Jordan Journal of Applied Science-Humanities Series

Volume 31
Issue 2 *31-2*

Article 4

2022

The accuracy of estimating the item difficulty parameter of the one parameter logistic model using the pairwise method in light of sample size and test length change

Mohamed Ahmad Ali Al-Yassin
Yarmouk University, Irbid, Jordan., maha_jawa2001@yahoo.com

VOL 31 NO 2 ISSN: 2708-9126

Jordan Journal of Applied Science

Humanities Series

Applied science Private University

The accuracy of estimating the item difficulty parameter of the one parameter logistic model using the pairwise method in light of sample size and test length change.

دقة تقدير معلمة صعوبة الفقرة للنموذج اللوجستي أحادي المعلمة باستخدام طريقة المزاوجة في ضوء تغير حجم العينة وطول الاختبار

Mohamed Ahmad Ali Al-Yassin[1*], Amal Ahmad Mahmoud Al Zoubi.[2]
Yarmouk University, Irbid, Jordan.[12]

ARTICLE INFO

Article history:
Received 31 Jan 2021
Accepted 15 Mar 2021
Published 01 Apr 2022

https://doi.org/10.35192/jjoas-h.v31i2.314

***Corresponding author at** Yarmouk University, Jordan.

Mohamed Ahmad Ali Al-Yassin.

Email: maha_jawa2001@yahoo.com.

Keywords:

Pairwise method

Item response theory

Test length

Sample Size

الكلمات المفتاحية:

طريقة المزاوجة

نظرية استجابة الفقرة

طول الاختبار

حجم العينة

ABSTRACT

This study aimed at investigating The accuracy of estimating the item difficulty parameter of the one parameter logistic model using the pairwise method in light of sample size and test length change. To achieve the study objective, the data of binary responses were generated by five different sample size (50, 100, 250, 500, 1000) examinee, and four different length test (20, 30, 40, 70) item using simra function in the pairwise package in R program by determining the level of the ability and difficulty parameter is between (-3, 3) normally distributed with a mean 0 and a standard deviation 1. To answer the questions of the study, the researcher used R program for estimating the difficulty parameter using the pairwise method, and estimating the ability using the weighted likelihood estimation method. The results of the study revealed statistically significant differences at $\alpha = 0.05$ between the standard errors of the estimates of the difficulty parameter due to the sample size for sample size 1000 examinee. The results of the study revealed statistically significant differences at $\alpha = 0.05$ between the standard errors of the estimates of the ability parameter due to the sample size and test length and interaction. The study recommended conduct studies which compare between the likelihood method, the Bayes method, and the pairwise method in the accuracy of estimating the parameters of the item.

هدفت الدراسة إلى الكشف عن دقة تقدير معلمة صعوبة الفقرة للنموذج أحادي المعلمة باستخدام طريقة المزاوجة في ضوء تغير حجم العينة وطول الاختبار. ولتحقيق هدف الدراسة، تم توليد بيانات ثنائية الاستجابة لأربعة مستويات من طول الاختبار (٢٠، ٣٠، ٤٠، ٧٠) فقرة، وخمسة مستويات من حجم العينة (٥٠، ١٠٠، ٢٥٠، ٥٠٠، ١٠٠٠) مفحوص باستخدام دالة simra الموجودة في حزمة المزاوجة، وذلك بتحديد مستوى قدرة المفحوصين وصعوبة الفقرات بحيث تكون بين (-٣، ٣) موزعة توزيعاً طبيعياً بوسط حسابي ٠ وانحراف معياري ١. استخدمت برمجية لغة R (Version R3.6.3)؛ لتقدير معلمة الصعوبة للفقرات باستخدام طريقة المزاوجة، وتقدير القدرة باستخدام طريقة الأرجحية الموزونة. وأظهرت نتائج الدراسة وجود فروق ذات دلالة احصائية عند مستوى الدلالة $\alpha = 0.05$ بين الأخطاء المعيارية لتقديرات معلمة الصعوبة تعزى لمتغير حجم العينة ولصالح حجم العينة ١٠٠٠ مفحوص. وأشارت النتائج إلى وجود فروق ذات دلالة احصائية عند مستوى الدلالة $\alpha = 0.05$ بين الأخطاء المعيارية لتقديرات معلمة القدرة تعزى لمتغير حجم العينة وطول الاختبار والتفاعل بينهما. وأوصت الدراسة بإجراء دراسات تقارن بها بين طريقة الأرجحية العظمى وطريقة بييز وطريقة المزاوجة في دقة تقدير معالم الفقرة.

المقدمة

يعتبر حجم عينة المفحوصين وطول الاختبار المطلوبان لتقديرات معالم نماذج نظرية استجابة الفقرة (Item response theory, [IRT]) عاملان مهمان؛ لأنهما يؤثران على دقة وكفاءة تقديرات معالم الفقرة والقدرة، ويعد التقدير الدقيق للمعالم هو أمر بالغ الأهمية؛ بسبب استخدام نتائج التقديرات لاتخاذ قرارات بشأن المفحوصين في العديد من المجالات التي يمكن أن تؤثر بشكل كبير في حياتهم كالتقدم في مراحل المدرسة والالتحاق بالجامعة والتوظيف.وتعد الموضوعية في القياس من الأمور الأساسية التي تهتم بها العلوم الطبيعية والعلوم الإنسانية على حد سواء، وكلما كان القياس موضوعياً زادت الدقة في فهم ظاهرة موضوع القياس والتنبؤ بها. ويقصد بالموضوعية في القياس النفسي والتربوي ثبات قياس الظاهرة السلوكية باختلاف الأداة المستخدمة في القياس، وألا يتأثر باختلاف المفحوصين الذين يتم تقدير السمة لديهم من خلال هذه الأداة، وأن يتم تدريج الأداة بوحدة قياس تتوافق مع تدريج مستويات الظاهرة السلوكية موضوع القياس (علام، ٢٠٠٥).

ولتلافي العيوب التي ظهرت في النظرية التقليدية في القياس كعدم الموضوعية في القياس، وأنها تقوم على افتراض أن تباين أخطاء القياس هو نفسه لمستويات القدرة جميعها، ومن أجل الوصول إلى قياس أكثر موضوعية، حاول علماء القياس الاستفادة من التقدم التكنولوجي للتوصل إلى طرق سيكومترية جديدة تساعد في التغلب على جوانب القصور في النظرية التقليدية، وتتمثل نظرية استجابة الفقرة (IRT) الاتجاه المعاصر في القياس النفسي والتربوي (Hambleton & Swaminathan, 1985). وتستند نظرية IRT على افتراضات أساسية هي: افتراض أحادي البعد، وافتراض الاستقلال الموضعي، وافتراض منحنى خاصية الفقرة، وافتراض التحرر من السرعة، وانبثقت عن نظرية IRT مجموعة من النماذج التي تعرف بنماذج السمات الكامنة، وتعد النماذج أحادية البعد الأكثر شيوعاً في تصميم الاختبارات والمقاييس التربوية والنفسية وبنائها، وهي الأكثر ملاءمة للفقرات ثنائية الاستجابة، ويعد عدد المعالم الذي توصف الفقرة به الفارق الأساسي بين نماذج السمات الكامنة أحادية البعد (Hambleton & Swaminathan, 1985).

ولقد زاد الاهتمام باستخدام نماذج نظرية IRT لتقدير معالم الفقرات وقدرات المفحوصين، لامتلاكها العديد من الميزات كإعطاء تقديرات دقيقة لمعالم الفقرة والقدرة، ولكنها تحتوي على بعض العيوب؛ إذ تتطلب استخدام حجوم عينات كبيرة للحصول على دقة لتقديرات المعالم (Lord, 1968)، حيث أن حجوم العينات الكبيرة غالبا غير متوفرة في البيئات التعليمية، لذلك فإن المهتمين بالقياس بحثوا عن طرق جديدة وبرمجيات أكثر تطوراً لاستخدامها في تقدير المعالم وفق نظرية IRT بوجود عينات صغيرة، ومنها طريقة المزاوجة Pairwise Method من خلال برمجية لغة R(Heine & Tarnai, 2015). وتعد النماذج ثنائية الاستجابة من أشهر النماذج أحادية البعد استخداماً في بناء الاختبارات والمقاييس، كالنموذج اللوجستي أحادي أو ثنائي أو ثلاثي المعلمة (de Gruijter & Van der Kamp, ٢٠٠٥)، ويفترض النموذج أحادي المعلمة أن لكل فقرة معلمة صعوبة خاصة بها، بينما لجميع الفقرات القدرة التمييزية نفسها بين المفحوصين، كما يفترض عدم لجوء المفحوصين للتخمين عند الإجابة على الفقرات (Hambleton, 1989).

ويمكن وصف هذا النموذج بالمعادلة الرياضية ١ الآتية:

$$P_i(\theta) = \frac{1}{1+e^{-Da(\theta - b_i)}} \quad ; i= ١, ٢, ٣,..., n \ (1)$$

حيث Pi(): احتمال الاستجابة الصحيحة للمفحوص ذو القدرة ()على الفقرة (i)، bi: معامل الصعوبة للفقرة (i)، : معلمة القدرة للمفحوص، a: معامل تمييز الفقرة وهو ثابت لجميع الفقرات في هذا النموذج، D: عامل التدريج وهو ثابت قيمته ١,٧.

ويعتمد احتمال إجابة المفحوص على الفقرة إجابة صحيحة في النموذج أحادي المعلمة على قدرة المفحوص ومعلمة الصعوبة، ولأن قيمة احتمال إجابة المفحوص على الفقرة غير معلومة، و إجابات المفحوصين على الفقرات معلومة، فإن أساليب تقدير المعالم تهدف الى تحديد قيمة القدرة لكل مفحوص وقيمة معلمة الصعوبة للفقرة من خلال الاعتماد على إجابات المفحوصين على الفقرات. وهنالك عدة عوامل تؤثر في دقة تقدير معالم الفقرات تأثيراً مباشراً مثل: طرق التقدير، وحجم العينة للمفحوصين، وطول الاختبار، وتوزيع السمة الكامنة، وجودة البيانات، والتوزيعات لمعالم الفقرة (Barnes & Wise, 1991). وأشار هامبلتون وسوامنيثان (Hambleton & Swaminathan, 1985) أن هناك ثلاثة طرق رئيسة لتقدير معالم الفقرة والقدرة وفق نماذج نظرية IRT تستخدم طريقة الأرجحية العظمى وهي: الأرجحية العظمى المشتركة Joint Maximum Likelihood (JML) والأرجحية العظمى الهامشية Marginal Maximum Likelihood (MML) والأرجحية العظمى المشروطة Conditional Maximum Likelihood (CML)، وأشار لورد (Lord, 1986) إلى أن طريقة بييز (Bayes) تعد من الأساليب الإحصائية المستخدمة لتقدير معالم الفقرات والقدرة.

ويؤدي التقدير المتزامن لكل من معالم الفقرة وقدرة المفحوص في نماذج نظرية IRT إلى تعقيدات إحصائية في التقدير، مما جعل إجراء التقدير محط تركيز أساسي لأبحاث القياس النفسي (Birnbaum, 1969). وأشار وورم (Warm, ١٩٨٩) إلى طريقة تقدير الأرجحية الموزونة Weighted Likelihood Estimation (WLE) التي تعتمد على توزين دالة الأرجحية لتقدير معلمة قدرة المفحوص للتقليل من التحيز في تقدير القدرة بطريقة الأرجحية العظمى، ولا تعتبر طريقة الأرجحية الموزونة من طرق تقدير بييز، لأنه لم يتم وضع افتراضات حول توزيع معلمة القدرة.

واقترح شوبن (Choppin, ١٩٨٥) تقنيتين يمكن استخدامهما لتقدير معالم الفقرات وفق النموذج أحادي المعلمة، فالتقنية الأولى هي الأرجحية العظمى التي تعتمد على خوارزمية تكرارية مثل طريقة نيوتن رافسون، أما التقنية الثانية هي طريقة المزاوجة التي لا تعتمد على خوارزمية تكرارية. ويعتبر نموذج ثيرستون للمقارنات الزوجية أساساً لطريقة المزاوجة، والذي أشار إلى هذه الطريقة هو العالم راش (Rasch, 1960)، وطورها العالم شوبن (Choppin, 1985) لتصبح طريقة عملية لمعايرة الفقرات داخل بنوك الأسئلة، واستخدامها لتقدير معلمة صعوبة الفقرات. إن المبدأ الأساسي في تطبيق المقارنات الزوجية هو الحكم على العديد من الأشياء (مثل الفقرات) على شكل أزواج وتحديد أي من الزوجين مفضلين من قبل المفحوص، ثم الحصول على مقياس متري لمجموعة من الفقرات، ومن مميزات طريقة المزاوجة: تجنب العمليات الحسابية الطويلة في تحليل البيانات الكبيرة، ومقارنة فقرتين فقط في كل مرة، والتعامل مع مصفوفات البيانات غير المكتملة دون أي متاعب بناءً على أسلوبها الرقمي، وتقدير معالم النموذج المستخدم بوجود بيانات مفقودة، وتعتبر دقتها في تقدير معلمة الصعوبة أفضل من دقة التقدير باستخدام طريقة الأرجحية العظمى عند استخدامها مع العينات الصغيرة (Heine & Tarnai, 2015).

ويمكن اشتقاق طريقة المزاوجة من خلال معادلة النموذج اللوجستي الذي صاغها راش (Rasch, 1960) كما في معادلة ٢:

$$P(x_{vi}) = \frac{e^{x_{iv}(\theta_v - b_i)}}{1 + e^{(\theta_v - b_i)}} \quad (2)$$

حيث x_{vi} : استجابة المفحوص v على الفقرة i (٠، ١)، θv : قدرة المفحوص v، b_i : صعوبة الفقرة i.

وعلى فرض أن عينة من المفحوصين (v=1....n)، يجب أن تجيب على فقرتين (i,j)، وبتحقق افتراض الاستقلالية العشوائية المشروطة للفقرتين المقدمتين للمفحوص وفق نموذج راش، فإن الإجابات الأربعة المتوقعة على الفقرتين ستكون إما (٠،١،٢)، وتعطي احتمالات الاجابة لها كالآتي:

$$P(x_{vi} = 0, x_{vj} = 0) = \frac{1}{1 + e^{(\theta_v - b_i)}} \times \frac{1}{1 + e^{(\theta_v - b_j)}} \quad (3)$$

$$P(x_{vi} = 1, x_{vj} = 0) = \frac{e^{(\theta_v - b_i)}}{1 + e^{(\theta_v - b_i)}} \times \frac{1}{1 + e^{(\theta_v - b_j)}} \quad (4)$$

$$P(x_{vi} = 0, x_{vj} = 1) = \frac{1}{1 + e^{(\theta_v - b_i)}} \times \frac{e^{(\theta_v - b_j)}}{1 + e^{(\theta_v - b_j)}} \quad (5)$$

$$P(x_{vi} = 1, x_{vj} = 1) = \frac{e^{(\theta_v - b_i)}}{1 + e^{(\theta_v - b_i)}} \times \frac{e^{(\theta_v - b_j)}}{1 + e^{(\theta_v - b_j)}} \quad (6)$$

حيث تشير معادلة ٣ إلى أن المفحوص أجاب إجابة خاطئة عن كلا الفقرتين، ومعادلة ٤ تشير إلى أن المفحوص أجاب إجابة صحيحة عن الفقرة i وإجابة خاطئة عن الفقرة j، كما أن معادلة ٥ تشير إلى أن المفحوص أجاب إجابة خاطئة عن الفقرة i وإجابة صحيحة عن الفقرة j، وتشير معادلة ٦ إلى أن المفحوص أجاب إجابة صحيحة عن كلا الفقرتين.

ولإيجاد الاحتمال المشترك لفقرتين الذي يحقق الدرجة ١ من خلال جمع المعادلتين (٤، ٥) التي تعطى بمعادلة ٧:

$$P(x_{vi} + x_{vj} = 1) = \frac{e^{(\theta_v - b_i)}}{\left(\left(1 + e^{(\theta_v - b_i)}\right) \times \left(1 + e^{(\theta_v - b_j)}\right)\right)} + \frac{e^{(\theta_v - b_j)}}{\left(\left(1 + e^{(\theta_v - b_i)}\right) \times \left(1 + e^{(\theta_v - b_j)}\right)\right)} \quad (7)$$

وللحصول على الاحتمالية المشروطة بالدرجة ١ على الفقرة i، من خلال قسمة معادلة ٤ على المعادلة ٧ وتعطى بمعادلة ٨:

$$P(x_{vi} = 1 | x_{vi} + x_{vj} = 1) = \frac{\frac{e^{(\theta_v - b_i)}}{1 + e^{(\theta_v - b_i)}} \times \frac{1}{1 + e^{(\theta_v - b_j)}}}{\frac{e^{(\theta_v - b_i)}}{\left(\left(1 + e^{(\theta_v - b_i)_i}\right) \times \left(1 + e^{(\theta_v - b_j)}\right)\right)} + \frac{e^{(\theta_v - b_i)}}{\left(\left(1 + e^{(\theta_v - b_i)}\right) \times \left(1 + e^{(\theta_v - b_j)}\right)\right)}} \quad (8)$$

ويمكن تبسيط معادلة ٨ رياضيا، وإلغاء معلمة القدرة كما في معادلة ٩:

$$P(x_{vi}=1 | x_{vi}+x_{vj}=1)=\frac{e^{b_i}}{e^{b_i}+e^{b_j}} \quad (9)$$

وبالمثل يمكن صياغة الاحتمال المشروط بالدرجة ١ على الفقرة j، ويمكن تقدير الاحتمال المشروط بإجابة المفحوص إجابة صحيحة على الفقرة i وإجابة خاطئة على الفقرة j في معادلة ٩ من خلال بيانات حقيقية، ويمكن التعبير عنها من خلال معادلة ١٠:

$$\frac{f_{i,j}}{n_{i,j}}=\frac{e^{b_i}}{e^{b_i}+e^{b_j}} \quad (10)$$

حيث أن: $f_{i,j}$ عدد المفحوصين الذين أجابوا إجابة صحيحة على الفقرة i وإجابة خاطئة على الفقرة j، $f_{j,i}$: عدد المفحوصين الذين أجابوا إجابة صحيحة على الفقرة j وإجابة خاطئة على الفقرة i، $n_{i,j}$: ($n_{i,j}=f_{i,j}+f_{j,i}$) عدد المفحوصين الذين حققوا مجموع الدرجة ١ على الفقرتين ويمكن اعتبارها حجم عينة المفحوصين n، وبالمثل يمكن إعادة صياغة الاحتمال المشروط بالدرجة ١ على الفقرة j.

وبالتالي يمكن تقدير صعوبة الفقرة من خلال الاحتمالية المشروطة، إذ يمكن إعادة كتابة معادلة ٩ كما في معادلة ١١:

$$\frac{f_{i,j}}{f_{j,i}}=\frac{e^{b_i}}{e^{b_j}} \quad (11)$$

وبأخذ اللوغاريتم الطبيعي للمعادلة السابقة لينتج نسب لوغاريتم مرجح النجاح Odds كما في معادلة ١٢:

$$\ln\left(\frac{f_{i,j}}{f_{j,i}}\right)=b_i-b_j \quad (12)$$

وبالمثل يمكن كتابة المعادلة للفقرة j كما يأتي (Heine & Tarnai, 2015):

$$\ln\left(\frac{f_{j,i}}{f_{i,j}}\right)=b_j-b_i \quad (13)$$

وبعد عملية تقدير معالم الفقرات وقدرات المفحوصين، فإنه يجب التحقق من مدى دقة التقدير، فقد أشار لورد (Lord, 1980) إلى وجود معايير مختلفة للكشف عن دقة تقدير المعالم ومن أبرزها: الخطأ المعياري في التقدير، ودالة المعلومات للفقرات والاختبار، والفاعلية النسبية للاختبار. يعد الخطأ المعياري في التقدير مؤشراً إحصائياً جيداً على دقة تقدير معلمة القدرة ومعالم الفقرة، ويرتبط الخطأ المعياري في التقدير عكسياً مع الجذر التربيعي لدالة معلومات الاختبار، ويكون الخطأ المعياري في التقدير أقل ما يمكن عند مستويات القدرة التي تناظر اقصى معلومات للاختبار (Hambleton et al., 1991). ويشير وورم (Warm, 1978) إلى أن الخطأ المعياري في التقدير هو القيمة المتوقعة للانحراف المعياري لأخطاء التقدير، وكلما قلت قيمته كان مؤشراً على دقة التقدير. وكلما قل مقدار الخطأ المعياري في التقدير، فإن ذلك يؤدي إلى زيادة الثبات وبالتالي زيادة دقة الاختبار (Reeve & Fayers, 2004). طور العالم ايفرون (Efron, 1979) طريقة لاختيار عينات بالإرجاع سميت بطريقة البوتستراب (Bootstrap) وهي اختيار مجموعة من عينات عشوائية التي تشتمل على (n) من العناصر المسحوبة بالإرجاع بشكل عشوائي من (N) من البيانات الأصلية، وتقوم هذه الطريقة في تقدير التباينات والأخطاء المعيارية وفترات الثقة والقيمة الاحتمالية. ويمكن تقدير الأخطاء المعيارية لمعالم الفقرات باستخدام طريقة البوتستراب عند استخدام طريقة المزاوجة في تقدير معالم الفقرات، إذ أنها لا تستند إلى دالة الأرجحية العظمى في تقدير معالم الفقرات (Finch & French, 2019). إن عدد عينات البوتستراب التي يجب سحبها تختلف باختلاف الغاية التي استخدمت من أجلها تقنية البوتستراب، مثلاً قد تبلغ عدد عينات البوتستراب ١٠٠٠ عينة. ودرس كلاً من ايمون وماسوم (Imon & Masoom, 2005) مسألة تحديد عدد عينات البوتستراب المطلوب سحبها من المجتمع، وأشارا إلى عدم وجود اتفاق بين الباحثين الإحصائيين حول عدد التكرارات الكافية التي تحقق الغاية من عملية تقدير المعالم، ويمكن تقدير الخطأ المعياري في تقدير معلمة الصعوبة كما في المعادلة الآتية:

$$S(\theta_b)=\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\theta_b-\theta_B)^2} \quad (14)$$

حيث $S(\theta_b)$: الخطأ المعياري لتقدير صعوبة الفقرة، θ_b : معلمة صعوبة الفقرة الحقيقية، θ_B : معلمة صعوبة الفقرة المقدرة، B: عدد عينات البوتستراب (Heine & Tarnai, ٢٠١٥).

الدراسات السابقة

يتناول هذا الجزء مراجعة للدراسات ذات الصلة بمشكلة البحث، وقد تم عرض هذه الدراسات وفق التسلسل الزمني ، فمن الدراسات الأجنبية ذات الصلة بمشكلة الدراسة، دراسة هيونغ وآخرون (Huang et al., 2001) التي هدفت إلى مقارنة فعالية بعض البرمجيات الجاهزة (Bilog- MG٣, PIC) في معايرة فقرات الاختبارات المخصصة لاختيار المفحوصين لمهن معينة والتي تتضمن فقراتها أجزاء لحقل الاختصاص وأخرى مشتركة، وتركز هذه الدراسة في جانب منها على فحص أثر حجم عينة المفحوصين على دقة تقدير معالم الفقرات، ولتحقيق هذا الغرض تم استخدام النموذج اللوجستي ثلاثي المعلمة لمعايرة ٣٦٠ فقرة باستخدام طريقة المحاكاة حيث تم افتراض أن أول ١٢٠ فقرة تشكل الجزء المشترك من الاختبار بينما باقي الفقرات شكلت أربعة اختبارات يتضمن كل منها ٦٠ فقرة، وقد تم اختيار حجم العينة ما بين ٢٥٠ مفحوص و ١٥٠٠ مفحوص، واستخدم معيار جذر معدل مربعات الفروق (RMSD) لمقارنة دقة تقديرات المعالم. أظهرت النتائج أن أخطاء التقدير لمعلمة الصعوبة ومعلمة التمييز تكون أكبر عندما يقل حجم عينة المفحوصين وذلك عند استخدام برمجية Bilog-MG3.

وأجرى الدرابيع (٢٠٠١) دراسة هدفت إلى البحث في فعالية النموذج اللوجستي أحادي المعلمة في دقة تقدير قدرة المفحوص، وصعوبة الفقرة باختلاف حجم عينة المفحوصين والاختبار. تكونت عينة المفحوصين في الدراسة من (٥٠، ١٠٠، ٥٠٠) مفحوص، وعدد فقرات الاختبار (٢٥، ٥٠، ٣٠٠) فقرة. أشارت نتائج الدراسة إلى وجود فروق ذات دلالة إحصائية لتفاعل كل من حجم عينة المفحوصين وطول الاختبار على دقة تقدير قدرة المفحوص. في حين لم تظهر النتائج وجود فروق إحصائية في دقة تقدير معلمة القدرة تعزى لحجم عينة المفحوصين، ووجود فروق إحصائية في دقة تقدير معلمة القدرة تعزى لطول الاختبار لصالح طول الاختبار الكبير، أما فيما يتعلق بدقة تقدير معلمة الصعوبة، فقد بينت النتائج بأن هناك فروقا إحصائية تعزى لتفاعل كل من حجم العينة وطول الاختبار، وفروق إحصائية في دقة تقدير معلمة الصعوبة تعزى لحجم عينة المفحوصين وطول الاختبار لصالح العينات الكبيرة للمفحوصين والاختبار.وفي دراسة عبابنة (٢٠٠٤) التي هدفت إلى دراسة أثر حجم العينة وطريقة انتقائها وعدد الفقرات وطريقة انتقائها على دقة تقدير معالم الفقرة والقدرة وفق النموذج اللوجستي ثلاثي المعلمة. اختار اختبار القدرة العقلية المؤلف من أربع اختبارات فرعية هي: اختبار المفردات، واختبار المتشابهات، واختبار المتضادات، واختبار الحساب، وطبقت ثلاث اختبارات بأطوال مختلفة (٢٠، ٣٠، ٧١) فقرة على ثلاثة عينات من الفحوصين (٢٥٠، ٥٠٠، ١٠٠٠) مفحوص، وقد استخدمت برمجية Bilog لتحليل البيانات وتقدير معالم فقرات الاختبار وقدرات المفحوصين. أشارت النتائج إلى أن دقة تقديرات معالم الفقرة تزداد بزيادة حجم عينة المفحوصين، وأشارت النتائج إلى أن دقة تقديرات معلمة القدرة تزداد بزيادة طول الاختبار.

وأجرى فيتزباترك (Fitzpatrick, 2008) دراسة هدفت إلى معرفة أثر تخفيض طول اختبارات التهيئة على معدل إتقان الطلبة. تكونت الاختبارات في الدراسة من حجوم مختلفة (٥، ١٠، ١٥) فقرة طبقت على عينات مختلفة من المفحوصين تتراوح ما بين (٣٣٠٠) مفحوص إلى (٢٦٠٠٠) مفحوص، وتم تحليل البيانات وفق النموذج اللوجستي أحادي المعلمة. أظهرت نتائج الدراسة زيادة في تباينات الاختلافات في معدلات إتقان المفحوصين عند استخدام أطوال اختبارات قصيرة، كما أن هناك اختلافات وعدم استقرار في معلمة الصعوبة عند استخدام اختبار طوله أقل من ١٥ فقرة، وأوصت الدراسة باستخدام اختبارات طولها أكبر من ١٥ فقرة لزيادة الاستقرار في تقدير معالم الفقرات.وأجرى دي لوتوري وهونغ (De la Torre & Hong, 2010) دراسة هدفت إلى التعرف إلى أثر حجم العينة على دقة تقدير معالم الفقرة والقدرة في اختبارات مطورة حسب نماذج نظرية IRT. واستخدمت الدراسة نموذج higher order item response theory (HO - IRT) في توليد مجموعة من البيانات ضمن ظروف اختبار مختلفة. حيث تم توليد البيانات لحجوم عينات مكونة من ٥٠٠ مفحوص و١٠٠٠ مفحوص، وطول اختبارات مكونة من ١٠ فقرات و٢٠ فقرة باستخدام طريقة مونتي كارلو من أجل التعرف إلى أثر حجم العينة في دقة تقدير معالم الفقرة والقدرة في الاختبار. وأشارت النتائج إلى أن حجم العينة وطول الاختبار يؤثران في تقدير معالم الفقرة ولصالح العينة ١٠٠٠ مفحوص والاختبار ٢٠ فقرة، حيث كان الوسط الحسابي للخطأ المعياري في تقدير معلمة الصعوبة يتراوح بين ٠,٠٦ و٠,٠٢. وأشارت النتائج أيضاً إلى عدم تأثر دقة تقدير معلمة القدرة بحجم العينة وتأثرها بطول الاختبار، حيث كانت الأخطاء المعيارية أقل ما يمكن عند استخدام طول اختبار ٢٠ فقرة مقارنة مع طول الاختبار ١٠ فقرات، وكان الوسط الحسابي للخطأ المعياري في تقدير معلمة القدرة يتراوح بين ٠,٥٨ و٠,٤٨.

وفي دراسة قام بها جين وآخرون (Chen et al., ٢٠١٤) هدفت إلى تقييم نتائج تحليل نموذج راش باستخدام حجوم عينات صغيرة، ولغرض الدراسة تم استخدام حجوم عينات مختلفة مكونة من (٣٠، ٥٠، ١٠٠، ٢٥٠) مفحوص واختبار مكون من ١٠ فقرات. واستخدم برنامج Mplus لتحليل البيانات. أشارت نتائج التحليل أنه عند استخدام عينات صغيرة (٣٠، ٥٠) مفحوص يؤدي إلى أخطاء معيارية في تقدير المعالم أكبر من نتائج التحليل عند استخدام عينات (١٠٠، ٢٥٠) مفحوص.وأجرى جيانغ وآخرون (Jiang et al., ٢٠١٦) دراسة هدفت إلى معرفة الحجم المناسب لتقدير معالم الفقرة وفق نموذج الاستجابة المتدرجة متعددة الأبعاد Multidimensional Graded Response Model، وتم استخدام طريقة الأرجحية العظمى الهامشية لتقدير المعالم باستخدام برمجية flexMIRT، ولأغراض الدراسة تم توليد بيانات بحجوم عينات مختلفة مكونة من (٥٠٠، ١٠٠٠، ١٥٠٠، ٢٠٠٠) مفحوص، وأطوال اختبارات مختلفة مكونة من (٣٠، ٩٠، ٢٤٠) فقرة. أشارت النتائج إلى أن أقل حجم عينة يقدم تقديرات دقيقة لمعالم الفقرة هو ٥٠٠ مفحوص عند استخدام اختبارات مكونة من ٣٠ فقرة و٩٠ فقرة، وعند استخدام اختبار طوله ٢٤٠ فقرة فمن الضروري استخدام عينة مكونة من ١٠٠٠ مفحوص على الأقل، كما وأشارت النتائج إلى أن زيادة حجم العينة أكبر من ١٠٠٠ مفحوص لا يزيد من دقة تقدير المعالم.

وفي دراسة أجراها ساهين وانيل (Sahin & Anil, 2017) هدفت إلى معرفة أثر حجم العينة وطول الاختبار على دقة تقدير معالم نماذج نظرية استجابة الفقرة، وتم استخدام طريقة الأرجحية العظمى الهامشية لتقدير معالم الفقرة. واستخدم برمجية Xcalibre 4.1 لتقدير المعالم. ولتحقيق هدف الدراسة تم بناء ثلاث اختبارات لغوية ذات أطوال مختلفة مكونة من (١٠، ٢٠، ٣٠) فقرة وطبقت على تسعة حجوم عينات مختلفة مكونة من (١٥٠، ٢٥٠، ٣٥٠، ٥٠٠، ٧٥٠، ١٠٠٠، ٢٠٠٠، ٣٠٠٠، ٥٠٠٠) مفحوص. أشارت النتائج إلى أنه يمكن استخدام حجم عينة مكونة من ١٥٠ مفحوص على الأقل مع اختبارات مكونة من (١٠، ٢٠، ٣٠) فقرة لتقدير معلمة الصعوبة بدقة وفق النموذج أحادي المعلمة، وأشارت النتائج إلى أن أوساط الأخطاء المعيارية في تقدير معلمة الصعوبة تتناقص بزيادة حجم العينة حيث تراوحت بين ٠,٣٣ و٠,٠١.

المعالجة الإحصائية

تم استخدام العديد من المعالجات الإحصائية للإجابة عن أسئلة الدراسة، وذلك وفق الخطوات التالية:

التحقق من افتراضات نموذج نظرية الاستجابة للفقرة أحادي المعلمة

أ- أحادية البعد (Unidimensionality):

تم التحقق من افتراض أحادية البعد للاختبارات المستخدمة في هذه الدراسة بعدة مؤشرات اعتمدت على التحليل العاملي للمكونات الأساسية (Principal Components)، ومنها: نسبة التباين المفسر للعامل الأول، ونسبة الفرق بين الجذر الكامن الأول والجذر الكامن الثاني إلى الفرق بين الجذر الكامن الثاني والجذر الكامن الثالث، ونسبة الجذر الكامن الأول إلى الجذر الكامن الثاني. ودراسة العلاقة بين الجذر الكامن الثاني وبقية الجذور التي تليه بالإضافة إلى رسم بياني للجذور الكامنة للعوامل المكونة للاختبار (Scree Plot) باستخدام برمجية SPSS. حيث تبين من تحليل البيانات تحقق افتراض أحادية البعد لجميع الاختبارات. ويبين جدول ١ قيم الجذور الكامنة ونسبة التباين المفسر للاختبار المكون من ٢٠ فقرة والذي طبق على عينات مختلفة مكونة من (٥٠، ١٠٠، ٢٥٠، ٥٠٠، ١٠٠٠) مفحوص.

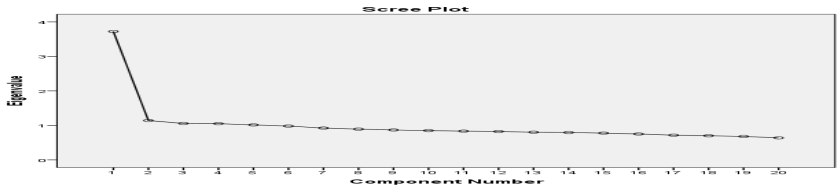
جدول ١ نتائج التحليل العاملي لفقرات الاختبار ٢٠ ولعينات الدراسة الخمسة

حجم العينة	العامل	قيمة الجذر الكامن	التباين المفسر %	الجذر الكامن الأول / الجذر الكامن الثاني	الجذر الكامن الأول – الجذر الكامن الثاني / الجذر الكامن الثاني – الجذر الكامن الثالث
١٠٠٠	الأول	٣,٨٥٦	٣٨,٣٦٢	٣,٤٧٧	٣٦,٦٢٧
	الثاني	١,١٠٩	٥,٥٤٧		
	الثالث	١,٠٣٤	٥,١٦٩		
	الرابع	١,٠١٤	٥,٠٧١		
٥٠٠	الأول	٣,٥٠٢	٢٦,٤٤٩	٢,٨٦١	٢٢,١١٧
	الثاني	١,٢٢٤	٦,١٢١		
	الثالث	١,١٢١	٥,٦٠٣		
	الرابع	١,٠٨٤	٥,٤١٩		
	الخامس	١,٠٦١	٥,٣٠٥		
	السادس	١,٠١	٥,٠٥٢		
	السابع	١,٠٠٥	٥,٠٢٦		
٢٥٠	الأول	٣,٨٨٩	٢٧,٠٧٧	٢,٨٩٤	٢٩,٥٩٣
	الثاني	١,٣٤٤	٦,٧٢٠		
	الثالث	١,٢٥٨	٦,٢٩١		
	الرابع	١,١٧٣	٥,٨٦٥		
	الخامس	١,٠٥	٥,٥٢٦		
	السادس	١,٠٠٥	٥,٢٧٣		
	السابع	١,٠١١	٥,٠٠٥		
١٠٠	الأول	٣,٨٢٤	٣١,٩١٤	٢,٤٧٤	٢١,٤٩١
	الثاني	١,٥٤٦	٧,٧٢٨		
	الثالث	١,٤٤٠	٧,١٩٨		
	الرابع	١,٣٣٦	٦,٦٧٩		
	الخامس	١,١٩٧	٥,٩٨٦		
	السادس	١,١٧٧	٥,٨٨٤		
	السابع	١,٠٢٨	٥,١٣٩		
٥٠	الأول	٤,٧٥٣	٢٩,٨٣٤	٢,٦١٠	١٦,٩٤٨
	الثاني	١,٨٢١	٩,١٠٣		
	الثالث	١,٦٤٨	٨,٢٤٠		
	الرابع	١,٤٣٦	٧,١٨٢		
	الخامس	١,٢٣٠	٦,١٥٢		
	السادس	١,٢٠٩	٦,٠٤٧		
	السابع	١,١٣٧	٥,٦٨٥		
	الثامن	١,٠١٨	٥,٠٨٩		

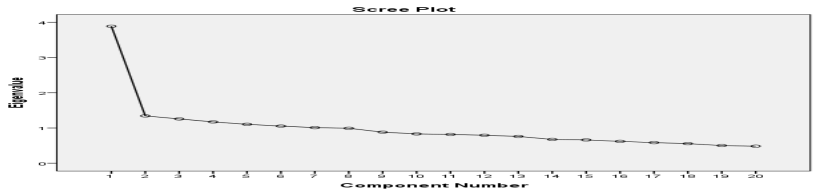
يتبين من جدول ١ أن قيم الجذر الكامن للعامل الأول لجميع حجوم العينات في الاختبار المكون من ٢٠ فقرة يفسر أعلى من ٢٠% من التباين الكلي، وهذا يطابق ما اقترحه ريكاس (Reckase, 1979) من أنه إذا فسر العامل الأول على الأقل نسبة ٢٠% من التباين المفسر فذلك يدل على تحقق افتراض أحادية البعد. وبلغت نسبة الجذر الكامن الأول إلى الجذر الكامن الثاني أكبر من ٢ لجميع حجوم العينات في الاختبار المكون من ٢٠ فقرة، وهذا يطابق ما اقترحه هاتي (Hattie, 1985) من أن النسبة المرتفعة للجذر الكامن الأول إلى الجذر الكامن الثاني يعد دليلاً على افتراض أحادية البعد، وحدد لينارس (Linacre, 2008) تلك النسبة بحيث تكون أكبر أو تساوي القيمة ٢ للدلالة على أحادية البعد. كما بلغت قيمة الفرق بين الجذرين الكامنين الأول والثاني إلى الفرق بين الجذرين الكامنين الثاني والثالث أعلى من ٦ لجميع حجوم العينات للاختبار المكون من ٢٠ فقرة، وهذا يطابق ما اقترحه هاتي (Hattie, 1985) بأن نسبة طرح الجذر الكامن الثاني من الجذر الكامن الأول إلى طرح الجذر الكامن الثالث من الجذر الكامن الثاني مرتفعة، وجميعها أعلى من القيمة (٦) بما يفيد من تحقق افتراض أحادي البعد لجميع اختبارات الدراسة في العينات الخمسة. ويتعزز افتراض أحادية البعد من خلال اختبار فحص العوامل الذي يظهر في شكل ١. وتبين الأشكال (أ، ب، ج، د، هـ) في شكل ١ مخططات القيم الفارزة للجذور الكامنة (Scree Plot) التي أظهرت أحادية البعد للاختبار المكون من ٢٠ فقرة والذي طبق على حجوم عينات الدراسة.

الشكل (١) : مخطط قيم الجذور الكامنة في الاختبار ٢٠ فقرة والمطبق على عينات الدراسة الخمسة.

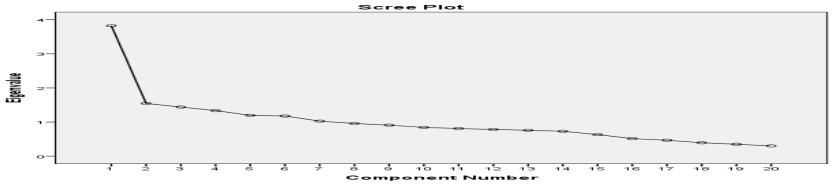
أ



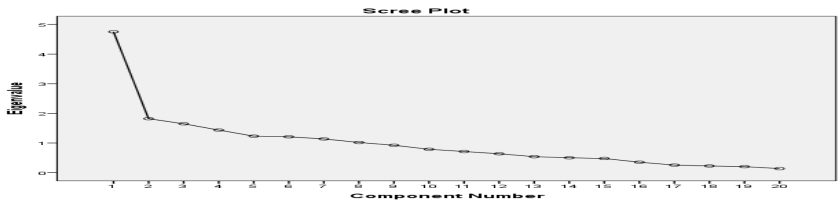
ب



ج



د



هـ

الاستقلال الموضعي (Local Independence):

تم استخدام حزمة sirt في برمجية لغة R لفحص الاستقلال الموضعي لفقرات الاختبارات والمفحوصين، من خلال مؤشر Q3 باستخدام حزمة المزاوجة (Pairwise Package)، التي طورها العالم (Heine & Tarnai, 2015)، وبينت نتائج التحليل تحقق افتراض الاستقلال الموضعي، ويبين جدول ٢ نتائج اختبار الاستقلال الموضعي لفقرات الاختبارات ولعينات الدراسة الخمسة.

جدول ٢ نتائج اختبار الاستقلال الموضعي لفقرات الاختبارات الأربعة ولعينات الدراسة الخمسة.

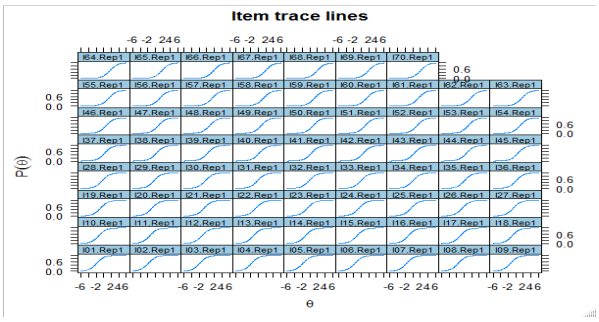
الوسط الحسابي لقيم Q3 للاختبار				
العينة	٢٠	٣٠	٤٠	٧٠
١٠٠٠	-0.029	-0.011	-0.005	0.007
٥٠٠	-0.019	-0.003	0.006	0.016
٢٥٠	-0.013	-0.007	0.016	0.029
١٠٠	0.024	0.025	0.045	0.054
٥٠	0.048	0.065	0.070	0.087

يلاحظ من جدول ٢ أن قيم الوسط الحسابي لـ Q3 لجميع الاختبارات الأربعة ولكل العينات الخمسة تراوحت ما بين (-٠,٠٢٩، ٠,٠٨٧)، وبينت نتائج التحليل في جدول ٢ أن قيم معامل Q3 لم تتجاوز مستوى انتهاك افتراض الاستقلال الموضعي الذي حدد ما بين (-٠,٢، ٠,٢) حسب ما اقترحه جين وثيسن (Chen & Thissen,1997).

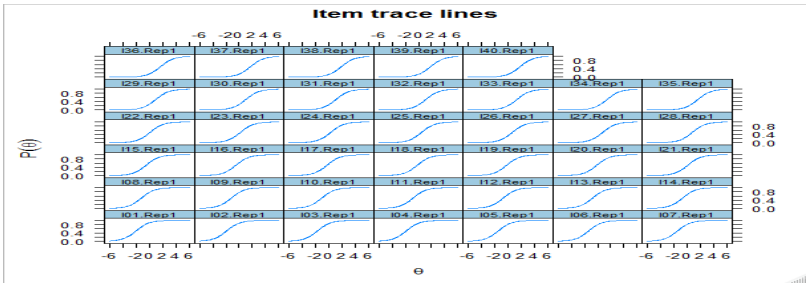
افتراض منحنى خصائص الفقرة (ICC):

للتحقق من افتراض اطراديه السمة الكامنة وفقاً للنموذج أحادي المعلمة في نظرية IRT، فقد تم رسم منحنى خصائص الفقرة لكل فقرة من فقرات الاختبارات الأربعة التي طبقت على عينات الدراسة الخمسة باستخدام حزمة mirt في برمجية لغة R، حيث أظهرت أشكال منحنيات خصائص الفقرة أن زيادة قدرة المفحوص يقابلها زيادة احتمالية الاستجابة الصحيحة للمفحوص على الفقرة في جميع فقرات الاختبارات الأربعة المطبقة على جميع عينات الدراسة الخمسة، وبينت أشكال منحنيات خصائص الفقرة أنها لا تختلف باختلاف طول الاختبار وحجم العينة، وتبين الأشكال (أ، ب، ج، د) في شكل ٢ منحنيات خصائص الفقرة لاختبارات الدراسة الأربعة عند حجم العينة ١٠٠٠ مفحوص .

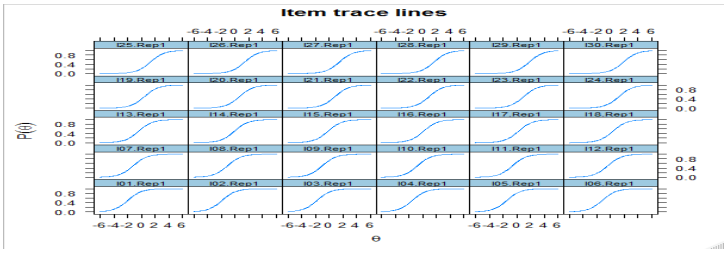
شكل (٢)منحنيات خصائص الفقرة لاختبارات الدراسة عند حجم العينة ١٠٠٠ مفحوص



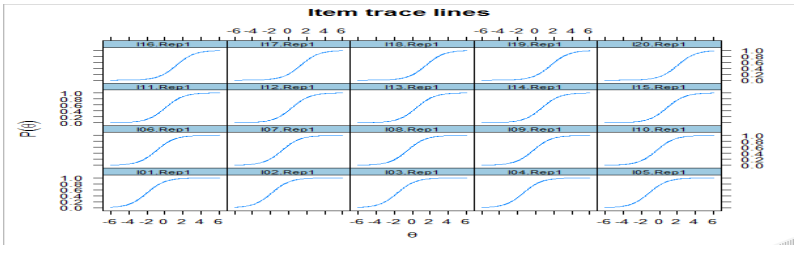
أ



ب



ج



د

ثبات الاختبارات في عينات الدراسة

تم التحقق من ثبات الاختبارات الأربعة التي طبقت على عينات الدراسة الخمسة من خلال حساب معامل أوميغا ومعامل الثبات التجريبي باستخدام برمجية لغة .ويبين جدول ٣ قيم معاملات الثبات لجميع الاختبارات في عينات الدراسة.

جدول ٣ قيم معامل ثبات أوميغا ومعامل الثبات التجريبي للاختبارات الأربعة في عينات الدراسة الخمسة

معامل ثبات أوميغا W و (معامل الثبات الامبريقي) للاختبار طول الاختبار				
حجم العينة	٢٠	٣٠	٤٠	٧٠
١٠٠٠	٠,٧٦ (٠,٨٠)*	٠,٨٢ (٠,٨٥)	٠,٨٨ (٠,٨٩)	٠,٩١ (٠,٩٢)
٥٠٠	٠,٨٠ (٠,٨١)	٠,٨٤ (٠,٨٦)	٠,٨٦ (٠,٨٧)	٠,٩٣ (٠,٩٣)
٢٥٠	٠,٧٣ (٠,٧٨)	٠,٨٥ (٠,٨٧)	٠,٨٦ (٠,٨٨)	٠,٩٣ (٠,٩٣)
١٠٠	٠,٨٠ (٠,٨٢)	٠,٧٦ (٠,٨٠)	٠,٩٠ (٠,٩١)	٠,٩١ (٠,٩٢)
٥٠	٠,٨٥ (٠,٨٥)	٠,٨٣ (٠,٨٣)	٠,٨٨ (٠,٨٩)	٠,٩٤ (٠,٩٤)

*: القيم داخل القوس تمثل معامل الثبات الامبريقي.

يتضح من جدول ٣ أن قيم معامل ثبات أوميغا قد تراوحت بين ٠,٧٣ و٠,٩٤، ويتبين أيضاً أن قيم معاملات الثبات التجريبي قد تراوحت بين ٠,٨٠ و٠,٩٤، وهذا يبين أن قيم معامل ثبات أوميغا والثبات التجريبي كانا مرتفعان لجميع اختبارات الدراسة والمطبقة على حجوم عينات مختلفة.

٢. تم تقدير معلمة صعوبة الفقرة بطريقة المزاوجة باستخدام حزمة المزاوجة (Pairwise Package)، وللحكم على دقة تقدير معلمة الصعوبة للفقرات تم استخدام مؤشر الخطأ المعياري في التقدير باستخدام طريقة البوتستراب لفحص دقة تقدير معلمة صعوبة الفقرة باستخدام دالة pairSE في برمجية لغة R.

٣. تقدير معلمة قدرات الأفراد وفق النموذج أحادي المعلمة باستخدام طريقة الأرجحية الموزونة باستخدام حزمة (pairwise package).

نتائج الدراسة ومناقشتها

أولاً: النتائج الخاصة بسؤال الدراسة الأول الذي نص على: هل تختلف دقة تقدير معلمة صعوبة الفقرة باستخدام طريقة المزاوجة باختلاف حجم عينة المفحوصين وطول الاختبار وفق النموذج أحادي المعلمة ؟

للإجابة عن السؤال الأول في الدراسة، تم تقدير أوساط قيم معلمة الصعوبة للفقرات باختلاف حجم عينة المفحوصين وطول الاختبار باستخدام طريقة المزاوجة وفق النموذج أحادي المعلمة، وإيجاد قيم أوساط الجذر التربيعي لمعدل مربعات الأخطاء (RMSE) باستخدام طريقة البوتستراب في برمجية لغة R، وتم تحديد عدد عينة البوتستراب بحجم العينة الكلية (N). يبين جدول ٤ ملخص نتائج صعوبة الفقرات والخطأ المعياري في تقدير معلمة الصعوبة باختلاف حجم العينة وطول الاختبار.

جدول ٤ أعلى قيمة وأدنى قيمة لمعلمة صعوبة الفقرة وقيم الوسط الحسابي للخطأ المعياري في تقدير صعوبة الفقرات باختلاف عينة المفحوصين وطول الاختبار

طول الاختبار												
	٧٠			٤٠			٣٠			٢٠		
حجم العينة	اعلى قيمة	أدنى قيمة	الوسط الحسابي للخطأ المعياري	اعلى قيمة	أدنى قيمة	الوسط الحسابي للخطأ المعياري	اعلى قيمة	أدنى قيمة	الوسط الحسابي للخطأ المعياري	اعلى قيمة	أدنى قيمة	الوسط الحسابي للخطأ المعياري
١٠٠٠	٢،١٠	-٢،٠١	٠,٠٠٢	١,٨٦	-٢،٠٨	٠,٠٠٢	١,٨٨	-١,٩٨	٠,٠٠٢	٢،٠٧	-١,٩٧	٠,٠٠٢
٥٠٠	٢،٠١	-٢،٠١	٠,٠٠٥	٢،٠٣	-١,٩٦	٠,٠٠٥	١,٩٩	-١,٩٨	٠,٠٠٥	٢،٠٧	-١,٩٤	٠,٠٠٥
٢٥٠	٢،٣٣	-٢،١٧	٠,٠١	٢،٢٧	-٢،٣٥	٠,٠١	١,٨٩	-٢،٠٣	٠,٠١	١,٩٥	-٢،٣٧	٠,٠١
١٠٠	٢،٤٩	-٢،٢٥	٠,٠٣	٣،٢٢	-٢،٢١	٠,٠٣	٢،٤٤	-٢،٢١	٠,٠٣	٣	-٢	٠,٠٣
٥٠	٣،٦٢	-٢،٤٤	٠,٠٥	٢،١٨	-٢،٤٤	٠,٠٥	١,٨٤	-٢،١٩	٠,٠٥	٢،٣٤	-٢،٥٩	٠,٠٦

يلاحظ من جدول ٤ أن الوسط الحسابي للأخطاء المعيارية لتقديرات صعوبة الفقرات كانت صغيرة باختلاف حجم العينة وطول الاختبار، اذ تراوحت قيمة الوسط الحسابي بين ٠,٠٠٢ و ٠,٠٦، وهذا يدل على دقة تقدير صعوبة الفقرة باستخدام طريقة المزاوجة وخاصة عند استخدام عينات صغيرة من طول الاختبار وحجم عينة المفحوصين، وأظهرت النتائج أن تقديرات معلمة الصعوبة باستخدام طريقة المزاوجة كانت تعطي أخطاء معيارية قليلة في جميع اختبارات الدراسة والمطبقة على حجوم عينات مختلفة وقد تراوحت بين ٠,٠٠٢ و٠,٠٥. ويفسر الباحثان هذه النتيجة بأن استخدام حزمة المزاوجة الحديثة لتقدير معلمة الصعوبة والخطأ المعياري في تقديرها أدى إلى الحصول على نتائج دقيقة في تقدير معلمة صعوبة الفقرة وفق النموذج أحادي المعلمة وهذا يؤكد ما أشار إليه زانغ (Zhang, ٢٠١٤) إلى أن طريقة البوتستراب هي أفضل طريقة لتقدير الأخطاء المعيارية لتقديرات المعلمة وخاصة عند استخدام حجوم عينات صغيرة.

ويلاحظ من جدول ٤ ثبات قيمة الوسط الحسابي للخطأ المعياري في التقدير بتغير طول الاختبار وثبات حجم العينة، ويعزو الباحثان هذه النتيجة إلى أن تقدير معلمة صعوبة فقرة معينة لا تتأثر بأداء المفحوصين على فقرة أخرى وصعوبتها، وهذا يدل على جودة البيانات المولدة في هذه الدراسة ومطابقتها للنموذج أحادي المعلمة. ويبين جدول ٥ نتائج تحليل التباين الثنائي لقيم الأخطاء المعيارية لتقديرات معلمة الصعوبة.

جدول ٥ نتائج تحليل التباين الثنائي لقيم الأخطاء المعيارية لتقديرات معلمة الصعوبة المقدرة باختلاف حجم العينة وطول الاختبار

مصدر التباين	مجموع المربعات	درجات الحرية	متوسط مجموع المربعات	قيمة ف المحسوبة	الدلالة الإحصائية
طول الاختبار	٠,٠٠٠١٣	٣	٤,٣٣E-٥	٠,٨٢	٠,٤٩
حجم العينة	٠,٢٣٩٧٠	٤	٠,٠٦	١١٤٠,٦٨	٠,٠٠
طول الاختبار x حجم العينة	٠,٠٠٠٦٢	١٢	٥,١٧E-٥	٠,٩٨	٠,٤٦
الخطأ	٠,٠٤١٠٤	٧٨٠	٤,٢٦E-٥		
الكلي	٠,٦٣٦٧٧٣	٧٩٩			

يلاحظ من جدول ٥ عدم وجود فروق ذات دلالة إحصائية عند مستوى الدلالة $\alpha=0.05$ بين الأخطاء المعيارية لتقديرات الصعوبة المقدرة باستخدام طريقة المزاوجة تعزى لمتغير طول الاختبار، والتفاعل بين حجم العينة وطول الاختبار؛ مما يعني عدم تأثر دقة تقدير صعوبة الفقرات بطول الاختبار والتفاعل بين حجم العينة وطول الاختبار، كما يبين جدول ٥ وجود فروق ذات دلالة احصائية عند مستوى الدلالة $\alpha=0.05$ بين الأخطاء المعيارية لتقديرات الصعوبة تعزى لمتغير حجم العينة؛ أي أنه يوجد تأثيراً لمتغير حجم العينة في دقة تقدير معلمة الصعوبة. ويبين جدول ٦ نتائج اختبار شافيه (Scheffe) للمقارنات البعدية لتحديد لصالح أي من حجوم العينات قد كانت الفروق بينها دالة إحصائياً.

جدول ٦ نتائج اختبار شافيه للمقارنات البعدية لمتغير حجم العينة

العينة i	العينة j	وسط الاختلاف بين العينتين	الدلالة الإحصائية
٥٠	١٠٠	٠,٠٢٩١٤٤	٠٠٠.
	٢٥٠	٠,٠٤٤٠٨٨	٠٠٠.
	٥٠٠	٠,٠٤٩٣١٣	٠٠٠.
	١٠٠٠	٠,٠٥١٩١٣	٠٠٠.
١٠٠	٢٥٠	٠,٠١٤٩٤٤	٠٠٠.
	٥٠٠	٠,٠٢٠١٦٩	٠٠٠.
	١٠٠٠	٠,٠٢٢٧٦٩	٠٠٠.
٢٥٠	٥٠٠	٠,٠٠٥٢٢٥	٠٠٠.
	١٠٠٠	٠,٠٠٧٨٢٥	٠٠٠.
٥٠٠	١٠٠٠	٠,٠٠٢٦	٠٣٧.

يبين جدول ٦ وجود فروق ذات دلالة إحصائية بين حجم العينة ٥٠ من جهة وحجوم العينات (١٠٠، ٢٥٠، ٥٠٠، ١٠٠٠) مفحوص على التوالي، ولصالح حجوم العينات الأربعة، كما تبين وجود فروق ذات دلالة إحصائية بين حجم العينة ١٠٠ من جهة وحجوم العينات (٢٥٠، ٥٠٠، ١٠٠٠) مفحوص على التوالي، ولصالح حجوم العينات الثلاثة، كما تبين وجود فروق ذات دلالة إحصائية بين حجم العينة ٢٥٠ من جهة وحجوم العينات (٥٠٠، ١٠٠٠) مفحوص على التوالي، ولصالح حجم العينة ٥٠٠ و١٠٠٠، كما تبين أيضاً وجود فروق ذات دلالة إحصائية بين حجم العينة ٥٠٠ من جهة حجم العينة ١٠٠٠ مفحوص، ولصالح حجم العينة ١٠٠٠ مفحوص.

أشارت النتائج أنه يوجد فروق ذات دلالة احصائية عند مستوى الدلالة $\alpha=0.05$ بين الأخطاء المعيارية لتقديرات معلمة الصعوبة المقدرة باستخدام طريقة المزاوجة تعزى لمتغير حجم العينة لصالح حجوم العينات الكبيرة، مما يعني تأثر دقة تقدير معلمة صعوبة الفقرات بحجم عينة المفحوصين ولصالح العينات الكبيرة، كما أظهرت نتائج المقارنات البعدية باستخدام اختبار شافييه أنه كلما زاد حجم العينة قلت الأخطاء المعيارية في تقدير صعوبة الفقرة. ويمكن تفسير هذه النتائج بأن تقدير معلمة صعوبة الفقرة في نظرية IRT يحتاج إلى عدد كاف من المفحوصين فزيادة حجم العينة يزيد من دقة تقدير معلمة الصعوبة، حيث أن زيادة حجم عينة المفحوصين المطبق عليهم فقرات الاختبار يؤدي إلى تقديرات دقيقة لمعلمة صعوبة الفقرات وهذا ما أشار إليه هامبلتون (Hambleton, ١٩٨٩) بأن نظرية IRT تحتاج إلى حجوم عينات كبيرة للحصول على تقديرات دقيقة لمعالم الفقرة وقدرة المفحوصين.

تتفق هذه النتائج مع نتائج دراسة ساهين وانيل (Sahin & Anil, 2017) التي أشارت إلى أن دقة التقدير لمعلمة صعوبة الفقرة تزداد بزيادة حجم عينة المفحوصين. كما واتفقت نتائج الدراسة مع نتائج دراسة عبابنة (٢٠٠٤) حيث أشارت النتائج إلى أن دقة تقديرات معالم الفقرة تزداد بزيادة حجم عينة المفحوصين، إذ أن متوسط الأخطاء المعيارية لتقديرات معالم الفقرة تكون أقل ما يمكن عند استخدام حجم عينة ١٠٠٠ مفحوص. كما وتتفق هذه النتائج مع نتائج دراسة هيونغ وآخرون (Huang et al, 2001) والتي أشارت إلى أن أخطاء التقدير لمعلمة الصعوبة ومعلمة التمييز تكون أكبر عندما يقل حجم عينة المفحوصين. وتتفق هذه النتائج مع نتيجة دراسة فينش وفرينش (Finch & French, 2019) التي أشارت أن بزيادة حجم عينة المفحوصين تقل الأخطاء المعيارية في تقدير معلمة الصعوبة، وإن زيادة طول الاختبار ليس بالضرورة أن يؤدي إلى دقة أفضل في تقدير معلمة الصعوبة.

كما أشارت النتائج إلى عدم وجود فروق ذات دلالة إحصائية عند مستوى الدلالة $\alpha=0.05$ بين الأخطاء المعيارية لتقديرات معلمة الصعوبة المقدرة باستخدام طريقة المزاوجة تعزى لمتغير طول الاختبار. ويعزو الباحثان هذه النتائج بأن تقدير معالم الفقرة في نظرية IRT لا يحتاج إلى عدد كبير من الفقرات، أي أن دقة تقدير معلمة الصعوبة لا تتأثر بطول الاختبار وهذا ما يؤكده رايت و ستون (Wright & Stone, ١٩٧٩) إلى عدم وجود طول اختبار محدد يقدم أقل خطأ معياري في تقدير معالم الفقرة. واتفقت هذه النتيجة مع ما توصلت إليه دراسة فيتزباترك (Fitzpatrick, 2008) أن هناك اختلافات وعدم استقرار في معلمة الصعوبة عند استخدام اختبار طوله أقل من ١٥ فقرة، وأوصت الدراسة باستخدام اختبارات طولها أكبر من ١٥ فقرة لزيادة الاستقرار في تقدير معلمة صعوبة الفقرات.

كما وأشارت النتائج أيضاً إلى عدم وجود فروق ذات دلالة إحصائية عند مستوى الدلالة $\alpha=0.05$ بين الأخطاء المعيارية لتقديرات معلمة الصعوبة المقدرة باستخدام طريقة المزاوجة تعزى للتفاعل بين حجم العينة وطول الاختبار، مما يعني عدم تأثر دقة تقدير معلمة الصعوبة بالتفاعل بين حجم عينة المفحوصين وطول الاختبار. واختلفت هذه النتيجة مع نتيجة دراسة الدرابيع (٢٠٠١) بأن هناك فروقا إحصائية تعزى لتفاعل كل من حجم العينة وطول الاختبار على دقة تقدير معلمة صعوبة الفقرة.

إن نتائج هذه الدراسة مشجعة لاستخدام طريقة المزاوجة في تقدير معلمة الصعوبة مع حجوم عينات صغيرة؛ لأن الكثير من المجالات العملية وخاصة التربوية تتطلب استخدام حجوم عينات صغيرة، مما يجعل جمع عينات كبيرة بما يكفي لتقدير معالم نماذج نظرية IRT أمراً صعباً جداً. حيث أشارت نتائج هذه الدراسة إلى أن طريقة المزاوجة تعطي دقة في تقدير معلمة الصعوبة عند استخدامها مع العينات الصغيرة والكبيرة على حد سواء.

ثانياً: النتائج الخاصة بسؤال الدراسة الثاني الذي نص على: هل تختلف دقة تقير معلمة القدرة باستخدام طريقة الأرجحية الموزونة باختلاف حجم عينة المفحوصين وطول الاختبار وفق النموذج أحادي المعلمة ؟

للإجابة عن السؤال الثاني في الدراسة، فقد قدرت قيم معلمة القدرة للمفحوصين باختلاف حجم عينة المفحوصين وطول الاختبار باستخدام طريقة تقدير الأرجحية الموزونة وفق النموذج أحادي المعلمة، وإيجاد قيم الأخطاء المعيارية في تقدير القدرة، وذلك باستخدام برمجية لغة R. ويبين جدول ٧ ملخص نتائج قدرات المفحوصين والخطأ المعياري في تقدير القدرة باختلاف حجم العينة وطول الاختبار.

جدول ٧ أعلى قيمة وأدنى قيمة لمعلمة القدرة وقيم وسطها الحسابي وخطأ تقديرها باختلاف عينة المفحوصين وطول الاختبار

طول الاختبار																
	70				40				30				20			
حجم العينة	اعلى قيمة	أدنى قيمة	وسط القدرة	وسط الخطأ	اعلى قيمة	أدنى قيمة	وسط القدرة	وسط الخطأ	اعلى قيمة	أدنى قيمة	وسط القدرة	وسط الخطأ	اعلى قيمة	أدنى قيمة	وسط القدرة	وسط الخطأ
1000	3.26	-3.26	-0.06	0.29	3.87	-3.88	-0.02	0.39	3.56	-3.58	0	0.44	4.38	-4.38	0.01	0.56
500	3.83	-3.28	0.05	0.30	2.62	-2.91	0.06	0.39	4.78	-4.79	0.09	0.46	3.20	-3.39	-0.05	0.56
250	3.57	-2.67	0.06	0.30	2.93	-2.62	-0.04	0.39	4.74	-3.59	0.14	0.46	3.17	-2.10	0.08	0.55
100	1.90	-2.38	-0.18	0.30	3.45	-2.66	0.01	0.40	2.29	-2.29	-0.06	0.44	4.73	-3.27	-0.03	0.58
50	2.95	-2.06	0.10	0.30	2.44	-1.99	0.32	0.40	2.66	-2.07	0.17	0.45	2.24	-4.59	-0.41	0.62

يلاحظ من جدول ٧ أن قيمة الوسط الحسابي للأخطاء المعيارية تراوحت بين ٠,٢٩ و٠,٦٢ باختلاف طول الاختبار وحجم عينة المفحوصين، كما أن أقل قيمة للخطأ المعياري عند حجم عينة ١٠٠٠ مفحوص وطول اختبار ٧٠ فقرة وأعلى قيمة عند حجم عينة ٥٠ وطول اختبار ٢٠، أي أن كلما زاد طول الاختبار قل الخطأ المعياري في تقدير معلمة القدرة. وقام الباحثان بإجراء تحليل تباين ثنائي Interaction ANOVA باستخدام برمجية (SPSS version 22) لقيم الأخطاء المعيارية لتقديرات معلمة قدرة المفحوصين المقدرة باختلاف متغيري (حجم العينة، طول الاختبار)، ويبين جدول ٨ نتائج تحليل التباين الثنائي.

جدول ٨ نتائج تحليل التباين الثنائي لقيم الأخطاء المعيارية لتقديرات معلمة قدرة المفحوصين المقدرة باختلاف حجم العينة وطول الاختبار

مصدر التباين	مجموع المربعات	درجات الحرية	متوسط مجموع المربعات	قيمة ف المحسوبة	الدلالة الإحصائية
طول الاختبار	٢٧,٤٠٧	3	9.136	2352.54	0.00
حجم العينة	0.109	4	0.027	7.045	0.00
طول الاختبار× حجم العينة	0.259	12	0.022	5.56	0.00
الخطأ	29.44	7580	0.004		
الكلي	101.09	7599			

يلاحظ من جدول ٨ وجود فروق ذات دلالة إحصائية عند مستوى الدلالة $\alpha=0.05$ بين الأخطاء المعيارية لتقديرات معلمة القدرة المقدرة باستخدام طريقة الأرجحية الموزونة تعزى لمتغيري طول الاختبار وحجم العينة، والتفاعل بين حجم العينة وطول الاختبار. ولتحديد لصالح أي من حجوم العينات وطول الاختبار قد كانت الفروق بينها دالة إحصائياً، فقد قام الباحثان بإجراء مقارنات بعدية باستخدام اختبار شافيه (Scheffe). ويبين جدول ٩ نتائج اختبار شافيه للمقارنات البعدية لمتغير حجم العينة.

جدول ٩ نتائج اختبار شافيه للمقارنات البعدية لمتغير حجم العينة

العينة i	العينة j	وسط الاختلاف بين العينتين	الدلالة الإحصائية
٥٠	١٠٠	٠,٠١٢٤	٠,٢٦١
	٢٥٠	٠,٠١٨٦	٠,٠٠٥
	٥٠٠	٠,٠١٦٧	٠,٠١١
	١٠٠٠	٠,٠٢٠٧	٠,٠٠٠
١٠٠	٢٥٠	٠,٠٠٦٢	٠,٥٨٥
	٥٠٠	٠,٠٠٤٣	٠,٨١٥
	١٠٠٠	٠,٠٠٨٣	٠,١٦٧
٢٥٠	٥٠٠	-٠,٠٠١٩	٠,٩٥٨
	١٠٠٠	٠,٠٠٢١	٠,٩٢٣
٥٠٠	١٠٠٠	٠,٠٠٤	٠,٢٣٠

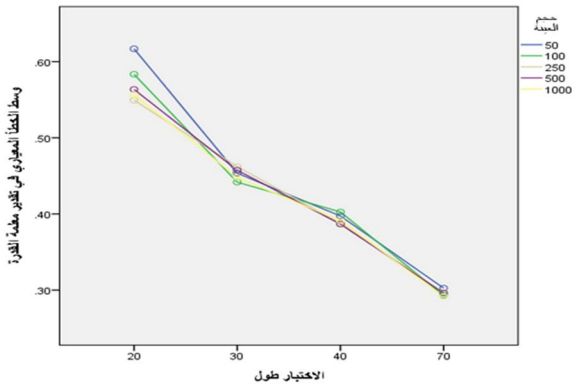
يبين جدول ٩ وجود فروق ذات دلالة إحصائية بين حجم العينة ٥٠ من جهة وحجوم العينات (١٠٠، ٢٥٠، ٥٠٠، ١٠٠٠) مفحوص على التوالي، ولصالح حجوم العينات (٢٥٠، ٥٠٠، ١٠٠٠) مفحوص، كما تبين عدم وجود فروق ذات دلالة إحصائية بين حجم العينة ١٠٠ من جهة وحجوم العينات (٥٠، ٢٥٠، ٥٠٠، ١٠٠٠) مفحوص على التوالي، كما تبين عدم وجود فروق ذات دلالة إحصائية بين حجم العينة ٢٥٠ من جهة وحجوم العينات (٥٠٠، ١٠٠٠) مفحوص على التوالي، كما تبين أيضاً عدم وجود فروق ذات دلالة إحصائية بين حجم العينة ٥٠٠ من جهة حجم العينة ١٠٠٠ مفحوص، مما يعني عدم تأثر دقة تقدير معلمة القدرة عند استخدام عينات ١٠٠ مفحوص فأكثر. ويمكن عزو هذه النتائج إلى أن دقة تقدير قدرات المفحوصين تتسم بالاستقرار عندما يتم استخدام عينات بأحجام كبيرة، كما تتأثر الدقة في تقدير قدرات المفحوصين بوضع مطابقة البيانات للنموذج اللوجستي أحادي المعلمة ويكون عندئذ من غير المؤكد الحصول على دقة أكبر عند زيادة حجوم عينات كبيرة. كما وتشير النتائج إلى أن استخدام عينات صغيرة ٥٠ مفحوص يؤثر سلباً على دقة تقدير قدراتهم وقد يستلزم زيادة دقة تقدير قدرات المفحوصين بشكل طفيف زيادة العينة المستخدمة في تقدير معلمة القدرة، وهذا ما أشار إليه هامبلتون وكوك (Hambleton & Cook, ١٩٨٣) إلى أن زيادة حجم العينة وطول الاختبار تزيد دقة التقدير لقدرة المفحوص. ويبين جدول ١٠ نتائج اختبار شافيه للمقارنات البعدية لمتغير طول الاختبار.

جدول ١٠ نتائج اختبار شافيه للمقارنات البعدية لمتغير طول الاختبار

الاختبار i	الاختبار j	وسط الاختلاف بين العينتين	الدلالة الإحصائية
٢٠	٣٠	٠,١٠٨٣	٠,٠٠٠
	٤٠	٠,١٧٠١	٠,٠٠٠
	٧٠	٠,٢٦٦٧	٠,٠٠٠
٣٠	٤٠	٠,٠٦١٨	٠,٠٠٠
	٧٠	٠,١٥٨٤	٠,٠٠٠
٤٠	٧٠	٠,٠٩٦٦	٠,٠٠٠

يبين جدول ١٠ وجود فروق ذات دلالة إحصائية بين طول الاختبار ٢٠ فقرة من جهة وأطوال الاختبارات (٣٠، ٤٠، ٧٠) فقرة من جهة أخرى، ولصالح أطوال الاختبارات (٣٠، ٤٠، ٧٠) فقرة، كما تبين وجود فروق ذات دلالة إحصائية بين طول الاختبار ٣٠ فقرة من جهة وأطوال الاختبارات (٤٠، ٧٠) فقرة على التوالي، ولصالح أطوال الاختبارات (٤٠، ٧٠) فقرة، كما تبين وجود فروق ذات دلالة إحصائية بين طول الاختبار ٤٠ فقرة من جهة وطول الاختبار ٧٠ فقرة، ولصالح طول الاختبار ٧٠ فقرة، أي كلما زاد طول الاختبار تزداد دقة تقدير معلمة القدرة، مما يشير إلى زيادة طول الاختبار ساهم بشكل واضح في خفض الخطأ المعياري لتقديرات معلمة القدرة، ويعزو الباحثان هذه النتائج إلى أن زيادة طول الاختبار يؤدي إلى اجابة المفحوص على عدد كبير من الفقرات تؤدي إلى تقدير قدرته بشكل أفضل، وهذا ما يؤكده لورد (Lord, ١٩٨٠) إلى أن طول الاختبار يؤثر على جودة تقدير المعالم. تبين من نتائج التحليل على أن العامل الأهم في التأثير على دقة تقديرات معلمة القدرة وفق النموذج اللوجستي أحادي المعلمة هو طول الاختبار وليس حجم العينة عندما يكون حجم العينة متوسطاً (١٠٠، ٢٥٠) مفحوص أو كبيراً (٥٠٠، ١٠٠٠) مفحوص، حيث أشارت النتائج إلى أنه عند استخدام حجم عينة ١٠٠ مفحوص فأكثر فإن الوسط الحسابي لتقدير معلمة القدرة يبقى ثابتاً مع ثبات حجم العينة، وتتفق هذه النتيجة مع نتيجة دراسة عبابنة (٢٠٠٤) التي أشارت إلى أن دقة تقديرات معلمة القدرة تزداد بزيادة طول الاختبار. وأشارت النتائج أيضاً إلى وجود فروق ذات دلالة إحصائية عند مستوى الدلالة $\alpha=0.05$ بين الأخطاء المعيارية لتقديرات معلمة الصعوبة المقدرة باستخدام طريقة الأرجحية الموزونة تعزى للتفاعل بين حجم العينة وطول الاختبار، إن نتائج هذا التفاعل يشير إلى أنه عند استخدام حجم عينة ١٠٠ مفحوص فأكثر فإن دقة تقدير معلمة القدرة تميل إلى الثبات، وهذا الثبات يزداد عندما يكون طول الاختبار أكبر من ٤٠ فقرة. وتتفق هذه النتيجة مع نتائج دراسة الدرابيع (٢٠٠١) التي أشارت إلى وجود فروق ذات دلالة إحصائية لتفاعل كل من حجم عينة المفحوصين وطول الاختبار على دقة تقدير قدرة المفحوص. في حين لم تظهر النتائج في هذه الدراسة الحالية وجود فروق إحصائية في دقة تقدير معلمة القدرة تعزى لحجم عينة المفحوصين، ووجود فروق إحصائية في دقة تقدير معلمة القدرة تعزى لطول الاختبار لصالح طول الاختبار الكبير. واختلفت مع هذه الدراسات في أنها أظهرت بأن أفضل دقة لتقدير معلمة القدرة المفحوص عندما كان حجم العينة ١٠٠ مفحوص وطول الاختبار ٧٠ فقرة. ولتوضيح أثر التفاعل بين حجم العينة وطول الاختبار في دقة تقدير معلمة القدرة، فقد تم تمثيل هذا الأثر بيانياً كما في شكل ٣.

شكل (٣) الرسم البياني لأثر التفاعل بين حجم العينة وطول في دقة تقدير معلمة الصعوبة



التوصيات

في ضوء النتائج التي تم التوصل إليها في هذه الدراسة يمكن الخروج بالتوصيات:

١. نظراً لتمتع طريقة المزاوجة بالدقة الكبيرة في تقدير معلمة الصعوبة وخاصة عند استخدامها مع العينات الصغيرة من المفحوصين وطول الاختبار، يوصي الباحثان بتقدير معلمة الصعوبة وفق النموذج ثنائي المعلمة والنموذج ثلاثي المعلمة.

٢. يوصي الباحثان بإجراء المزيد من الدراسات المستقبلية باستخدام بيانات حقيقية لتقدير معلمة صعوبة الفقرة.

٣. بسبب انخفاض الأخطاء المعيارية في تقدير معلمة الصعوبة عند استخدام طريقة المزاوجة، ويوصي الباحثان بدراسة تهدف للمقارنة بين طريقة الأرجحية العظمى وطريقة بييز وطريقة المزاوجة في دقة تقدير معالم الفقرة.

٤. يوصي الباحثان بدراسة تهدف للمقارنة بين البرمجيات الإحصائية (R, Bilog) لتقدير معلمة الصعوبة باستخدام طرق الأرجحية العظمى وطريقة بييز.

المراجع:

- الدرابيع، ماهر.(٢٠٠١). فعالية النموذج اللوجستي ذي المعلمة الواحدة "نموذج راش" في دقة تقدير قدرة الفرد ومعامل
- صعوبة الفقرة باختلاف حجم العينة وطول الاختبار. مجلة دراسات-العلوم الإنسانية 28(١)، ١٩٧-٢٠٨.
- عبابنة، عماد. (٢٠٠٤). أثر حجم العينة وطريقة انتقائها وعدد الفقرات وطريقة انتقائها على دقة تقدير معالم الفقرة والقدرة
- لاختبار قدرة عقلية باستخدام نظرية استجابة الفقرةأطروحة دكتوراه غير منشورةجامعة عمان العربية للدراسات العليا، عمان.
- علام، صلاح الدين (٢٠٠٥). نماذج الاستجابة للمفردات الاختيارية أحادية البعد ومتعددة الأبعاد وتطبيقاتها في القياس النفسي
- والتربوي الطبعة الأولى دار الفكر العربي

- Barnes, L., & Wise, S. (1991). The utility of a modified one-parameter IRT model with small samples. *Applied Measurement in Education*, 4(2), 143-157.
- Birnbaum, A. (1969). Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology*, 6(2), 258-276.
- Chen, W., Lenderking, W., Jin, Y., Wyrwich, W., Gelhorn, H., & Revicki, A. (2014). Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behaviour item bank data. *Quality of Life Research,* 23(2), 485–493.
- Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. Journal of Educational and Behavioral Statistics, 22(3), 265-289.
- Choppin, B. (1985). A fully conditional estimation procedure for Rasch model parameters. Evaluation in Education, 9(1), 29–42.
- de Gruijter, D. & Van der Kamp, L. (2005). Statistical test theory for education and psychology. Graduate School of Education, University of Leiden.
- De la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size a higher-order IRT model approach. *Applied Psychological Measurement,* 34(4), 267–285.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7, 1-26.
- Finch, H., & French, B. (2019). A comparison of estimation techniques for IRT models with small samples. Applied Measurement in Education, 32(2), 77-96.
- Fitzpatrick, A. (2008). The impact of anchor test configuration on student proficiency rates. *Educational Measurement: Issues and* Practice, 27(4), 34-40.
- Hambleton, R. (1989). Principles and Selected applications of item response theory, In Linn, Robert, L. (Ed.), Educational Measurement (3rd ed., PP. 147-201). American Council on Education, Macmillan Publishing Company.
- Hambleton, R., & Cook, L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. In D. J. Weiss & R. D. Bock (Eds.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 31-49). Academic Press.
- Hambleton, R. & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Kluwer, Nijhoff Publishing.
- Hambleton, R., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Sage Publications.
- Hattie, John. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied psychological measurement,* 9(2), 139-164.
- Heine, J., & Tarnai, C. (2015). Pairwise Rasch model item parameter recovery under sparse data conditions. *Psychological Test and Assessment Modeling*, 57(1), 3-36.

- Huang, C., Lohss, W., Lin, C., & Shin, D. (2001). Item calibrations for licensure tests with multiple specialty components. Submitted to division ID: *Educational Measurement, Psychometrics and Assessment.* Enabled Tiger a web – based Manuscripts Processing system. Michigan State University.
- Imon, A., & Masoom, A. (2005). Bootstrapping regression residuals. *Journal of the Korean Data and Information Science Society,* 16(3), 665-682.
- Jiang, S. Wang, C., & Weiss, D. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in Psychology,* 7(109), 1-10.
- Lord, F. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 28(4), 989-1020.
- Lord, F. (1980). Applications of item response theory to practical testing problems. Lawrence Erlbaum Associates, publishers Hillsdale.
- Lord, F. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement,* 23(2), 157-162.
- Linacre, j. (2008). WINSTEPS Rasch measurement (3.63.2). Chicago, IL:MESA press.
- Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Danish Institute of Educational Research.
- Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of educational statistics,* 4(3), 207-230.
- Reeve, B., & Fayers, P. (2004). *Applications of item response theory (IRT) modelling for building and evaluating questionnaires measuring* patient-reported outcome-s. Bethesda MD: Advances in Health Outcomes Measurement. National Cancer Institute.
- Sahin, A., Anil, D., 2017. The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory and Practice,* 17 (1), 321-335.
- Warm, T. (1978). A Primer of item response theory. Coast Guard Institute.
- Warm, T. (1989). Weight likelihood estimation of ability in item response theory. *Psychometrikia,* 54(3), 427-450.
- Wright, B., & Stone, M. (1979). Best test design. MESA Press.
- Zhang, G. (2014). Estimating standard errors in exploratory factor analysis. Multivariate Behavioral Research, 49(4), 339-353.