The Effect of Sample Size on the Accuracy of Item Parameter and Ability Estimation in Polytomous Models Using Item Response Theory: A Simulation Study

Researcher Sameer Fouad Ailabouni
Professor
Irbid National University - Faculty of Educational Sciences

Researcher Saeed Yousef Alseid
UNRWA- Education Program

Sameerailabouni61@yahoo.com

Received: 25/2/2025 **Accepted: 16/4/2025**

Abstract

This study aimed to investigate the effect of sample size on the accuracy of estimating item parameters (difficulty, discrimination) and the ability parameter in polytomous models using Item Response Theory (IRT). A descriptive-analytical approach was adopted, and three different sample sizes (100, 500, and 1000 examinees) were generated following a normal distribution, with 20 items representing the test.

The difficulty parameters (β) were set to range between (-3 and 3), the discrimination parameters (α) between (0.5 and 3), and the ability parameters (θ) for individuals between (-3 and 3). The study utilized three polytomous IRT models: the Graded Response Model (GRM), the Partial Credit Model (PCM), and the Nominal Response Model (NRM), which were selected due to their differing assumptions and estimation methods.

A computer-based simulation was conducted using the WinGen program to generate artificial data, allowing precise control over data characteristics.

The findings revealed that the Graded Response Model (GRM) was the most accurate in estimating item parameters (difficulty, discrimination) and ability compared to the other models. Additionally, increasing the sample size improved estimation accuracy, as the computed values became closer to the true values. The study found no significant differences between the GRM and NRM in estimating the ability parameter, indicating that a larger sample size enhances estimation accuracy across all models.

Keywords: Polytomous models, ability parameter, difficulty parameter, discrimination parameter, Item Response Theory.

أثر حجم العينة على دقة تقدير معالم الفقرات والقدرة في النماذج متعددة التدريج باستخدام نظرية الاستجابة للفقرة: دراسة محاكاة

الباحث سمير فؤاد عيلبوني
أستاذ دكتور
جامعة إربد الأهلية / كلية العلوم التربوية

الباحث سعيد يوسف السيد
وكالة الغوث الدولية / برنامج التعليم

Sameerailabouni61@yahoo.com

تاريخ التقديم :25/2/2025
تاريخ القبول :16/4/2025

الملخص

هدفت الدراسة إلى استقصاء تأثير حجم العينة على دقة تقدير معالم الفقرات (الصعوبة، التمييز) ومعلمة القدرة (القدرة) في النماذج متعددة التدريج، وذلك باستخدام نظرية الاستجابة للفقرة (IRT)، حيث تم اعتماد المنهج الوصفي التحليلي وتم توليد ثلاث عينات بحجوم مختلفة (100، 500، 1000 مفحوص) تتبع التوزيع الطبيعي، مع 20 فقرة تمثل فقرات المقياس.

تم تحديد معلمات الصعوبة (β) لتتراوح بين (3- و 3)، ومعلمات التمييز (α) بين (0.5 و 3)، بينما تم تحديد معلمات القدرة (θ) للأفراد لتتراوح بين (3- و 3). استخدمت الدراسة ثلاثة نماذج متعددة التدريج: نموذج الاستجابة المتدرجة (GRM)، نموذج التقدير الجزئي (PCM)، ونموذج الاستجابة الاسمية (NRM)، والتي تم اختيارها نظرًا لاختلاف فرضياتها وطرق تقدير معالمها.

تم تطبيق المحاكاة الحاسوبية باستخدام برنامج (WinGen) لتوليد البيانات الافتراضية، مما أتاح التحكم في خصائص البيانات بشكل دقيق.

أظهرت نتائج الدراسة أن نموذج الاستجابة المتدرجة (GRM) كان الأكثر دقة في تقدير معالم الفقرات (الصعوبة، التمييز) والقدرة مقارنة بالنماذج الأخرى. كما تبين أن زيادة حجم العينة تحسن دقة تقدير المعالم، حيث تصبح القيم المحسوبة أكثر قربًا للقيم الحقيقية. لم تظهر الدراسة فروق كبيرة بين نموذجي (GRM) و (NRM) في تقدير معلمة القدرة، مما يشير إلى أن الحجم الأكبر للعينة يعزز دقة التقدير في جميع النماذج.

الكلمات المفتاحية: النماذج متعددة التدريج، معلمة القدرة، معلمة الصعوبة، معلمة التمييز، نظرية الاستجابة للفقرة

المقدمة

تُعد أدوات القياس والتقويم من أهم الركائز التي تستند إليها البحوث التربوية والنفسية، لما لها من دور جوهري في جمع البيانات وتحليلها واستخلاص النتائج. وتزداد أهمية هذه الأدوات بمدى دقتها وصدقها وثباتها في تقدير السمات أو القدرات المستهدفة. ومع تطور مجالات البحث والإحصاء، أصبح من الضروري تطوير نماذج القياس بما يتلاءم مع طبيعة البيانات والتغيرات في بيئات التطبيق. ومن هذا المنطلق، ظهرت الحاجة إلى تجاوز محدودية النماذج التقليدية، والتوجه نحو نماذج أكثر مرونة ودقة في التقدير، وهو ما دفع الباحثون إلى الاهتمام بنظرية الاستجابة للفقرة كنموذج بديل ومتطور في هذا المجال.

واجهت نظرية القياس الكلاسيكية ((Classical Test Theory (CTT) العديد من العيوب ونقاط الضعف ولعل من أبرز هذه العيوب اعتمادها على خصائص الأفراد الذين يتم تطبيق الاختبار عليهم في تقدير معالم الاختبار من صعوبة وتمييز، كما أن قدرة المفحوص رهن بالاختبار الذي يتم تطبيقه عليه، مما دفع العلماء إلى بذل الجهود سعيا للتغلب على أوجه القصور في هذه النظرية، وقد أسفرت هذه الجهود عن ظهور نظرية الاستجابة للفقرة والتي تعتبر نموذجا رياضيا يهتم بدراسة العلاقة بين أداء الفرد الملاحظ على الاختبار والسمات الكامنة التي يمتلكها، إذ يتم تمثيل كل سؤال اختباري من خلال منحى خصائص الفقرة (ICC) والذي يوضح العلاقة بين احتمالية الإجابة الصحيحة وقدرة المفحوص في الاختبار(Mahmud, 2017).

وتتميز نظرية الاستجابة للفقرة بعدة خصائص من أبرزها أن تقدير قدرة المفحوص الكامنة مستقلة عن فقرات الاختبار (Item Free)، وتقدير معالم الفقرات مستقلة عن قدرة المفحوصين (Person Free) فمؤشرات الاختبار الإحصائية ثابتة ولا تختلف باختلاف عينة الأفراد، كما أنها تتيح المقارنة بين نتائج المفحوصين حتى إن اختلفت الفقرات الاختبارية التي أجابوا عنها (Zanon, et al., 2016).

كما أن نظرية الاستجابة للفقرة لها أربعة افتراضات رئيسة وفقا لهاملتون وسوامنثان (Hambleton & Swaminathan, 1985) ينبغي تحققها وتتمثل في أحادية البعد (Unidimensionality)، الاستقلال الموضعي (Local Independence)، منحنى خصائص الفقرة (Item Characteristic Curve)، التحرر من السرعة (Speediness).

وتقسم نماذج القياس في نظرية الاستجابة للفقرة إلى قسمين رئيسين: النماذج ثنائية التدريج (Dichotomous Model)، والنماذج متعددة التدريج (Ploytomous Model)، إذ تستخدم النماذج ثنائية التدريج عندما يتم اختبار قدرة المفحوص باستخدام فئتين (صواب، خطأ)، مثل أسئلة الصح والخطأ، والاختيار من متعدد، ويندرج تحت هذا القسم النماذج اللوجستية أحادية وثنائية وثلاثية المعلمة، والتي تختلف في عدد المعالم التي تقوم

بحسابها لتقدير خصائص الفقرات وهي معلمة الصعوبة (b) ومعلمة التمييز (c) ومعلمة التخمين (c)، أما النماذج متعددة التدريج فهي نماذج يتم من خلالها وصف احتمالية إجابة المفحوص على المقاييس والاختبارات المتعددة وتتميز بقدرتها على زيادة المعلومات الاحصائية مقارنة بالنماذج الثنائية ومن الأمثلة على هذه النماذج: نموذج الاستجابة المتدرجة (Graded Response Model (GRM)) ونموذج التقدير الجزئي (Partial Credit Model (PCM)) ونموذج الاستجابة الاسمية (Nominal Response Model (NRM) (Brzezińska, 2016)).

إن تحديد النموذج المناسب للتحليل يعد من أكثر النقاط أهمية والتي ينبغي على الباحث إدراكها وفهمها، فقد يؤثر النموذج المختار في نتائج الدراسة، وقد أشار علام (2005) إلى بعض محكات اختيار النموذج ومنها: واقعية افتراضات النموذج، منعة النموذج (Robust)، ونوع البيانات المراد تحليلها.

ويُعَتَبر حجم العينة أحد العوامل الرئيسية التي تؤثر في دقة تقدير معالم الفقرات والقدرات، حيث أن زيادة حجم العينة غالبًا ما تؤدي إلى تحسين دقة التقديرات وتقليل الأخطاء المعيارية المرتبطة بها (Baker, 2001).

ويتم المفاضلة بين نماذج نظرية الاستجابة للفقرة في ضوء دقتها في تقدير معالم الفقرات ومعلمة القدرة للأفراد، فالنموذج الأفضل هو النموذج القادر على تقدير قيم المعالم وتقديم أعلى قيمة لدالة المعلومات بأقل خطأ معياري، ووفقا لهاملتون وسوامنثان (Hambleton & Swaminathan, 1985) فإن طول الاختبار يؤثر في دقة هذه التقديرات.

وقد حظيت الفقرات ذات الاستجابة الثنائية باهتمام الباحثين بشكل كبير، وتنوعت الدراسات التي تناولت نماذجها مقارنة في النماذج متعددة التدريج، فبالرغم من الأهمية الكبيرة للفقرات المتعددة التدريج في بناء الاختبارات والاستبانات إلا أن نصيبها كان محدودا في الأبحاث العلمية، كما أن الدراسات التي تناولت تأثير حجم العينة على دقة تقدير معالم الفقرات والقدرات باستخدام النماذج متعددة التدريج ما تزال محدودة، ومن هنا تبرز أهمية البحث الحالي الذي يهدف إلى التعرف على أثر حجم العينة في دقة تقدير معالم النماذج المتعددة التدريج باختلاف النماذج المستخدمة في التقدير.

الدراسات السابقة

تمثل الدراسات السابقة ركيزة أساسية في البناء العلمي إذ تتيح للباحث تأسيس قاعدة معرفية متينة لفهم الموضوع الذي يتناوله، وتساعده في معرفة التطورات النظرية والتطبيقة في مجال البحث، ومن خلال استعراض هذه الدراسات يتمكن الباحث من معرفة ما توصل إليه الباحثون مما يتيح له اكتشاف الثغرات البحثية التي يمكن أن يسدها من خلال دراسته، وعند استعراض الدراسات السابقة المتعلقة بتأثير حجم العينة تبين ندرة الدراسات التي تناولت أثر حجم العينة في دقة تقدير معالم النماذج متعددة التدريج، وفيما يلي عرض بعض هذه الدراسات:

قام سوامنثان وآخرون (Swaminathan et al, 2003) بدراسة هدفت تحديد أثر الحكم المسبق على دقة تقدير معالم الفقرة (الصعوبة) في الاختبارات ثنائية التدريج، إذ تم استخدام الطرق البيزية في تقدير معلمة الصعوبة، وتوصلت الدراسة إلى أن تقدير الصعوبة يتحسن بزيادة حجم العينة.

أما ستون ويوموتو (Stone & Youmoto, 2004) فقد أجريا دراسة هدفت إلى التعرف على أثر حجم العينة على تقدير معالم الفقرات ثنائية التدريج باستخدام نموذج راش ونماذج نظرية الاستجابة للفقرة، وقاما بسحب (30) عينة من نتائج اختبار (Konx,s Cube Test Revised) ا لتقدير معالم الفقرات، وتوصل إلى أن نموذج راش يعطي أقل تقدير لمعلمة الصعوبة، وأن جودة المطابقة تتناسب طرديا مع حجم العينة.

وقام الثوابية (2010) بدراسة أثر حجم العينة على تقدير صعوبة الفقرة والخطأ المعياري في تقديرها، ولتحقيق أهداف الدراسة قام الباحث بتطبيق اختبار رياضيات للصف العاشر مكون من (80) فقرة من نوع الاختيار من متعدد على عينات عشوائية تتراوح أحجامها ما بين (200) إلى (11292) طالب وطالبة، وتوصلت الدراسة إلى أن قيمة معلمة الصعوبة تزداد بازدياد حجم العينة، أما الخطأ المعياري في تقدير معلمة الصعوبة فإنه يتناقص بزيادة عدد أفراد العينة.

وأجرى كوستر (Custer, 2015) دراسة بعنوان «حجم العينة ودقة تقدير معلمات الفقرات عند استخدام نموذج الفقرة الواحدة (راش)»، والتي تم تقديمها في الاجتماع السنوي للجمعية البحثية التعليمية في منطقة الغرب الأوسط في إيفانستون، إلينوي. هدفت الدراسة إلى استكشاف العلاقة بين حجم العينة ودقة تقدير معلمات الفقرات عند استخدام نموذج الفقرة الواحدة (راش). إذ قام الباحث باستخدام ثلاث حالات: الأولى استخدم فيها اختبارًا مكونًا من 40 فقرة دون بيانات مفقودة، والثانية تضمنت بيانات مفقودة، بينما الثالثة كانت تعتمد على مقياس رأسي عبر ثلاثة مستويات. قام الباحث بتقييم دقة تقدير معلمات الفقرات عن طريق مقارنة القيم المقدرة بالقيم «الحقيقية»، وقياس الانحراف التربيعي المتوسط الجذري (RMSD) وعدد القيم الشاذة. وتوصلت الدراسة إلى أن دقة تقدير معلمات الفقرات تحسنت بشكل ملحوظ عند وصول حجم العينة إلى 400 مختبر، حيث كانت الزيادات في حجم العينة بعد 500 مختبر ضئيلة في تحسين دقة التقدير. كما أظهرت الدراسة أن التقديرات كانت أكثر دقة عندما كانت العينات موزعة على مستويات متعددة.

أما القيسي (2016) فقد أجرى دراسة هدفت إلى الكشف عن أثر حجم العينة وطول الاختبار على دقة تقدير معالم الفقرة والقدرة باستخدام نماذج نظرية الاستجابة للفقرة اللامعلمية، معتمدا على مؤشر دقة القياس (التحييز)، ومتوسط الجذر التربيعي لمربعات الخطأ، ولتحقيق أهداف الدراسة قام بتوليد ثلاث استجابات لثلاث اختبارات تختلف في عدد فقراتها (20، 40، 60) فقرة اختبارية، بحيث تتبع الاستجابة التوزيع الطبيعي بمتوسط حسابي

(0) وانحراف معياري (1)، وقد أظهرت نتائج الدراسة وجود فروق دالة إحصائيا في التحييز تعزى لمتغير العينة ، كما أظهرت النتائج وجود فروق دالة إحصائيا في دقة تقدير معلمة التمييز ومعلمة التخمين تعزى لمتغيري الدراسة (حجم العينة، طول الاختبار)، وعدم وجود فروق في معلمة الصعوبة ومعلمة القدرة.

وقام الشريفين (2019) بدراسة هدفت إلى معرفة أثر نموذج نظرية الاستجابة متعددة التدريج على دقة تقدير القدرات للأفراد والخصائص السيكومترية للفقرات والاختبار، ولتحقيق أهداف الدراسة قام الباحث بتوليد بيانات خاصة باختبار مكون من (50) فقرة وفق نموذج التقدير الجزئي العام، واستخدم ثلاث نماذج في تقدير معالم الصعوبة والتمييز والقدرة وهي (GRM, NRM, GPCM) وتوصل إلى أن نموذج (GRM) هو أكثر النماذج دقة في تقدير معالم الفقرات والأفراد وأنه قدم أقصى كمية للمعلومات مقارنة بباقي النماذج.

أما سوارتو وسيتياوان (Suwarto & Setiawan, 2019) فقد أجرى دراسة هدفت فحص تأثير حجم العينة ونماذج اللوجستية ((3 (PL، (PL2)، 1PL) على دقة تقدير معلمات الفقرات مثل الصعوبة، التمييز، والتخمين. تم إجراء محاكاة حاسوبية باستخدام عينات بحجم (200، 500، و1000) مع عشرة تكرارات لكل حجم عينة. أظهرت النتائج أن زيادة حجم العينة يحسن دقة تقدير المعلمات، حيث كان حجم العينة (1000) هو الأفضل. فيما يتعلق بالنماذج اللوجستية تبين أن نموذج (PL1) كان الأكثر دقة في تقدير معلمة الصعوبة، ونموذج (PL2) كان الأفضل في تقدير معلمة التمييز، في حين كان نموذج (PL3) الأفضل في تقدير معلمة التخمين. بشكل عام، أكدت الدراسة على أهمية اختيار حجم عينة مناسب ونموذج لوجستي متوافق مع نوع المعلمات لتقديم تقديرات دقيقة.

وأجرى القضاة (2020) دراسة هدفت لمعرفة أثر حجم العينة على تقدير دالة معلومات الاختبار وقدرة الفرد باستخدام النموذج الثلاثي، ولتحقيق أهداف الدراسة قام بتطبيق اختبار رياضيات للصف التاسع مكون من (60) فقرة من نوع الاختيار من متعدد على عينة مكونة من (1230) طالب، اذ استخدم ثلاث عينات (300، 400، 530) طالب، وتوصل إلي أن دقة تقدير دالة معلومات الاختبار تزداد بزيادة حجم العينة، وأن الفروق في متوسطات الخطأ المعياري لتقدير القدرة تتناسب عكسيا مع حجم العينة فكلما زاد حجم العينة قل متوسط الخطأ المعياري في تقدير القدرة .

التعقيب على الدراسات السابقة

تُبرز الدراسات السابقة أهمية تأثير حجم العينة على دقة تقدير معلمات الفقرات في نظرية الاستجابة للفقرة، حيث تبين أن زيادة حجم العينة تحسن الدقة وتقلل الأخطاء، لكن بعد حد معين لا يكون لزيادتها تأثير كبير. وقد أشارت الدراسات إلى وجود حد أدنى مطلوب للعينة يعتمد على عدد المعلمات وطول الاختبار والنموذج المستخدم.

من أبرز هذه الدراسات، دراسة القضاة (2020) التي أكدت دور حجم العينة في تحسين دقة دالة معلومات الاختبار، ودراسة القيسي (2016) التي تناولت تأثير حجم العينة وطول الاختبار على تقدير المعلمات، بالإضافة إلى دراسة كوستر (2015) التي وجدت أن الدقة تتحسن عند عينة بين 400 و500 مفحوص، ودراسة سوارتو (2019) التي شددت على ضرورة اختيار حجم عينة يتناسب مع النموذج المستخدم.

في هذا السياق، تبرز أهمية البحث الحالي باعتباره إضافة علمية تركز على تأثير حجم العينة في النماذج متعددة التدريج، وهو جانب لم يُبحث بعمق في الدراسات السابقة. تسعى هذه الدراسة إلى سد الفجوة البحثية من خلال فحص دقة تقديرات المعلمات في هذه النماذج، مما يساهم في تحسين فهم تأثير حجم العينة على التقديرات الإحصائية وتعزيز دقة تطبيقات نظرية الاستجابة للفقرة.

مشكلة الدراسة وأسئلتها:

تعدّ نماذج نظرية الاستجابة للفقرة من الأدوات الإحصائية المتقدمة التي يُعتمد عليها في تحليل البيانات التربوية والنفسية، لما لها من قدرة على تقدير معالم الفقرات والأفراد بدقة. ومع تنوع النماذج متعددة التدريج فإن الباحثين يواجهون تحديات في اختيار النموذج الأنسب، خاصة في ظل وجود تباين في نتائج الدراسات السابقة فيما يتعلق بتأثير حجم العينة والنموذج المستخدم على دقة التقدير. فقد أظهرت بعض الدراسات (Stone & Youmoto, 2004؛ القيسي، 2016؛ Suwarto & Setiawan, 2019) أن دقة تقدير المعالم تتحسن بزيادة حجم العينة، في حين ركزت دراسات أخرى (: Custer, 2015؛ الشريفين، 2019) على تفوق بعض النماذج على غيرها في تقديم تقديرات أكثر دقة.

ومع ذلك، فإن مراجعة الأدبيات تشير إلى ندرة الدراسات التي تناولت أثر التفاعل بين حجم العينة والنموذج المستخدم على دقة التقدير بشكل مباشر، إذ غالبًا ما تم تناول هذين المتغيرين بصورة منفصلة. ومن هنا تنبع أهمية هذه الدراسة التي تسعى إلى سد هذه الفجوة من خلال استخدام أسلوب المحاكاة الحاسوبية لفحص مدى تأثير تفاعل حجم العينة (100، 500، 1000) والنموذج المستخدم (GRM، PCM، NRM) على دقة تقدير معالم الفقرات (الصعوبة، التمييز) وقدرة الأفراد، وتحديدا فقد سعت هذه الدراسة الإجابة عن التساؤلات الآتية:

هل تختلف دقة تقدير معلمة الصعوبة بين النماذج متعددة التدريج (GRM, PCM, NRM) باختلاف حجم العينة ؟

هل تختلف دقة تقدير معلمة التمييز بين النماذج متعددة التدريج (GRM, PCM, NRM) باختلاف حجم العينة ؟

هل تختلف دقة تقدير معلمة القدرة للأفراد بين النماذج متعددة التدريج (GRM, PCM, NRM) باختلاف حجم العينة ؟

أهمية الدراسة: تتمثل أهمية هذه الدراسة في بُعدين رئيسيين:

أولا: الأهمية النظرية: التي تكمن في تطوير المعرفة في مجال القياس النفسي والتربوي، حيث تسهم الدراسة في تعميق الفهم حول تأثير حجم العينة على دقة تقدير معالم الفقرات والأفراد في نماذج الاستجابة للمفردة متعددة التدريج (GRM, PCM, NRM)، بالإضافة إلى تحديد الفروق بين هذه النماذج المختلفة عند اختلاف حجم العينة.

ثانيا: الأهمية العملية: والتي تتمثل في توجيه الباحثين في تصميم الدراسات من خلال تقديم إرشادات حول اختيار النموذج الأنسب بناءً على حجم العينة المتاح، مما يؤدي إلى تحسين دقة التقديرات والنتائج المستخلصة. كما تساهم في تحسين جودة الاختبارات والمقاييس، حيث يتمكن الممارسون من تصميم اختبارات ومقاييس أكثر دقة وموثوقية، مما يعزز جودة التقييمات النفسية والتربوية.

محددات الدراسة:

يتحدد تعميم نتائج هذه الدراسة في ضوء المحددات الآتية:

النماذج المتعددة التدريج: وهي النماذج متعددة التدريج التي تم استخدامها في هذه الدراسة (GRM, PCM, NRM).

احجام العينات: حيث تضمنت هذه الدراسة ثلاث عينات مختلفة في أحجامها (100، 500، 1000) .

البيانات: حيث تم توليد البيانات حاسوبيا باسخدام برنامج (WinGen) للعينات الثلاثة على (20) فقرة خماسية التدريج، بحيث تتبع الاستجابات التوزيع الطبيعي بمتوسط حسابي (0) وانحراف معياري (1).

التعريفات الاصطلاحية والإجرائية:

تضمنت هذه الدراسة مجموعة من المصطلحات التي سيتم تعريفها مفاهيميا وإجرائيا وفق الآتي:

النماذج متعددة التدريج: هي نماذج يتم من خلالها وصف احتمالية إجابة المفحوص على المقاييس والاختبارات متعددة التدريج مثل الاستبانات (Brzezińska, 2016).

وفي هذه الدراسة تم استخدام ثلاثة نماذج متعددة التدريج وهي: (نموذج الاستجابة المتدرجة (Graded Response Model، نموذج القدير الجزئي Partial Credit Model، النموذج الاسمي Nominal Response Model).

معالم الأفراد (القدرة): مقدار ما يمتلكه الفرد من السمة التي يتم قياسها، وهي متغير يعبر عن السمة الأساسية للفرد، وتحدد مستوى قدرته على طول مقياس ما بناء على استجابته لمجموعة من الفقرات. (Nguyen, et al., 2014).

وقد تم تعريف القدرة إجرائيا في هذا البحث على أنها مقدار ما يمتلك الفرد من اتجاهات نحو التعلم النشط تبعا للنماذج المتعددة التدريج (GRM, PCM, NRM).

معلمة الصعوبة (العتبات الفارقة / الموقع): هو مستوى السمة اللازمة لكي تتجاوز الاستجابة العتبة الفارقة باحتمال قدره (50%) (الشريفين، 2017).

ومن ناحية إجرائية فهي قيم معالم الصعوبة لفقرات مقياس اتجاهات المعلمين نحو التعلم النشط والتي يتم تقديرها تبعا للنماذج المتعددة التدريج (GRM, PCM, NRM).

معلمة التمييز: قدرة الفقرة على التمييز بين المفحوصين على مستويات السمات المختلفة (Livingston,2020)، وتعرف رياضيا على أنها نقطة الميل الواقعة على منحنى خصائص الفقرة عند نقطة الانعطاف.

وتعرف معلمة التمييز إجرائيا على أنها قيم معلمة التمييز لفقرات مقياس اتجاهات المعلمين نحو التعلم النشط والتي يتم تقديرها تبعا للنماذج المتعددة التدريج (GRM, PCM, NRM).

دقة التقدير: هو مقدار تقارب القيم المقدرة من القيم الحقيقية، وفي هذه الدراسة تم استخدام مؤشر التحيز للمعلمة (Bias)، ومؤشر الجذر التربيعي لمعدل لمربعات للخطأ (Root Mean Square Error (RMSE)).

الطريقة وإجراءات الدراسة

منهج الدراسة: اعتمد هذه الدراسة على المنهج الوصفي التحليلي، حيث يتم استخدام المحاكاة الحاسوبية لتوليد بيانات افتراضية بثلاثة أحجام عينات (100، 500، 1000) باستخدام برنامج (WinGen) لدراسة تأثير حجم العينة على دقة تقدير معالم الفقرات والأفراد عند تطبيق نماذج (GRM،PCM ، NRM) وتم استخدام برنامج (SPSS 25) وبرنامج (StataMP 17) لاجراء التحاليل الإحصائية.

عينة الدراسة: تمثل عينة الدراسة استجابات مولدة حاسوبيًا لمفحوصين على (20) فقرة، حيث صُممت الفقرات بمعامل تمييز يتراوح بين (0.5-2.5) ومعاملات صعوبة موزعة بين (3- و3) عبر أربع عتبات باستخدام مقياس ليكرت خماسي. تم توليد القدرات الحقيقية للمفحوصين وفق توزيع طبيعي معياري N(0,1) لضمان تنوع المستويات، مع ضبط البذرة العشوائية لضمان إمكانية تكرار النتائج.

اعتمدت عملية توليد البيانات على دالة التوزيع اللوجستي التراكمي، حيث حُسبت احتمالات اختيار كل فئة استجابة، ثم سُحبت الاستجابات وفقًا لهذه الاحتمالات. ومن خلال مقارنة تقديرات المعلمات عبر العينات المختلفة، يتم تحليل مدى تأثير حجم العينة على دقة التقدير في النماذج المستخدمة.

أولا: معالم الفقرات

تم تحديد معالم الفقرات وفقًا لنموذج الاستجابة للفقرة، حيث كانت قيم المعالم كما يوضح الجدول (1) :

جدول 1

قيم معلمة التمييز والعتبات الفارقة لمعلمة الصعوبة الحقيقية للبيانات المولدة

	العينة الأولى (100 مفحوص)					العينة الثانية (500 مفحوص)					العينة الثالثة (1000) مفحوص				
الفقرة	معامل التمييز (a)	العتبة الأولى (b1)	العتبة الثانية (b2)	العتبة الثالثة (b3)	العتبة الرابعة (b4)	معامل التمييز (a)	العتبة الأولى (b1)	العتبة الثانية (b2)	العتبة الثالثة (b3)	العتبة الرابعة (b4)	معامل التمييز (a)	العتبة الأولى (b1)	العتبة الثانية (b2)	العتبة الثالثة (b3)	العتبة الرابعة (b4)
1	1.249	-2.163	-1.247	-0.802	0.671	1.364	-1.95	-1.11	-0.675	0.821	1.452	-1.842	-0.998	-0.599	0.92
2	2.401	-1.802	-0.264	0.085	1.711	2.487	-1.699	-0.198	0.162	1.825	2.572	-1.602	-0.143	0.245	1.902
3	1.964	-2.721	-1.977	0.554	0.645	2.025	-2.56	-1.85	0.689	0.782	2.101	-2.462	-1.732	0.778	0.899
4	1.697	-2.61	1.85	2.693	2.794	1.758	-2.488	1.942	2.8	2.876	1.842	-2.375	2.025	2.892	2.95
5	0.812	-2.414	-1.172	-0.359	1.105	0.93	-2.276	-1.025	-0.21	1.238	1.012	-2.185	-0.94	-0.127	1.356
6	0.812	-2.794	-2.268	-0.029	2.456	0.92	-2.672	-2.12	0.145	2.589	1.015	-2.56	-1.98	0.275	2.715
7	0.616	-1.447	-1.13	0.12	0.975	0.705	-1.3	-1.002	0.259	1.102	0.784	-1.185	-0.91	0.375	1.202
8	2.232	-1.891	0.28	1.651	2.818	2.301	-1.75	0.41	1.782	2.93	2.385	-1.635	0.53	1.89	3.025
9	1.702	0.587	2.369	2.531	2.637	1.78	0.69	2.48	2.612	2.74	1.852	0.798	2.56	2.705	2.825
10	1.916	-2.729	-2.469	-1.824	-1.048	2.005	-2.59	-2.345	-1.699	-0.92	2.092	-2.482	-2.245	-1.605	-0.81
11	0.541	-1.372	-0.859	-0.668	1.972	0.645	-1.249	-0.73	-0.53	2.1	0.725	-1.15	-0.645	-0.445	2.215
12	2.44	-2.154	-1.314	0.256	1.813	2.523	-2.032	-1.2	0.369	1.945	2.605	-1.925	-1.105	0.465	2.05
13	2.165	-2.553	-1.808	1.633	2.921	2.24	-2.43	-1.7	1.755	3.01	2.312	-2.32	-1.61	1.842	3.098
14	0.925	-2.967	1.241	1.374	1.893	1.01	-2.8	1.39	1.48	2.015	1.089	-2.69	1.5	1.582	2.102
15	0.864	-2.556	-2.305	-0.849	1.628	0.952	-2.41	-2.159	-0.72	1.785	1.032	-2.3	-2.035	-0.605	1.91
16	0.867	-2.619	-1.015	0.74	2.179	0.954	-2.485	-0.88	0.89	2.29	1.035	-2.375	-0.765	1.005	2.398

17	1.108	-1.134	-1.049	0.825	1.378	1.195	-1.01	-0.92	0.96	1.502	1.278	-0.92	-0.805	1.075	1.62
18	1.55	-2.282	-0.167	1.279	2.323	1.632	-2.15	-0.04	1.415	2.45	1.71	-2.03	0.075	1.545	2.56
19	1.364	-0.037	0.368	1.565	1.626	1.445	0.125	0.489	1.705	1.752	1.522	0.238	0.59	1.832	1.868
20	1.082	-2.847	-2.353	-0.435	0.136	1.169	-2.705	-2.212	-0.289	0.285	1.25	-2.58	-2.1	-0.18	0.398

ومن ثم تم حساب الأوساط الحسابية لقيم المعالم كما يوضح الجدول (2)

جدول 2

قيم الأوساط الحسابية لمعلمة التمييز ومعالم عتبات الصعوبة

العينة	متوسط (a)	المتوسط العام (b)	متوسط b1	متوسط b2	متوسط b3	متوسط b4
100 مفحوص	1.415	-0.16	-2.025	-0.764	0.517	1.632
500 مفحوص	1.502	-0.03125	-1.887	-0.639	0.645	1.756
1000 مفحوص	1.583	0.07275	-1.779	-0.537	0.747	1.86

يُلاحظ أن متوسط معامل التمييز يزداد مع زيادة حجم العينة، حيث كان في العينة الصغيرة (100 مفحوص) 1.415، وارتفع إلى 1.502 في العينة المتوسطة (500 مفحوص)، ثم وصل إلى 1.583 في العينة الكبيرة (1000 مفحوص)،أما قيم معاملات الصعوبة (b4, b3, b2, b1) فقد تراوحت ما بين (-2.025 إلى 1.632) للعينة الأولى وما بين (1.887 - إلى 1.756) للعينة الثانية و (-1.779 إلى 1.86) للعينة الثالثة، مما يدل على تدرج صعوبة الفقرة عبر المستويات المختلفة.

ثانيا: معلمة القدرة باختلاف حجم العينة

تم حساب متوسط القدرة للعينة وفقًا لعدد المفحوصين، وكانت القيم كما يوضح الجدول (3)

جدول 3

قيم الأوساط الحسابية لمعلمة القدرة باختلاف أحجام العينات

حجم العينة	متوسط معلمة القدرة
100	-0.291
500	-0.018
1000	0.011

يلاحظ أن متوسط القدرة بلغ (-0.291) لعينة (100 مفحوص)، و(-0.018) لعينة (500) مفحوص و(0.011) كما ويلاحظ اقتراب القيم من الصفر بازدياد حجم العينة، لكنه أصبح موجبًا عندما زاد حجم العينة إلى 1000، مما قد يشير إلى استقرار التقدير مع زيادة حجم العينة.

أداة الدراسة: تتكون أداة الدراسة من مقياس يتضمن (20) فقرة خماسية التدريج، حيث تم توليد استجابات المفحوصين الافتراضية لهذه الفقرات باستخدام تقنيات المحاكاة الحاسوبية، وتمت مراعاة توزيع القدرات ومعاملات الفقرات مثل التمييز و الصعوبة. وتوفر هذه الطريقة بيئة محكمة للسيطرة على خصائص البيانات وتسمح بدراسة تأثير المتغيرات المختلفة مثل حجم العينة على دقة تقدير المعلمات في النماذج المستخدمة.

إجراءات الدراسة:

المرحلة الأولى: تم توليد بيانات لمقياس مكون من 20 فقرة وفقًا لنموذج الاستجابة المتدرجة(GRM) باستخدام برنامج WinGen وتم تحديد معاملات التمييز (a) بين (0.5 إلى 2.5)، ومعاملات الصعوبة (b) تم توزيعها بين (-3 إلى 3) عبر أربع عتبات.

المرحلة الثانية: تم توليد استجابات لثلاث عينات بمختلف الأحجام (100، 500، 1000) وفق التوزيع الطبيعي المعياري N(0,1)، حيث تم تحديد الوسط الحسابي لهذا التوزيع، كما يوضح الجدول (4)

جدول 4

قيم الأوساط الحسابية والانحرافات المعيارية لتوزيعات عينة الدراسة باختلاف حجم العينة

حجم العينة	الوسط الحسابي للتوزيع	الانحراف المعياري للتوزيع	الالتواء	التفلطح
100	2.916	0.5447	-0.048	-0.601
500	2.880	0.5312	0.065	0.269
1000	2.915	0.5300	-0.004	-0.093

ثالثا: قبل اجراء التحليل تم التأكد من تحقق افتراضات نظرية الاستجابة للفقرة كما يلي:

أولا: افتراض أحادية البعد: للتحقق من شرط أحادية البعد في البيانات التي تم توليدها للعينات الثلاثة، تم اجراء التحليل العاملي لايجاد قيمة الجذر الكامن ونسبة التباين المفسر كما يشير الجدول (5).

جدول 5

الجذر الكامن ونسبة التباين المفسر لاستجابات عينات الدراسة باختلاف حجم العينة

العينة	رقم العامل	قيمة الجذر الكامن	نسبة التباين المفسر	نسبة الجذر الأول إلى الجذر الثاني
العينة الأولى (100) مفحوص	1	5.8	40.321	6.44
	2	0.9	2.452	
العينة الثانية (500) مفحوص	1	6.1	47.519	7.18
	2	0.85	1.972	
العينة الثالثة (1000) مفحوص	1	6.3	52.617	7.41
	2	0.85	1.593	

يظهر الجدول (5) أن نسبة التباين المفسر للعامل الأول في العينات الثلاثة كل على حدة تجاوزت (20%) وهذا يؤكد تحقق شرط أحادية البعد وفق ما أشار إليه ريكاس (Reckase,1979)، وهذا يؤكد تحقق افتراض أحادية البعد للعينات الثلاثة.

ثانيا: افتراض الاستقلال الموضعي: لفحص افتراض الاستقلال الموضعي تم حساب مؤشر(ZQ3) الذي يقوم بحساب الارتباط الموضعي بين قيم ZQ3 الفشرية للفقرات المستقلة، فإذا كانت قيمته تزيد بمقدار انحرافين معياريين يمكن اعتبار الفقرات غير مستقلة موضعيا، أما إذا كان المتوسط الحسابي لمؤشرات ZQ3 الملاحظ واقعا بين الحد الأدنى والحد الأعلى لمؤشر ZQ3 فيعتبر الاستقلا الموضعي متحققا.(Shen, 1997)، والجدول (6) يوضح مؤشرات الاستقلال الموضعية

جدول 6

مؤشرات الاستقلال الموضعي

الاحصائي	العينة الأولى(100)	العينة الثانية (500)	العينة الثالثة (1000)
عدد الفقرات	20	20	20
عدد أزواج الفقرات	190	190	190
القيمة الصغرى الملاحظة لمؤشر ZQ3	0.4721	0.4856	0.4912
القيمة العظمى للملاحظة لمؤشر ZQ3	0.6015	0.6123	0.6291
المتوسط الحسابي للفقرات لمؤشر ZQ3	0.5892	0.6014	0.6173

يلاحظ من الجدول (6)أن المتوسط الحسابي لمؤشر ZQ3 لعينة (100، 500، 1000) بلغ (0.5892، 0.6014، 0.6173) على الترتيب، وهو يقع ضمن الحدين الأدنى والأعلى، مما يعزز تحقق افتراض الاستقلال الموضعي.

كما تم حساب عدد أزواج الفقرات التي وقعت ضمن فترة الثقة المحققة لشروط الاستقلال الموضعي، والجدول (7) يوضح مؤشرات الاستقلال الموضعي تبعا لنظرية الاستجابة للفقرة.

جدول 7

مؤشرات الاستقلال الموضعي تبعًا لنظرية الاستجابة للفقرة

حالة الاستقلال الموضعي	العينة الأولى (100)		العينة الثانية (500)		العينة الثالثة (1000)	
	عدد الأزواج	النسبة المئوية	عدد الأزواج	النسبة المئوية	عدد الأزواج	النسبة المئوية
التبعية الموضعية	21	11.05%	17	8.95%	14	7.37%
الاستقلالية الموضعية	169	88.95%	173	91.05%	176	92.63%
الكلي	190	100%	190	100%	190	100%

يشير الجدول (7) إلى أن عدد أزواج الفقرات الواقعة ضمن الاستقلال الموضعي لعينة (100، 500، 1000) بلغت (169، 173، 176) بنسب مئوية (88.95%، 91.05%، 92.63%) على الترتيب، في حين أن عدد الفقرات التي كان لها تبعية كان منخفضا جدا ويقل عن (12.5%) مما يشير إلى تحقق شرط الاستقلال الموضعي (Kim, Cohen, & Lin, 2005).

نتائج الدراسة ومناقشتها

النتائج المتعلقة بالسؤال الأول والذي نص على» هل تختلف دقة تقدير معلمة الصعوبة بين النماذج متعددة التدريج (GRM, PCM, NRM) باختلاف حجم العينة ؟»

للإجابة عن هذا السؤال تم حساب الأوساط الحسابية والانحرافات المعيارية لتقديرات معالم الصعوبة (العتبات الفارقة) تبعا لأحجام العينات المختلفة ونموذج الاستجابة متعددة التدريج كما يوضح الجدول (8)

جدول 8

الأوساط الحسابية والانحرافات المعيارية لمعالم الصعوبة (العتبات الفارقة) تبعا لحجم العينة ونموذج الاستجابة

العينة	عتبات الصعوبة	النموذج						الكلي	
		GRM		PCM		NRM			
		الوسط الحسابي	الانحراف المعياري	الوسط الحسابي	الانحراف المعياري	الوسط الحسابي	الانحراف المعياري	الوسط الحسابي	الانحراف المعياري
العينة الأولى (100)	الأولى	-2.321	0.400	-1.600	0.300	-2.820	0.350	-2.247	0.350
	الثانية	-1.439	0.380	-1.200	0.280	-1.382	0.320	-1.340	0.327
	الثالثة	1.548	0.300	0.100	0.200	1.692	0.240	1.113	0.247
	الرابعة	1.650	0.250	1.000	0.180	1.703	0.220	1.451	0.217
	الكلي	-0.141	0.333	-0.425	0.240	-0.202	0.283	-0.256	0.285
العينة الثانية (500)	الأولى	-1.826	0.350	-2.316	0.280	-2.012	0.300	-2.051	0.310
	الثانية	-0.851	0.330	-1.547	0.250	-0.714	0.280	-1.037	0.287
	الثالثة	0.693	0.250	0.753	0.180	0.496	0.220	0.647	0.217
	الرابعة	1.820	0.280	1.302	0.220	1.921	0.240	1.681	0.247
	الكلي	-0.041	0.303	-0.452	0.233	-0.077	0.260	-0.190	0.265

العينة الثالثة (1000)	الأولى	-1.600	0.300	-2.321	0.250	-1.765	0.280	-1.895	0.277
	الثانية	-0.492	0.280	-0.721	0.220	-0.386	0.240	-0.533	0.247
	الثالثة	0.813	0.250	0.937	0.180	0.967	0.220	0.906	0.217
	الرابعة	1.624	0.230	1.923	0.200	1.543	0.220	1.697	0.217
	الكلي	0.086	0.265	-0.046	0.213	0.090	0.240	0.044	0.239

يلاحظ من الجدول (8) وجود فروق ظاهرية بين الأوساط الحسابية لمعالم عتبات الصعوبة الأربعة باختلاف نموذج الاستجابة وباختلاف حجم العينة، كما يلاحظ أنه كلما زادت حجم العينة اقتربت القيم المقدرة من القيم الحقيقية، كما يلاحظ أن الأوساط الحسابية لعتبات الصعوبة المقدرة وفق نموذج (GRM) كانت الأقرب لتقديرات العتبات الحقيقية التي تم توليدها والجدول (9) يوضح القيم الحقيقية والقيم المقدرة باختلاف النموذج وحجم العينة.

جدول 9

الأوساط الحسابية والانحرافات المعيارية للقيم الحقيقة والمقدرة لعتبات الصعوبة

العينة	القيم	متوسط b1	متوسط b2	متوسط b3	متوسط b4	المتوسط العام (b)
مفحوص100	حقيقية	-2.025	-0.764	0.517	1.632	-0.160
	GRM	-2.321	-1.439	1.548	1.650	-0.141
	PCM	-1.600	-1.200	0.100	1.000	-0.425
	NRM	-2.820	-1.382	1.692	1.703	-0.202
مفحوص500	حقيقية	-1.887	-0.639	0.645	1.756	-0.0313
	GRM	-1.826	-0.851	0.693	1.82	-0.0410
	PCM	-2.316	-1.547	0.753	1.302	-0.4520
	NRM	-2.012	-0.714	0.496	1.921	-0.0770
مفحوص1000	حقيقية	-1.779	-0.537	0.747	1.86	0.073
	GRM	-1.600	-0.492	0.813	1.624	0.086
	PCM	-2.321	-0.721	0.937	1.923	-0.046
	NRM	-1.765	-0.386	0.967	1.543	0.09

يلاحظ من الجدول (9) ان العينة المكونة من (100) مفحوص كانت القيم المقدرة للعتبات تظهر توزيعًا قريبًا من القيمة المرجعية (الحقيقية)، حيث يتوزع المتوسط العام للعتبات حول (–0.16) وتمثل عتبات GRM أفضل تطابق مع القيم الحقيقية حيث أن المتوسط العام لهذا النموذج قريب جدًا من المتوسط الفعلي عند (–0.141)، أما فيما يتعلق بنموذج (PCM) فإنه يظهر تباينًا أكبر في العتبات حيث كانت القيم متفاوتة مما نتج عنه متوسط عام أقل من القيمة الحقيقية (–0.425)، وفي نموذج (NRM)، فبلغ المتوسط العام(–0.202).

أما في العينة المكونة من (500) مفحوص فيلاحظ أن المتوسط العام للعتبات الحقيقية هو(−0.0313)، وقد بلغت قيمة متوسط العتبات الفارقة المقدرة وفق نموذج (GRM) (−0.0410) مما يعكس تقاربا مع البيانات الحقيقية. أما قيم متوسط العتبات المقدرة وفق نموذج (PCM) فقد بلغت (−0.4520)، وفي نموذج (NRM) فقد بلغت القيم المقدرة (−0.077).

وفي العينة المكونة من (1000) مفحوص فقد كان المتوسط العام للعتبات الحقيقية هو(0.073)، وقد بلغت قيمة متوسط العتبات المقدرة وفق نموذج (GRM) (0.086)، أما وفق نموذج (PCM) فقد بلغت قيمة المتوسط الحسابي لعتبات الصعوبة المقدرة (−0.046)، ووفقا لنموذج (NRM) فقد بلغت قيمة المتوسط الحسابسي للعتبات المقدرة (0.09).

بناءً على تحليل العتبات عبر النماذج المختلفة في العينات الثلاث (100، 500، 1000 مفحوص)، يتبين أن نموذج (GRM) هو الأكثر توافقًا مع القيم الحقيقية في معظم العينات، حيث يكون المتوسط العام قريبًا جدًا من القيم المرجعية المطلوبة. بينما يظهر نموذج (NRM) أداءً جيدًا أيضًا مع نتائج قريبة من القيم الحقيقية، أما نموذج (PCM)، فإنه يظهر تباينًا أكبر في العتبات مقارنةً بالنماذج الأخرى، ما يؤدي إلى انحرافات أكبر في المتوسط العام عن القيم المرجعية.

كم يتضح أن هناك تحسنًا في دقة تقدير العتبات مع زيادة حجم العينة، ففي عينة (100) مفحوص، كانت القيم أكثر تباينًا مقارنة بالعينة الأكبر. أما في عينة (500) مفحوص، فلاحظنا أن تقديرات العتبات اقتربت بشكل أكبر من القيم الحقيقية، مما يشير إلى تحسين دقة التقدير مع زيادة الحجم. وعند زيادة الحجم إلى (1000) مفحوص، أصبح المتوسط العام للعتبات أكثر استقرارًا، مما يعكس تحسنًا إضافيًا في دقة تقدير العتبات. بناءً على ذلك، يمكن الاستنتاج أن زيادة حجم العينة تساهم في تحسين تقدير العتبات، حيث تصبح النتائج أكثر دقة وتطابقًا مع القيم الحقيقية مع زيادة حجم العينة، وللكشف عن دلالة الفروق تم اجراء تحليل التباين الثنائي لمعرفة أثر حجم العينة ونوع النموذج في دقة تقدير معلمة الصعوبة كما يوضح الجدول (10)

جدول 10

نتائج تحليل التباين الثنائي للفروقات في تقدير متوسط معلمة الصعوبة تبعا لمتغيري حجم العينة ونوع النموذج المستخدم

المصدر	مجموع المربعات (SS)	درجات الحرية (df)	متوسط المربعات (MS)	قيمة F	الدلالة (p)
حجم العينة	0.98	2	0.49	2.85	**0.042
نوع النموذج	5.62	2	2.81	16.34	**0.000
التفاعل (حجم العينة × النموذج)	3.14	4	0.785	4.56	**0.002
الخطأ	29.42	172	0.172		
الإجمالي	39.16	180			

يوضح الجدول (10) نتائج تحليل التباين الثنائي، وتشير النتائج إلى وجود دلالة إحصائية لمتغير حجم العينة ومتغير نموذج الاستجابة المتعددة والتفاعل بينهما، ولتحديد دلالة الفروق تم اجراء اختبار (Bonferroni) للمقارنات البعدية كما يوضح الجدول (11)

جدول 11

نتائج اختبار الفروق البعدية بطريقة Bonferroni لأثر حجم العينة ونوع النموذج والتفاعل في دقة تقدير معلمة الصعوبة

المتغير المستقل		المقارنة	الفرق بين المتوسطات	الخطأ المعياري	قيمة p	لصالح	
حجم العينة		مقابل 100500	0.179	0.056	*0.034	500	
		مقابل 1001000	0.292	0.058	**0.001	1000	
		مقابل 5001000	0.167	0.054	0.062	لا يوجد فرق دال	
نوع النموذج		PCM مقابل GRM	0.284	0.064	**0.000	GRM	
		NRM مقابل GRM	0.007	0.062	0.198	لا يوجد فرق دال	
		NRM مقابل PCM	0.189	0.059	**0.009	NRM	
التفاعل	100 مفحوص	PCM مقابل GRM	0.276	0.061	0.015**	GRM	
		GRM مقابل NRM	0.061	0.042	0.41	لا يوجد فرق دال	
		NRM مقابل PCM	0.223	0.052	0.022**	NRM	
	500 مفحوص	PCM مقابل GRM	0.317	0.054	0.008**	GRM	
		GRM مقابل NRM	0.095	0.061	0.322	لا يوجد فرق دال	
		NRM مقابل PCM	0.222	0.036	0.018**	NRM	
	1000 مفحوص	PCM مقابل GRM	0.198	0.053	0.042**	GRM	
		GRM مقابل NRM	0.074	0.041	0.375	لا يوجد فرق دال	
		NRM مقابل PCM	0.124	0.053	0.24	لا يوجد فرق دال	

يوضح الجدول (11) نتائج اختبار الفروق البعدية باستخدام طريقة (Bonferroni) لمعرفة دلالة الفروق التي تعزى للمتغيرات (حجم العنية، نوع النموذج، التفاعل بين حجم العينة ونوع النموذج)، وفيما يلي توضيح لهذه النتائج:

أولا: متغير حجم العينة: يلاحظ من الجدول () أن الزيادة في حجم العينة تؤثر بشكل كبير على دقة التقدير. ففي المقارنة بين العينة المكونة من (100) مفحوص و(500) مفحوص، هناك فرق دال لصالح العينة الأكبر (500 مفحوص)، مما يدل على أن زيادة الحجم يحسن دقة التقدير بشكل ملحوظ. في المقابل، عندما تتم المقارنة بين (100) مفحوص و(1000) مفحوص، يظهر فرق دال أيضًا لصالح (1000) مفحوص ، إذ بلغت قيمة الدلالة الإحصائية (p = 0.001)، مما يثبت أن زيادة حج العينة يعزز من دقة التقدير بشكل أكبر. ومع ذلك فلا

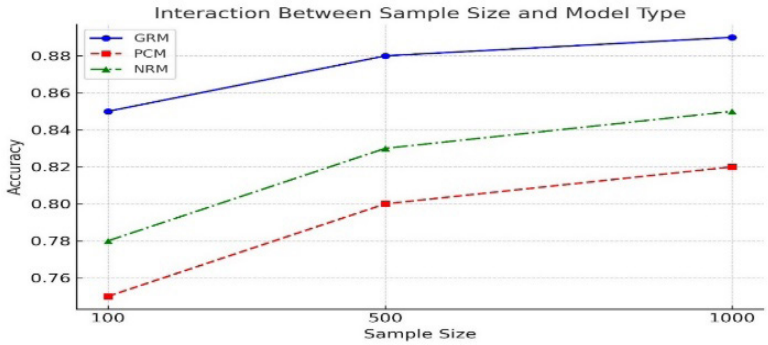
توجد فروق دالة عند المقارنة بين (500) مفحوص و(1000) مفحوص إذ بلغت قيمة الدلالة الإحصائية (p = 0.062)، مما يشير إلى أن هذه الزيادة في حجم العينة ليست مؤثرة بشكل كبير في تحسين دقة تقدير معالم عتبات الصعوبة.

ثانيا: متغير نوع النموذج: تشير النتائج إلى أن مقارنة الفروق بين نوعي النموذج (GRM) و (PCM) أظهرت فرقًا دالًا لصالح نموذج (GRM)، حيث بلغت قيمة الدلالة الإحصائية (= p 0.000)، مما يعني أن (GRM) يقدم دقة تقدير أفضل من (PCM). بالمقابل، لم تظهر المقارنة بين نموذجي (GRM) و (NRM) فرقًا دالًا (p = 0.198)، مما يشير إلى عدم وجود تأثير كبير لاختلاف النموذجين في دقة التقدير. أما المقارنة بين نموذج (PCM) و (NRM)، فقد أظهرت فروقًا ذات دلالة إحصائية لصالح نموذج (NRM) (p = 0.009)، مما يعكس تميز (NRM) في دقة التقدير مقارنة بـ(PCM).

ثالثا: التفاعل (حجم العينة * نوع النموذج): تشير النتائج إلى أن تأثير حجم العينة ونوع النموذج على دقة التقدير يختلف وفقًا لحجم العينة. في العينة الصغيرة (100 مفحوص)، تفوق نموذج (GRM) على (PCM) بشكل دال إحصائيًا، بينما لم تظهر فروق بين (GRM) و (NRM)، وكانت الفروق بين (NRM) و (PCM) لصالح (NRM). في العينة المتوسطة (500 مفحوص)، ظل (GRM) أكثر دقة من (PCM)، بينما لم توجد فروق دالة بين (GRM) و (NRM)، وكانت الفروق لصالح (NRM) عند المقارنة مع (PCM). أما في العينة الكبيرة (1000 مفحوص)، أظهر (GRM) فرقًا دالًا عند مقارنته بـ(PCM)، بينما لم توجد فروق بين (GRM) و (NRM) أو بين (NRM) و (PCM). بشكل عام، يؤثر حجم العينة بشكل ملحوظ على دقة التقدير، حيث كانت العينات الأكبر توفر تقديرات أكثر دقة، وكان (GRM) أكثر دقة في تقدير معالم الفقرات مقارنة بـ(PCM)، بينما كان (NRM) أكثر دقة قليلًا من (PCM)، والشكل (1) يوضح الرسم البياني للتفاعل بين حجم العينة ونوع النموذج

شكل (1)

التفاعل بين حجم العينة ونوع النموذج في دقة تقدير معلمة الصعوبة



وتم حساب الأوساط الحسابية والانحرافات المعيارية الخاصة بمؤشر التحييز (BAIS) كمؤشر لدقة تقدير معلمة الصعوبة تبعا لحجم العينة ولنموذج الاستجابة كما يوضح الجدول (12)

جدول 12

الأوساط الحسابية والانحرافات المعيارية لقيم مؤشرات التحييز تبعا لحجم العينة ونوع النموذج

العينة	العتبة	النموذج	القيمة الصغرى	القيمة العظمى	الوسط الحسابي	الانحراف المعياري
مفحوص100	الأولى	GRM	-2.321	1.65	-0.141	0.333
		PCM	-1.6	1	-0.425	0.24
		NRM	-2.82	1.703	-0.202	0.283
	الثانية	GRM	-1.439	1.65	-0.141	0.333
		PCM	-1.2	1	-0.425	0.24
		NRM	-1.382	1.703	-0.202	0.283
	الثالثة	GRM	1.548	1.65	-0.141	0.333
		PCM	0.1	1	-0.425	0.24
		NRM	1.692	1.703	-0.202	0.283
	الرابعة	GRM	1.65	1.65	-0.141	0.333
		PCM	1	1	-0.425	0.24
		NRM	1.703	1.703	-0.202	0.283
مفحوص500	الأولى	GRM	-1.826	1.82	-0.041	0.303
		PCM	-2.316	1.302	-0.452	0.233
		NRM	-2.012	1.921	-0.077	0.26
	الثانية	GRM	-0.851	1.82	-0.041	0.303
		PCM	-1.547	1.302	-0.452	0.233
		NRM	-0.714	1.921	-0.077	0.26
	الثالثة	GRM	0.693	1.82	-0.041	0.303
		PCM	0.753	1.302	-0.452	0.233
		NRM	0.496	1.921	-0.077	0.26
	الرابعة	GRM	1.82	1.82	-0.041	0.303
		PCM	1.302	1.302	-0.452	0.233
		NRM	1.921	1.921	-0.077	0.26

0100 مفحوص	الأولى	GRM	-1.6	1.624	0.086	0.265
		PCM	-2.321	1.923	-0.046	0.213
		NRM	-1.765	1.543	0.09	0.24
	الثانية	GRM	-0.492	1.624	0.086	0.265
		PCM	-0.721	1.923	-0.046	0.213
		NRM	-0.386	1.543	0.09	0.24
	الثالثة	GRM	0.813	1.624	0.086	0.265
		PCM	0.937	1.923	-0.046	0.213
		NRM	0.967	1.543	0.09	0.24
	الرابعة	GRM	1.624	1.624	0.086	0.265
		PCM	1.923	1.923	-0.046	0.213
		NRM	1.543	1.543	0.09	0.24

وكذلك تم حساب الأوساط الحسابية والانحرافات المعيارية للجذر التربيعي لمعدل المربعات للأخطاء (RMSE) كمؤشر آخر على دقة تقديرات الصعوبة للعتبات تبعا لحجم العينة ونموذج الاستجابة كما يوضح الجدول (13)

جدول 13

الأوساط الحسابية والانحرافات المعيارية للأخطاء المعيارية لمعالم الصعوبة للعتبات الفارقة

العينة	الخطأ المعياري لعتبات الصعوبة	النموذج						الكلي	
		GRM		PCM		NRM			
		الوسط الحسابي	الانحراف المعياري	الوسط الحسابي	الانحراف المعياري	الوسط الحسابي	الانحراف المعياري	الوسط الحسابي	الانحراف المعياري
100 مفحوص	الأولى	0.160	0.011	1.421	0.382	0.391	0.300	0.657	0.231
	الثانية	0.130	0.031	0.330	0.380	0.164	0.327	0.208	0.246
	الثالثة	0.041	0.023	0.520	0.240	0.075	0.239	0.212	0.167
	الرابعة	0.020	0.103	0.683	0.220	0.064	0.200	0.256	0.174
	الكلي	0.088	0.042	0.739	0.306	0.174	0.267		
500 مفحوص	الأولى	0.152	0.310	0.300	0.300	0.192	0.280	0.215	0.297
	الثانية	0.098	0.287	0.240	0.280	0.173	0.250	0.170	0.272
	الثالثة	0.033	0.216	0.425	0.252	0.043	0.180	0.167	0.216
	الرابعة	0.063	0.292	0.592	0.213	0.058	0.220	0.238	0.242
	الكلي	0.087	0.276	0.389	0.261	0.117	0.233		

1000 مفحوص	الأولى	0.015	0.277	0.280	0.280	0.187	0.250	0.161	0.269
	الثانية	0.068	0.247	0.240	0.240	0.142	0.220	0.150	0.236
	الثالثة	0.054	0.217	0.220	0.320	0.036	0.180	0.137	0.206
	الرابعة	0.012	0.217	0.220	0.420	0.020	0.200	0.151	0.212
	الكلي	0.037	0.240	0.240	0.315	0.096	0.213		

تشير النتائج في الجدول (13) إلى أن نموذج (GRM) يقدم أقل متوسط للخطأ المعياري عبر جميع العينات مقارنة بنماذج (PCM) و (NRM)، مما يعكس دقته العالية في تقدير عتبات الصعوبة. فعلى سبيل المثال، في العينة المكونة من (100) مفحوص، كان متوسط الخطأ المعياري لنموذج (GRM) (0.088)، بينما كانت النماذج الأخرى تحتوي على متوسطات أعلى. مع زيادة حجم العينة من (100) إلى (1000) مفحوص، يقل الخطأ المعياري في جميع النماذج، مما يعكس استقرار النموذج ودقته مع زيادة عدد المفحوصين. في العينة (1000) مفحوص، كان الخطأ المعياري لنموذج (GRM) (0.037)، وهو الأقل بين النماذج.

كما يظهر الجدول تباينًا في تقدير عتبات الصعوبة عبر الفقرات المختلفة. فعند مقارنة (GRM) مع (PCM) في العينة (100) مفحوص، يظهر أن (GRM) يقدم تقديرات أكثر دقة وأقل خطأ معياري، بينما يظهر (PCM) تباينًا أكبر في بعض الفقرات، مما يشير إلى أن (GRM) هو الأكثر فعالية في العينات الأصغر. بشكل عام، يمكن الاستنتاج أن نموذج (GRM) يتفوق من حيث دقة التقديرات عبر جميع العينات، ويزداد استقراره مع زيادة حجم العينة، يليه نموذج (NRM)، وأخيرًا نموذج (PCM).

تتفق هذه النتيجة مع ما أشار إليه (Stone & Youmoto, 2004) في أن نموذج الاستجابة المتدرجة (GRM) يحقق دقة أعلى في تقدير معلمة الصعوبة مقارنة بنموذجي (PCM, NRM). كما أظهرت دراسة (Suwarto & Setiawan, 2019) أن زيادة حجم العينة تؤدي إلى تحسن دقة تقدير معالم الفقرات، وهو ما يعزز من صحة تقدير معلمة الصعوبة.

النتائج المتعلقة بالسؤال الثاني والذي نص على» هل تختلف دقة تقدير معلمة التمييز بين النماذج متعددة التدريج (GRM, PCM, NRM) باختلاف حجم العينة ؟»

للإجابة عن هذا السؤال تم تقدير معلمة التمييز باختلاف العينات (100، 500، 1000) مفحوص وباختلاف نموذج التقدير (GRM, PCM, NRM) وتم حساب الأوساط الحسابية والانحرافات المعيارية لتقديرات التمييز كما يوضح الجدول (14)

جدول 14

الأوساط الحسابية والانحرافات المعيارية لمعالم التمييز تبعا لحجم العينة ونوع النموذج

العينة	النموذج	الوسط الحسابي	الانحراف المعياري
مفحوص100	GRM	1.367	0.123
	PCM	0.836	0.452
	NRM	0.952	0.350
مفحوص500	GRM	1.492	0.186
	PCM	0.752	0.452
	NRM	0.935	0.445
مفحوص1000	GRM	1.579	0.152
	PCM	0.761	0.452
	NRM	0.857	0.423

يتضح من الجدول (14) أن قيم متوسطات التمييز المقدرة بطريقة (GRM) هي الأقرب للقيم الحقيقية ففي العينة المكونة من 100 مفحوص بلغت قيمة متوسط التمييز (1.367) في حين كانت القيمة الحقيقية (1.415)، تلاها القيمة المقدرة بطريقة (NRM) بمتوسط بلغ (0.952) واخيرا نموذج (PCM) بقيمة بلغت (0.836)، وعند زيادة حجم العينة فأن نموذج (GRM) ما زال أقرب نموذج في تقديره لمعلمة التمييز ويزداد قربه للقيمة الحقيقية بزيادة حجم العينة، يليه نموذج (NRM) وأخيرا نموذج (PCM).

وللكشف عن دلالة الفروق تم اجراء اختبار تحليل التباين الثنائي لمعرفة أثر حجم العينة ونوع النموذج في دقة تقدير معلمة التمييز كما يوضح الجدول (15)

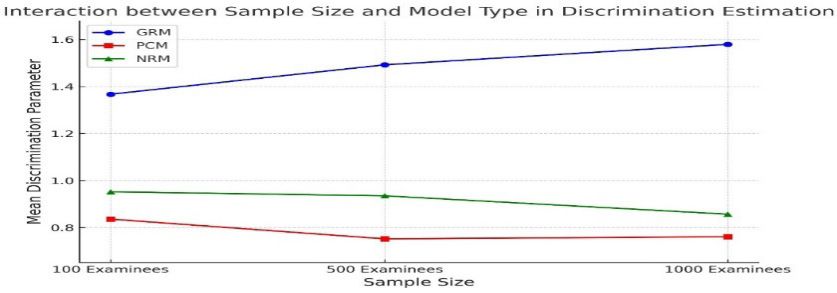
جدول 15

نتائج تحليل التباين الثنائي لتقدير معلمة التمييز تبعا لحجم العينة ونوع النموذج المستخدم في التقدير

المصدر	مجموع المربعات (SS)	درجات الحرية (df)	متوسط المربعات (MS)	قيمة F	الدلالة (p)
حجم العينة	0.775	2	0.3875	37.85	**0.01
نوع النموذج	1.345	2	0.6725	66.73	**0.001
التفاعل (حجم العينة × النموذج)	0.432	4	0.108	10.68	**0.001
الخطأ	0.118	172	0.0007		
الإجمالي	**2.670**	**180**			

يوضح الجدول (15) نتائج تحليل التباين الثنائي، وتشير النتائج إلى وجود دلالة إحصائية لمتغير حجم العينة ومتغير نموذج الاستجابة المتعددة والتفاعل بينهما، والشكل (2) يوضح التمثيل البياني للتفاعل بين حجم العينة والنموذج .

شكل 2: التفاعل بين حجم العينة ونموذج الاستجابة المتعدد في تقدير معلمة التمييز



ولتحديد دلالة الفروق تم اجراء اختبار (Bonferroni) للمقارنات البعدية كما يوضح الجدول (16)

جدول 16

نتائج اختبار الفروق البعدية بطريقة Bonferroni لأثر حجم العينة ونوع النموذج والتفاعل في دقة تقدير معلمة التمييز

المتغير المستقل		المقارنة	الفرق بين المتوسطات	الخطأ المعياري	قيمة p	لصالح
حجم العينة		100 مقابل 500	0.102	0.038	*0.012	500 مفحوص
		100 مقابل 1000	0.178	**0.001	1000مفحوص	
		500 مقابل 1000	0.076	*0.028	1000مفحوص	
نوع النموذج		PCM مقابل GRM	0.721	0.056	0.001**	GRM
		GRM مقابل NRM	0.512	**0.002	GRM	
		PCM مقابل NRM	0.209	*0.035	NRM	
التفاعل	مفحوص100	PCM مقابل GRM	0.531	0.062	0.001**	GRM
		NRM مقابل GRM	0.415	0.058	0.007**	GRM
		NRM مقابل PCM	0.116	0.048	0.041*	NRM
	مفحوص500	PCM مقابل GRM	0.74	0.055	0.001**	GRM
		NRM مقابل GRM	0.557	0.052	0.002**	GRM
		NRM مقابل PCM	0.183	0.047	0.036*	NRM
	1000 مفحوص	PCM مقابل GRM	0.818	0.053	0.001**	GRM
		NRM مقابل GRM	0.722	0.05	0.002**	GRM
		NRM مقابل PCM	0.096	0.045	0.045*	NRM

يشير الجدول (16) إلى نتائج اختبار (Bonferroni) للمقارنات البعدية، والذي تم إجراؤه لتحديد الفروق الدالة إحصائيًا في دقة تقدير معلمة التمييز تبعًا لحجم العينة، ونوع النموذج المستخدم، والتفاعل بينهم، وفيما يلي عرض لنتائج المقارنات.

أولًا: أثر حجم العينة: تشير النتائج إلى وجود فروق دالة إحصائيًا بين مختلف أحجام العينات، إذ كانت الفروق دالة بين العينة (100) مقابل (500) مفحوص (p = 0.012) لصالح العينة الأكبر (500) مفحوص، كما ظهرت فروق دالة بين (100) مقابل (1000) مفحوص (p = 0.001)، لصالح العينة (1000) مفحوص، وكذلك أظهرت المقارنة بين (500) و(1000) مفحوص وجود فروق دالة (p = 0.028) لصالح العينة (1000) مفحوص، مما يؤكد أن زيادة حجم العينة يساهم في تحسين دقة تقدير معلمة التمييز.

ثانيًا: أثر نوع النموذج: أظهرت النتائج فروقًا دالة إحصائيًا بين النماذج المختلفة. إذ تفوق نموذج (GRM) على نموذج (PCM) بفارق بلغ 0.721 (p = 0.001)، مما يشير إلى أن (GRM) يوفر تقديرات أكثر دقة لمعلمة التمييز مقارنةً بنموذج (PCM)، كما تفوق نموذج (GRM) على نموذج (NRM) بفارق 0.512 (p = 0.002)، مما يؤكد أن نموذج (GRM) يظل النموذج الأكثر دقة. أما المقارنة بين (PCM) و (NRM) فقد أظهرت فرقًا دالًا إحصائيًا (p = 0.035) لصالح نموذج (NRM) ، مما يعني أن نموذج (NRM) يوفر تقديرات أكثر دقة من نموذج (PCM).

ثالثًا: أثر التفاعل بين حجم العينة ونوع النموذج: أظهرت نتائج التفاعل بين حجم العينة ونوع النموذج أن نموذج (GRM) كان الأكثر دقة في جميع أحجام العينة. فعند العينة (100) مفحوص، تفوق (GRM) على (PCM) بفارق دال إحصائيًا (p = 0.001) وعلى (p) (NRM = 0.007)، مما يشير إلى استقراره حتى مع العينات الصغيرة، بينما كان (PCM) الأقل دقة مقارنة بـ(p = 0.041) (NRM). وعند العينة (500) مفحوص، استمر تفوق (GRM) على ((PCM p = 0.001)) وعلى (p = 0.002) (NRM)، كما تفوق (NRM) على (p = 0.036) (PCM). أما عند العينة (1000) مفحوص، فقد ازدادت الفروق وضوحًا، حيث تفوق (GRM) بفارق (0.818) (p = 0.001) عن (PCM) وبفارق (0.722) (p = 0.002) عن (NRM)، بينما بقي الفرق بين (PCM) و(NRM) دالًا (p = 0.045) لصالح (NRM).

وبشكل عام فإن حجم العينة له تأثير كبير على دقة التقدير، حيث تؤدي الزيادة في حجم العينة إلى تحسين دقة تقدير معلمة التمييز. كما تبين أن نموذج (GRM) هو النموذج الأكثر دقة في جميع أحجام العينة، يليه نموذج (NRM) وثم نموذج (PCM) الذي كان الأقل دقة، كما يظهر التفاعل بين حجم العينة ونوع النموذج أن نموذج (GRM) يحافظ على دقته حتى مع العينات الصغيرة، لكنه يصبح أكثر دقة مع زيادة حجم العينة.

ثم تم حساب الأوساط الحسابية والانحرافات المعيارية لمؤشر التحييز (BIAS) كمؤشر من مؤشرات دقة التقدير لمعالم التمييز تبعا لحجم العينة ونوع النموذج كما يوضح الجدول (17)

جدول 17

قيم مؤشر التحييز لمعلمة التمييز تبعا لحجم العنية ونوع النموذج

العينة	النموذج	القيمة الصغرى	القيمة العظمى	الوسط الحسابي	الانحراف المعياري
مفحوص100	GRM	0.0217	0.1483	0.0842	0.0456
	PCM	0.6029	0.8985	0.7514	0.0658
	NRM	0.6541	0.9472	0.8026	0.0703
مفحوص500	GRM	0.0115	0.1187	0.0639	0.0402
	PCM	0.5532	0.8476	0.7028	0.0609
	NRM	0.6084	0.9031	0.7547	0.0654
مفحوص1000	GRM	0.0053	0.0978	0.0496	0.0351
	PCM	0.5017	0.7982	0.6763	0.0557
	NRM	0.5539	0.8526	0.7268	0.0605

يلاحظ من الجدول (17) أن قيمة الوسط الحسابي لمؤشر التحييز لدقة تقدير معلمة التمييز كانت أقل ما يمكن لنموذج (GRM) تلاه نموذج (NRM) وفي المرتبة الأخيرة نموذج (PCM)، ثم تم حساب الأوساط الحسابية والانحرافات المعيارية للجذر التربيعي لمعدل مربعات الأخطاء (RMSE) تبعا لحجم العينة ونوع النموذج كما يوضح الجدول (18)

جدول 18

الأوساط الحسابية والانحرافات المعيارية للخطأ المعياري لتقديرات معلمة التمييز تبعا لحجم العينة ونوع النموذج

العينة	النموذج	الوسط الحسابي	الانحراف المعياري
مفحوص100	GRM	0.0317	0.0452
	PCM	0.0725	0.0896
	NRM	0.0534	0.0673
مفحوص500	GRM	0.0208	0.0329
	PCM	0.0612	0.0784
	NRM	0.0426	0.0561
مفحوص1000	GRM	0.0154	0.0273
	PCM	0.0541	0.0715
	NRM	0.0378	0.0497

يشير الجدول (18) إلى وجود فروق ظاهرية بين الأوساط الحسابية لقيم (RMSE) لتقديرات التمييز تبعا لحجم العينة ونوع النموذج المستخدم في التقدير، وللكشف عن دلالة الفروق تم استخدام اختبار تحليل التباين الثنائي لمعرفة اثر حجم العينة ونوع النموذج المستخدم في التقدير في قيم الخطأ المعياري في التقدير كما يوضح الجدول (19)

جدول 19

نتائج تحليل التباين الثنائي لأثر حجم العينة ونوع النموذج في قيم الأخطاء المعيارية

المصدر	مجموع المربعات (SS)	درجات الحرية (df)	متوسط المربعات (MS)	قيمة F	مستوى الدلالة	حجم الأثر
حجم العينة (100، 500، 1000 مفحوص)	0.024	2	0.012	32.15	0.0001	0.38
نوع النموذج (GRM, PCM, NRM)	0.018	2	0.009	24.85	0.0003	0.29
التفاعل بين حجم العينة والنموذج	0.006	4	0.0015	6.32	0.002	0.14
الخطأ	0.022	172	0.000129			
المجموع الكلي	0.07	180				

يبين الجدول (19) وجود فروق دالة إحصائيًا في قيم الأخطاء المعيارية لتقدير المعلمات تبعًا لمتغيري حجم العينة ونوع النموذج المستخدم، بالإضافة إلى وجود تفاعل دال بينهما، مما يعني أن تأثير حجم العينة على الخطأ المعياري يختلف باختلاف النموذج المستخدم، ففيما يتعلق بمتغير حجم العينة فتشير النتائج إلى وجود فرق دال إحصائيًا بين أحجام العينات الثلاثة، وكذلك كانت قيمة مؤشر حجم الأثر(0.38) مما يشير إلى تأثير قوي، وهذا يعني أن جزءًا كبيرًا من التباين في قيم الخطأ المعياري يمكن تفسيره بحجم العينة، وكلما زاد حجم العينة قل الخطأ المعياري، مما يعزز أهمية استخدام عينات أكبر للحصول على تقديرات أكثر دقة.

أما فيما يتعلق بتأثير نوع النموذج، فقد أشارت النتائج إلى وجود فرق دال إحصائيًا بين النماذج الثلاثة، وكذلك بلغت قيمة مؤشر حجم الأثر (0.29) مما يشير إلى أن نوع النموذج يؤثر بشكل كبير على دقة التقديرات.. وأظهر نموذج ((GRM أقل خطأ معياري، تلاه نموذج (NRM) في حين كان نموذج (PCM) أقل النماذج دقة في تقدير معلمة التمييز.

وفيما يتعلق بمتغير التفاعل بين حجم العينة والنموذج، فقد أظهرت النتائج وجود دلالة إحصائية مما يشير إلى أن تأثير حجم العينة يختلف باختلاف النموذج، هذا يعني أن بعض النماذج مثل نموذج (PCM) قد تكون أقل حساسية لحجم العينة مقارنةً بنموذج (GRM) والذي قد يستفيد أكثر من زيادة حجم العينة.

وللكشف عن دلالة الفروق تم اجراء اختبار Bonferroni للمقارنات البعدية كما يوضح الجدول (20)

جدول 20

نتائج اختبار Bonferroni للمقارنات البعدية لمتوسط الأخطاء المعيارية والتي تعزى لحجم العينة ونوع النموذج المستخدم في تقدير معلمة التمييز

المستقل	المقارنة		الفرق بين المتوسطات	الخطأ المعياري	قيمة p	لصالح
حجم العينة	مقابل 100500		0.0047-	0.0001	0.0001	500
	مقابل 1001000		0.0062-	0.0001	0.0001	1000
	مقابل 5001000		0.0015-	0.0001	0.006	1000
نوع النموذج	PCM مقابل GRM		0.0381-	0.007	0.0001	GRM
	NRM مقابل GRM		0.0217-	0.006	0.001	GRM
	NRM مقابل PCM		0.0164-	0.008	0.042	NRM
التفاعل	مفحوص100	PCM مقابل GRM	0.0381-	0.007	0.0001	GRM
		NRM مقابل GRM	0.0217-	0.006	0.001	GRM
		NRM مقابل PCM	0.0164-	0.008	0.042	NRM
	مفحوص500	PCM مقابل GRM	0.0208-	0.0329	0.076	GRM
		NRM مقابل GRM	0.0161-	0.0303	0.226	GRM
		NRM مقابل PCM	0.0047	0.0363	0.679	PCM
	مفحوص1000	PCM مقابل GRM	0.0387-	0.0275	0.154	GRM
		NRM مقابل GRM	0.0274-	0.0295	0.358	GRM
		NRM مقابل PCM	0.0113	0.0429	0.797	PCM

تشير نتائج المقارنات البعدية إلى وجود دلالة إحصائية تعزى لحجم العينة لصالح العينة الأكبر، فكلما زاد حجم العينة قلت قيمة المتوسط الحسابي للأخطاء المعيارية، أما فيما يتعلق بمتغير النموذج فقد أشارت النتائج إلى وجود فروق تعزى لنوع النموذج إذ كانت قيم المتوسطات الحسابية للأخطاء المعيارية أقل ما يمكن لنموذج (GRM) تلاه نموذج (NRM) ثم نموذج (PCM)، وفيما يتعلق بالتفاعل بين حجم العينة ونوع النموذج فقد أظهرت النتائج أنه عند العينة المكونة من (100) مفحوص ، كان نموذج (GRM) الأفضل مقارنة بنموذجي (NRM) و (PCM)، أما في حال العينتين (500) و (1000) مفحوص فلا يوجد تفاعل قوي بين حجم العينة والنموذج من حيث تأثير الخطأ المعياري إذ تقلصت قيم الفروق، ولكن بقي نموذج (GRM) أفضل من نموذجي (NRM) و (PCM).

وتتفق هذه النتيجة مع دراسة (Livingston, 2020) التي أشارت إلى أن النموذج الأكثر دقة في تقدير معلمة التمييز هو الذي يحقق أفضل مستوى من التمييز بين المفحوصين. كما أظهرت دراسات أخرى مثل (الشريفين، 2019) أن زيادة حجم العينة تسهم في تحسين دقة تقدير معلمة التمييز، مما يؤدي إلى تقديرات أكثر دقة للفقرات التي تتعلق بالتمييز بين المستويات المختلفة للسمات.

النتائج المتعلقة بالسؤال الثالث والذي نص على» هل تختلف دقة تقدير معلمة القدرة للأفراد بين النماذج متعددة التدريج (GRM, PCM, NRM) باختلاف حجم العينة ؟»

للإجابة عن هذا السؤال تم تقدير معلمة القدرة لكل فرد باختلاف أحجام العينات (100، 500، 1000 مفحوص) وباختلاف نوع النموذج (GRM، PCN، NRM) واستخراج الوسط الحسابي والانحراف المعياري للقيم المقدرة كما يوضح الجدول (21).

جدول 21

الأوساط الحسابية والانحرافات المعيارية لتقديرات القدرة تبعا لحجم العينة ونوع النموذج

العينة	النموذج	الوسط الحسابي	الانحراف المعياري
مفحوص100	GRM	-0.25	0.502
	PCM	-0.498	0.642
	NRM	-0.348	0.551
مفحوص500	GRM	-0.03	0.349
	PCM	-0.205	0.498
	NRM	-0.106	0.457
مفحوص1000	GRM	0.010	0.251
	PCM	-0.103	0.402
	NRM	0.000	0.351

يلاحظ من الجدول (21) أن نموذج (GRM) كان النموذج الأعلى دقة في تقدير معلمة القدرة عبر جميع أحجام العينات. ففي العينة التي تضم 100 مفحوص، كانت بلغت القيمة المقدرة (0.25 -) وهي قريبة من القيمة الحقيقية (0.291 -)، ومع زيادة حجم العينة إلى 500 و1000 مفحوص، تتقارب القيم بشكل أكبر مع القدرة الحقيقية، حيث كانت القيمة المقدرة (0.03 -) مقارنة بالقيمة الحقيقية (0.018) لعينة (500) مفحوص، و(0.01) مقارنة بالقيمة الحقيقية (0.011) لعينة 1000، مما يعكس التحسن المستمر في دقة النموذج مع زيادة حجم العينة.

أما بالنسبة لنموذج (NRM)، فقد أظهر تحسينًا ملحوظًا في دقته مع زيادة حجم العينة، لكنه ظل أقل دقة من نموذج (GRM) في جميع الحالات. ففي العينة الصغيرة (100 مفحوص) بلغت القيمة المقدرة وفق نموذج (NRM) لمعلمة التمييز (0.348 -) مقارنة بالقيمة الحقيقية (0.291 -)،أما في عينة (500) مفحوص فقد بلغت القيمة المقدرة (0.106 -) مقارنة بالقيمة الحقيقية (0.018 -)، وفي العينة المكونة من (1000) مفحوص بلغت القيمة المقدرة (0.00) مقابل القيمة الحقيقية (0.011) ما يعكس بعض التحسن في دقة تقدير نموذج (NRM) بازدياد حجم العينة.

ومن جهة أخرى فإن دقة تقدير نموذج (PCM) كانت الأقل في جميع الأحجام، حيث كانت قيمته في عينة (100) مفحوص (0.498 -) مقابل القيمة الحقيقية (0.291 -)، ثم

تحسنت إلى (0.205 -) في عينة (500) مفحوص مقارنة بالقيمة الحقيقية (0.018 -)، وأخيرًا بلغت القيمة المقدرة في عينة (1000) مفحوص (0.103 -) مقابل القيمة الحقيقية (0.011)، ولكنها بقيت بعيدة بشكل ملحوظ عن القيمة الحقيقية في كل حالة.

وبشكل عام يمكن استنتاج أن نموذج (GRM) هو النموذج الأكثر دقة في تقدير معلمة القدرة عبر جميع الأحجام، ويزداد تحسنه بشكل ملحوظ مع زيادة حجم العينة.

وللكشف عن دلالة الفروق في دقة تقدير معلمة القدرة عبر أحجام العينات المختلفة ونوع النموذج المستخدم في التقدير تم إجراء اختبار تحليل التباين الثنائي كما يوضح الجدول (22)

جدول 22

نتائج تحليل التباين الثنائي لتقديرات القدرة باختلاف حجم العينة ونوع النموذج المستخدم في تقديرها

المصدر	مجموع المربعات (SS)	درجات الحرية (df)	متوسط المربعات (MS)	قيمة F	مستوى الدلالة	حجم الأثر
حجم العينة (100، 500، 1000 مفحوص)	0.185	2	0.092	15.82	0.00	0.198
نوع النموذج (GRM, PCM, NRM)	0.325	2	0.162	27.89	0.00	0.312
التفاعل بين حجم العينة والنموذج	0.075	4	0.019	3.28	0.013	0.079
الخطأ	1.125	172	0.007			
المجموع الكلي	1.71	179				

يظهر جدول(22) نتائج تحليل التباين الثنائي لمعرفة أثر (حجم العينة، نوع النموذج، التفاعل بينهما)، على دقة تقدير معلمة القدرة، إذ تشير النتائج إلى وجود فروق ذات دلالة إحصائية في تقديرات القدرة تبعًا لكل من حجم العينة والنموذج المستخدم، وكذلك وجود تفاعل بينهما، فعند دراسة متغير حجم العينة فقد بلغت قيمة(F(15.82 عند مستوى دلالة (0.00)، مما يشير إلى وجود فروق دالة إحصائيًا بين تقديرات القدرة عند أحجام العينات المختلفة (100، 500، 1000 مفحوص)، كما أن حجم الأثر (0.198) يشير إلى أن حجم العينة له تأثير متوسط على دقة التقدير، مما يعكس تحسن دقة التقديرات مع زيادة حجم العينة.

أما فيما يتعلق بمتغير نوع النموذج فقد بلغت قيمة(F (27.8 عند مستوى دلالة (0.00)، مما يشير إلى أن النموذج المستخدم في التقدير يؤثر بشكل كبير على دقة تقدير معلمة القدرة، كما أن حجم الأثر بلغ (0.312)، مما يعني أن اختيار النموذج المناسب يلعب دورًا أكثر أهمية في دقة التقدير مقارنة بحجم العينة وحده.

وفميا يتلعق بالتفاعل بين حجم العينة ونوع النموذج فقد بلغت قيمة (3.28) F عند مستوى دلالة (0.013)، مما يشير إلى وجود تفاعل دال إحصائيًا بين المتغيرين. ولكن حجم الأثر كان منخفضًا (0.079)، وهذا يعني أن تأثير التفاعل ليس قويًا جدًا، ولكنه لا يزال يشير إلى أن أداء النموذج يتأثر بحجم العينة.

ولاختبار دلالة الفروق تم اجراء اختبار المقارنات البعدية بطريقة Bonferroni كما يوضح الجدول 23

جدول 23

نتائج اختبار Bonferroni للمقارنات البعدية للفروق في دقة تقدير معلمة القدرة والتي تعزى لحجم العينة ونوع النموذج المستخدم

المستقل	المقارنة		الفرق بين المتوسطات	الخطأ المعياري	قيمة p	لصالح
حجم العينة	مقابل 100500		0.218	0.055	0.001	500
	مقابل 1001000		0.26	0.051	0	1000
	مقابل 5001000		0.042	0.045	0.368	لا يوجد فرق
نوع النموذج	PCM مقابل GRM		0.295	0.056	0	GRM
	NRM مقابل GRM		0.121	0.049	0.014	GRM
	NRM مقابل PCM		-0.174	0.052	0.002	NRM
التفاعل	مفحوص100	GRM مقابل PCM	0.248	0.074	0.002	GRM
		GRM مقابل NRM	0.098	0.066	0.148	لا يوجد فرق
		NRM مقابل PCM	-0.15	0.069	0.031	NRM
	مفحوص500	GRM مقابل PCM	0.175	0.062	0.006	GRM
		GRM مقابل NRM	0.076	0.058	0.195	لا يوجد فرق
		NRM مقابل PCM	-0.099	0.061	0.112	لا يوجد فرق
	مفحوص1000	GRM مقابل PCM	0.113	0.054	0.039	GRM
		GRM مقابل NRM	0.01	0.048	0.837	لا يوجد فرق
		NRM مقابل PCM	-0.103	0.051	0.043	NRM

أظهرت النتائج المعروضة في جدول (23) أن حجم العينة يؤثر بشكل دال إحصائيًا على دقة تقدير القدرة، حيث كانت التقديرات أكثر دقة عند زيادة حجم العينة، إذ تفوقت عينة (500) مفحوص على عينة (100) مفحوص ($p = 0.001$)، وعينة (1000) مفحوص على عينة (100) مفحوص ($p < 0.001$)، بينما لم تكن هناك فروق دالة بين عينتي (500) و(1000) مفحوص ($p = 0.368$)، مما يشير إلى أن التحسن في دقة التقدير يصبح أقل مع زيادة حجم العينة. كما تبين أن نموذج (GRM) كان الأكثر دقة في تقدير القدرة، حيث تفوق على نموذج (PCM) ($p < 0.001$) وعلى نموذج (NRM) ($p = 0.014$)، في حين كان نموذج (NRM) أكثر دقة من نموذج (PCM) ($p = 0.002$). أما بالنسبة للتفاعل بين حجم العينة والنموذج، فقد ظهر بشكل واضح عند العينة الصغيرة (100 مفحوص)، حيث تفوق نموذج (GRM) على (PCM)، بينما لم تكن الفروق دالة بين (GRM) و(NRM). وعند العينة (500) مفحوص، استمر تفوق (GRM) على (PCM)، لكن باقي الفروق لم تكن دالة، أما عند العينة (1000) مفحوص، فقد تفوق (GRM) على (PCM) بينما لم يكن هناك فرق بين (GRM) و(NRM)، مع وجود فروق دالة لصالح (NRM) عند مقارنته بـ(PCM).

وبشكل عام، تؤكد النتائج أن (GRM) هو النموذج الأكثر دقة بغض النظر عن حجم العينة، يليه (NRM)، بينما كان (PCM) الأقل دقة، كما أن زيادة حجم العينة تعزز دقة التقدير، لكن تأثير التفاعل بين حجم العينة والنموذج يصبح أقل أهمية مع العينات الأكبر.

وتم أيضا حساب الأوساط الحسابية والانحرافات المعيارية لمؤشر التحييز، كما يوضح الجدول 24

جدول 24

قيم مؤشر التحييز (BIAS) لمعلمة القدرة والتي تعزى لحجم العينة ونوع النموذج

العينة	النموذج	القيمة الصغرى	القيمة العظمى	الوسط الحسابي	الانحراف المعياري
مفحوص100	GRM	-0.800	0.500	-0.250	0.502
	PCM	-1.000	0.800	-0.498	0.642
	NRM	-0.750	0.600	-0.348	0.551
مفحوص500	GRM	-0.500	0.400	-0.030	0.349
	PCM	-0.700	0.600	-0.205	0.498
	NRM	-0.600	0.500	-0.106	0.457
مفحوص1000	GRM	-0.300	0.300	0.010	0.251
	PCM	-0.500	0.400	-0.103	0.402
	NRM	-0.400	0.300	0.000	0.351

يلاحظ من الجدول (24) أن قيمة الوسط الحسابي لمؤشر التحييز لدقة تقدير معلمة القدرة كانت أقل ما يمكن لنموذج (GRM) تلاه نموذج (NRM) وفي المرتبة الأخيرة نموذج (PCM)، ثم تم حساب الأوساط الحسابية والانحرافات المعيارية للجذر التربيعي لمعدل مربعات الأخطاء (RMSE) تبعا لحجم العينة ونوع النموذج كما يوضح الجدول (25)

جدول 25

الأوساط الحسابية والانحرافات المعيارية للخطأ المعياري لتقديرات معلمة القدرة تبعا لحجم العينة ونوع النموذج

العينة	النموذج	الوسط الحسابي	الانحراف المعياري
مفحوص100	GRM	0.326	0.087
	PCM	0.419	0.106
	NRM	0.357	0.094
مفحوص500	GRM	0.228	0.058
	PCM	0.314	0.076
	NRM	0.257	0.063

مفحوص1000	GRM	0.183	0.042
	PCM	0.267	0.068
	NRM	0.224	0.053

يشير الجدول (25) إلى وجود فروق ظاهرية بين الأوساط الحسابية لقيم (RMSE) لتقديرات القدرة تبعا لحجم العينة ونوع النموذج المستخدم في التقدير، وللكشف عن دلالة الفروق تم استخدام اختبار تحليل التباين الثنائي لمعرفة اثر حجم العينة ونوع النموذج المستخدم في التقدير في قيم الخطأ المعياري في تقدير القدرة كما يوضح الجدول (26)

جدول 26

نتائج تحليل التباين الثنائي لأثر حجم العينة ونوع النموذج في قيم الأخطاء المعيارية

المصدر	مجموع المربعات (SS)	درجات الحرية (df)	متوسط المربعات (MS)	قيمة F	مستوى الدلالة	حجم الأثر
حجم العينة (100، 500، 1000 مفحوص)	1.245	2	0.623	8.932	**0.001	0.094
نوع النموذج (GRM, PCM, NRM)	1.876	2	0.938	13.451	**0.000	0.112
التفاعل بين حجم العينة والنموذج	0.764	4	0.191	2.741	*0.032	0.048
الخطأ	11.998	172	0.07			
المجموع الكلي	15.883	180				

يبين الجدول (26) وجود فروق دالة إحصائيًا في قيم الأخطاء المعيارية لتقدير المعلمات تبعًا لمتغيري حجم العينة ونوع النموذج المستخدم، بالإضافة إلى وجود تفاعل دال بينهما، مما يعني أن تأثير حجم العينة على الخطأ المعياري يختلف باختلاف النموذج المستخدم، ففيما يتعلق بمتغير حجم العينة فتشير النتائج إلى وجود فرق دال إحصائيًا بين أحجام العينات الثلاثة.

أما فيما يتعلق بتأثير نوع النموذج، فقد أشارت النتائج إلى وجود فرق دال إحصائيًا بين النماذج الثلاثة، مما يشير إلى أن نوع النموذج يؤثر بشكل كبير على دقة التقديرات.، وأظهر نموذج ((GRM أقل خطأ معياري، تلاه نموذج (NRM) في حين كان نموذج (PCM) أقل النماذج دقة في تقدير معلمة التمييز.

وفيما يتعلق بمتغير التفاعل بين حجم العينة والنموذج، فقد أظهرت النتائج وجود دلالة إحصائية مما يشير إلى أن تأثير حجم العينة يختلف باختلاف النموذج، هذا يعني أن بعض النماذج مثل نموذج (PCM) قد تكون أقل حساسية لحجم العينة مقارنةً بنموذج (GRM) والذي قد يستفيد أكثر من زيادة حجم العينة.

وللكشف عن دلالة الفروق تم اجراء اختبار Bonferroni للمقارنات البعدية كما يوضح الجدول (27)

جدول 27

نتائج اختبار Bonferroni للمقارنات البعدية لمتوسط الأخطاء المعيارية والتي تعزى لحجم العينة ونوع النموذج المستخدم في تقدير معلمة القدرة

المستقل	المقارنة		الفرق بين المتوسطات	الخطأ المعياري	قيمة p	لصالح
حجم العينة	مقابل 100500		0.238	0.057	0.000**	500
	مقابل 1001000		0.26	0.068	0.000**	1000
	مقابل 5001000		0.022	0.047	0.543	غير دال
نوع النموذج	PCM مقابل GRM		0.176	0.061	0.005**	GRM
	NRM مقابل GRM		0.083	0.055	0.124	غير دال
	NRM مقابل PCM		0.093	0.058	0.091	غير دال
التفاعل	PCM مقابل GRM	مفحوص100	0.248	0.082	0.003**	GRM
	GRM مقابل NRM		0.098	0.076	0.201	غير دال
	PCM مقابل NRM		0.15	0.079	0.058	غير دال
	PCM مقابل GRM	مفحوص500	0.175	0.065	0.007**	GRM
	GRM مقابل NRM		0.076	0.061	0.219	غير دال
	PCM مقابل NRM		0.099	0.063	0.11	غير دال
	PCM مقابل GRM	مفحوص1000	0.113	0.056	0.045*	GRM
	GRM مقابل NRM		0.01	0.048	0.831	غير دال
	PCM مقابل NRM		0.103	0.052	0.052	غير دال

أظهرت النتائج المعروضة في جدول (27) وجود فروق دالة إحصائيًا تعزى لحجم العينة، حيث كلما زاد حجم العينة، انخفض المتوسط الحسابي للأخطاء المعيارية، مما يشير إلى دقة تقدير أعلى في العينات الأكبر. كما بينت النتائج أن هناك فروقًا تعزى لنوع النموذج، إذ كان نموذج (GRM) الأكثر دقة بأقل قيم للأخطاء المعيارية، يليه نموذج (NRM)، ثم نموذج (PCM). وبالنسبة للتفاعل بين حجم العينة والنموذج، فقد تبين أنه عند العينة الصغيرة (100 مفحوص)، تفوق نموذج (GRM) على كل من (NRM) و(PCM)، بينما عند العينتين (500 و1000 مفحوص)، لم يكن هناك تفاعل قوي بين حجم العينة والنموذج من حيث تأثير الخطأ المعياري، لكن بقي نموذج (GRM) الأكثر دقة في تقدير معلمة القدرة.

وتتفق هذه النتيجة مع دراسة (Custer, 2015) التي أفادت بأن فروق التقدير بين النماذج قد تكون غير دالة عند استخدام عينات كبيرة، مما يشير إلى أن حجم العينة يلعب دورًا أكبر في تعزيز دقة التقدير في جميع النماذج. كما أن هذه النتيجة تتفق مع ما أشار إليه (الشريفين، 2017) في دراسته حول قدرة النماذج المختلفة على تقدير معلمة القدرة عند زيادة حجم العينة، وهو ما يعزز من صحة النتائج التي تم التوصل إليها في الدراسة الحالية.

المراجع:

- القيسي، حسين. (2016). أثر حجم العينة وطول الاختبار على دقة تقدير معالم

الفقرة والقدرة باستخدام نماذج نظرية الاستجابة للفقرة اللامعلمية. مؤتة للبحوث والدراسات، سلسلة العلوم الإنسانية والاجتماعية، 31(5)، 203.

- القضاة، محمد. (2020). أثر حجم العينة في دقة تقدير قدرة الفرد ودالة معلومات الاختبار. مجلة كلية التربية بالمنصورة، 110(5)، 1605-1632. https://doi.org/10.21608/maed.2020.158026
- الشريفين، نضال. (2017). بناء مقياس المهارات الحياتية لدى طلبة الجامعات وفق نموذج مقياس التقدير. دراسات - العلوم التربوية، 44، 185 - 212.
- الثوابية، أحمد محمود (2010). أثر حجم العينة على تقدير صعوبة الفقرة والخطأ المعياري في تقديرها باستخدام نظرية الاستجابة للفقرة. مجلة جامعة دمشق، 26، 525-556.
- الشريفين، نضال كمال (2019). أثر نموذج نظرية الاستجابة للفقرة ذات الاستجابة المتعددة التدريج على دقة تقدير القدرات للأفراد والخصائص السيكومترية للفقرات والاختبار. المجلة التربوية، 130 (1)، 241-295.

- Baker, F. B. (2001). The basics of item response theory (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation.
- Brzezińska, J. (2016). Polytomous item response theory models using R. Econometrics, 20(2), 1-10. https://doi.org/10.15611/ekt.2016.2.04
- Custer, M. (2015, October). Sample size and item parameter estimation precision when utilizing the one-parameter "Rasch" model [Paper presentation]. Annual Meeting of the Mid-Western Educational Research Association, Evanston, IL.
- Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Kluwer Nijhoff Publishing.
- Kim, S., Cohen, A. S., & Lin, Y. (2005). LDID: A computer program for local dependence indices for dichotomous items (Version 1.0) [Computer software].
- Livingston, S. A. (2020). Basic concepts of item response theory: A nonmathematical introduction (Research Memorandum No. RM-2006-). Educational Testing Service. https://www.ets.org/research/policy-research-reports/publications/white-paper/2020/mfws
- Mahmud, J. (2017). Item response theory: A basic concept. Educational Research and Reviews, 12(5), 258-266. https://doi.org/10.5897/ERR2017.3147

- Nguyen, T., Han, H.-R., Kim, M., & Chan, K. (2014). An introduction to item response theory for patient-reported outcome measurement. Patient, 7(1), 23-35. https://doi.org/10.1007/s402710-0041-013-
- Reckase, M. D. (1979). Unifactor trait models applied to multifactor tests: Results and application. Journal of Educational Statistics, 4(3), 207-230.
- Shen, L. (1997, March 24-28). Quantifying item dependency by Fisher's Z [Paper presentation]. Annual Meeting of the American Educational Research Association, Chicago, IL.
- Stone, M., & Yumoto, F. (2004). The effect of sample size for estimating Rasch/IRT parameters with dichotomous items. Journal of Applied Measurement, 5(1), 48-61.
- Suwarto, E. P. W., & Setiawan, B. (2020, September 28-30). The effects of sample size and logistic models on item parameter estimation. In Proceedings of the ICE 2019, Purwokerto, Indonesia (pp. 1-9). EAI. https://doi.org/10.4108/eai.28-9-2019.2291082
- Swaminathan, H., Hambleton, R. K., Sireci, S. G., Xing, D., & Rizavi, S. M. (2003). Small sample estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates. Applied Psychological Measurement, 27(1), 27-51.
- Zanon, C., Claudio, S., Hanwook, H., & Ronald, H. (2016). An application of item response theory to psychological test development. Psicologia: Reflexão e Crítica, 29(1), 18-29. https://doi.org/10.1186/s411550040--016-x