$) بالدلالة الإحصائية ($p = .031$) وهي أقل من مستوى الدلالة ($p = 0.05$)، مما يشير إلى أن الاختبار الأطول (15 فقرة) أسهم في تقدير أعلى لمعامل التمييز مقارنة بالاختبار ذي (10 فقرات). في المقابل، لم تظهر فروق دالة بين الاختبار ذي (5 فقرات) والاختبار ذي (10 فقرات) بالتالي يمكن الاستنتاج أن الزيادة في عدد فقرات الاختبار تُحدث فرقًا في دقة تقدير معلم التمييز .تدعم هذه النتائج الأدبيات التي تشير إلى أن زيادة طول الاختبار قد تُحسِّن من دقة تقدير معلمات الفقرات (Hambleton & Swaminathan, 1985؛ Embretson & Reise, 2000).

في المقابل أظهرت النتائج عدم وجود تفاعلات ثنائية دالة إحصائيًا بين بقية المتغيرات على تقدير معلم التمييز؛ فقد سجلت بقية التفاعلات الثنائية قيماً احتمالية أكبر من ($p > .05$) مما يشير إلى عدم وجود أثر لاحد المتغيرات المستقلة على متغير مستقل آخر. وفي السياق نفسه؛ نجد أيضا أن التفاعلات الثلاثية لم تؤثر بشكل دال إحصائيًا على معالم التمييز المقدَّر، إذ لم تكن هناك فروق

ذات دلالة إحصائية بين المجموعات المختلفة الناتجة عن هذه التفاعلات كما أظهرت قيمة (F) المحسوبة ($F(6, 216) = 1.140, p = .340$).

نتائج الدراسة:

1) أظهرت النتائج تفوّق نموذج التقدير الجزئي المعمم (GPCM) - في دقة تقدير معلم الصعوبة - على نموذج دلتا لتقدير الدرجات (DSM) عبر جميع مستويات الدراسة. فقد حقق نموذج التقدير الجزئي المعمم (GPCM) أقل قيم للتحيز (Bias) ومتوسط مربع الخطأ (MSE)، مما يشير إلى تقديرات دقيقة وغير متحيزة، إضافة إلى معاملات ارتباط مرتفعة مع القيم الحقيقية بلغت (0.99) في معظم الحالات. في المقابل، أظهر نموذج دلتا لتقدير الدرجات (DSM) أداءً أقل دقة مع قيم تحيز أكبر وأخطاء تقدير أعلى، وانخفاض في معامل الارتباط وصل إلى (0.65)، مما يعكس ضعفًا في دقة التقدير تحت تلك الظروف.
2) أظهرت نتائج تحليل التباين الثلاثي (Three Way ANOVA) أن نوع النموذج المستخدم في تقدير معالم صعوبة فئات الاستجابة له أثر دال إحصائيًا، إذ تفوق نموذج التقدير الجزئي المعمم (GPCM) على نموذج دلتا لتقدير الدرجات (DSM) من حيث دقة التقدير. في المقابل، لم يكن لكل من طول الاختبار وحجم العينة أو تفاعلاتها مع النموذج أثر معنوي على دقة التقدير، بالرغم من الفروقات التي لوحظت في التحليلات الوصفية.
3) أظهرت النتائج أن متوسط التحيز في نموذج التقدير الجزئي المعمم (GPCM) كان سلبيًا في المجمل، مما يشير إلى ميل التقديرات لأن تكون أقل من القيم الحقيقية، كما كانت قيم متوسط مربع الخطأ (MSE) منخفضة نسبيًا، مما يدل على محدودية الخطأ في التقدير، وأظهرت معاملات الارتباط قيمًا مرتفعة، مما يعكس دقة تقدير لمعلمة التمييز باستخدام نموذج التقدير الجزئي المعمم (GPCM)؛ أما نموذج دلتا لتقدير الدرجات (DSM)، فقد سجل قيم تحيز مرتفعة في معظم الحالات، وقيم متوسط مربع الخطأ (MSE) أعلى، ومعاملات الارتباط غير مستقرة، مما يعكس ضعف دقة تقدير معلم التمييز في هذا النموذج.
4) أظهرت نتائج تحليل التباين الثلاثي (Three Way ANOVA) أن نموذج التقدير كان العامل الأكثر تأثيرًا في تقدير معلم التمييز، إذ كان له تأثير كبير ودال إحصائيًا، فقد فسر حوالي (66.9%) من التباين في التقديرات. كما كان لطول الاختبار تأثير دال إحصائيًا بنسبة (3.5%)، بينما لم يظهر حجم العينة تأثيرًا ذا دلالة إحصائية على تقدير معلم التمييز. أما بالنسبة للتفاعلات بين

المتغيرات، فقد كان التفاعل بين طول الاختبار وحجم العينة ذو دلالة إحصائية، مما يشير إلى أن التفاعل بين هذين العاملين يؤثر على تقدير معلمة التمييز. أما التفاعلات الأخرى فلم تكن ذات دلالة إحصائية.

التوصيات:

1) توصي الدراسة باستخدام نموذج التقدير الجزئي المعمم (GPCM) عند تقدير معلمات الفقرات، سواء معلم الصعوبة أم معلم التمييز، وفقا لما أظهرت نتائج دقة التقدير وانخفاض في مستويات الانحياز وفرق مربع الخطأ، مقارنة بنموذج دلتا لتقدير الدرجات (DSM)، خصوصًا في الدراسات التي تتطلب دقة عالية في التقدير.
2) تشير النتائج إلى أن نوع النموذج هو العامل الأكثر تأثيرًا في دقة التقدير، لذلك توصي الدراسة الباحثين والممارسين التربويين باختيار نموذج التقدير بناءً على خصائص البيانات وطبيعة الفقرات.

المقترحات لدراسات مستقبلية:

1) اجراء المزيد من الدراسات على نموذج دلتا لتقدير الدرجات في الإطار الكامن (DSM-L) باستخدام بيانات حقيقية ومولدة لاختبارات طويلة أو لعينات كبيرة، إذ أظهرت النتائج تراجعًا في دقة التقدير في هذه الحالات.
2) توصي الدراسة بإجراء دراسات مستقبلية تهتم بالمقارنة بين نموذج دلتا لتقدير الدرجات ونماذج نظرية الاستجابة للفقرة متعددة التدريج مثل نموذج الاستجابة الاسمية ونموذج التقدير الجزئي المعمم.

قائمة المراجع العربية والانجليزية

اولاً: المراجع العربية:

بني عطا، ز. ص. إ. (2017). تقصي أثر طول الاختبار وحجم العينة على دقة طرق تقدير معالم الفقرات وقدرات الأفراد في برنامج بايلوج. *المجلة الدولية للبحث في التربية وعلم النفس*، *5*(2)، 579.606–

دي أيالا، آر .(2017). *النظرية والتطبيق في نظرية الاستجابة للفقرة*. دار جامعة الملك سعود . https://doi.org/10.33948/1158-030-002-008

دودين، ح. م .(2009). *التحليل الإحصائي المتقدم للبيانات باستخدام SPSS*. عمّان: دار الحامد للنشر والتوزيع.

Arabic References

Banī ʿAṭā, Z. Ṣ. I. (2017). taqaṣṣī Athar Ṭūl al-ikhtibār wa-ḥajm al-ʿayyinah ʿalá diqqat Ṭuruq taqdīr Maʿālim al-faqarāt wqdrāt al-afrād fī Barnāmaj bāylwj. *al-Majallah al-Dawlīyah lil-Baḥth fī al-Tarbiyah wa-ʿilm al-nafs, 5* (2), 579 – 606.

Dī ayālā, Ār. (2017). *al-naẓarīyah wa-al-taṭbīq fī Naẓarīyat al-istijābah llfqrh*. Dār Jāmiʿat al-Malik Saʿūd. https : / / doi. org / 10. 33948/1158-030-002-008

Dūdīn, Ḥ. M. (2009). al-Taḥlīl al-iḥṣā'ī almtqdm llbyānāt bi-istikhdām SPSS. ʿAmmān : Dār al-Ḥāmid lil-Nashr wa-al-Tawzīʿ.

ثانياً:المراجع الإنجليزية:

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561–573. https://doi.org/10.1007/BF03037732

Auné, S. E., Abal, F. J. P., & Attorresi, H. F. (2020). A psychometric analysis from the Item Response Theory: Step-by-step modelling of a Loneliness Scale. *Ciencias Psicológicas, 14*(1), e-2179. https://doi.org/10.22235/cp.v14i1.2179 (تم تعديل الترتيب)

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for dichotomously scored items. *Psychometrika, 35*(2), 179–197. https://doi.org/10.1007/BF02291262

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates.

Dai, S., Vo, T. T., Kehinde, O. J., He, H., Xue, Y., Demir, C., & Wang, X. (2021). Performance of Polytomous IRT Models With Rating Scale Data: An Investigation Over Sample Size, Instrument Length, and Missing Data. *Frontiers in Education,* 6. https://doi.org/10.3389/feduc.2021.721963

Dimitrov, D. M. (2016). An approach to scoring and equating tests with binary items: Piloting with large-scale assessments. *Educational and Psychological Measurement, 76*(6), 954–975. https://doi.org/10.1177/0013164416631100

Dimitrov, D. M., & Alsadaawi, A. (2018). Psychometric features of the General Teacher Test under the D-scoring model: The case of teacher certification assessment in Saudi Arabia. *World Journal of Social Science Research, 5*(2), 107–122. https://doi.org/10.22158/wjssr.v5n2p107

Dimitrov, D. M., & Atanasov, D. V. (2021). Latent D-scoring modeling: Estimation of item and person parameters. *Educational and Psychological Measurement, 81*(2), 388–404. https://doi.org/10.1177/0013164420941147

Dimitrov, D. M., & Luo, Y. (2019). A note on the D-scoring method adapted for polytomous test items. *Educational and Psychological Measurement, 79*(3), 545–557. https://doi.org/10.1177/0013164418786014

Dimitrov, D. M., Atanasov, D. V., & Luo, Y. (2020). Person-fit assessment under the D-scoring method. *Measurement: Interdisciplinary Research and Perspectives, 18*(3), 111–123. https://doi.org/10.1080/15366367.2020.1725733

Djidu, H , Heri Retnawati, H & Haryanto H. (2023). Ensuring Parameter Estimation Accuracy in 3PL IRT Modeling: The Role of Test Length and Sample Size. *JP3I (Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia), 12*(2), 177–190. https://doi.org/10.15408/jp3i.v12i2.34130

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Lawrence Erlbaum Associates Publishers.

Han, Z., Sinharay, S., Johnson, M. S., & Liu, X. (2023). The standardized S-X2 statistic for assessing item fit. *Applied Psychological Measurement, 47*(1), 3–18. https://doi.org/10.1177/01466216221108077

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer-Nijhoff Publishing.

Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in Psychology, 7*, Article 109. https://doi.org/10.3389/fpsyg.2016.00109

Lord, F. M. (1952). *A theory of test scores* (Psychometric Monograph No. 7). Psychometric Society.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174. https://doi.org/10.1007/BF02296272

Mills, C. N. (2002). Computerized Simulation in Research and Testing. *Applied Psychological Measurement, 26*(3), 217–231. https://doi.org/10.1177/0146621602026003003

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176. https://doi.org/10.1177/014662169201600206

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

Şahin, A., & Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Kuram ve Uygulamada Eğitim Bilimleri, 17*(1), 321–335. https://doi.org/10.12738/estp.2017.1.0270

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometrika Monograph Supplement No. 17). Psychometric Society.

Shen, L. (1997, March). *Quantifying item dependency by Fisher's Z*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.