تقصي أثر طول الاختبار وحجم العينة على دقة طرق تقدير معالم الفقرات وقدرات الأفراد في برنامج بايلوج

زايد صالح بني عطا

قسم علم النفس الإرشادي والتربوي

كلية التربية – جامعة اليرموك – إربد – الأردن

zaied_baniata@yahoo.com

Received: 28 Oct. 2016
Revised: 04 May 2017, Accepted: 03 Jun. 2017
Published online: 1 (October) 2017

International Journal of Research in Education and Psychology
ISSN 2210-1780
Int. J. Res. edu. Psy. 5, No. 2 (Oct. 2017)

تقصي أثر طول الاختبار وحجم العينة على دقة طرق تقدير معالم الفقرات وقدرات الأفراد في برنامج بايلوج

زايد صالح بني عطا[1]*

قسم علم النفس الإرشادي والتربوي

جامعة اليرموك – إربد – الأردن

الملخص

هدفت الدراسة إلى تقصي أثر طول الاختبار وحجم العينة على دقة طرق تقدير معالم الفقرات وقدرات الأفراد في برنامج بايلوج. ولتحقيق الهدف من الدراسة تم توليد بيانات ثنائية التدريج من خلال النموذج اللوجستي ثلاثي المعلمة بواقع ١٠٠ مرة لأربعة اختبارات طول كل منها (٨٠، ٦٠، ٣٠، ١٠) وأربعة مستويات لحجم العينة (١٥٠٠، ١٠٠٠، ٥٠٠، ٢٥٠) باستخدام برنامج (WINGEN). وباستخدام طرق التقدير (الأرجحية العظمى، توقع الاقتران، تعظيم الاقتران) المستخدمة في برنامج (Bilog- Mg3) تم تحليل البيانات المولدة من أجل تقدير معالم الفقرات وقدرة الأفراد وإيجاد قيم الجذر التربيعي لمتوسطات مربعات الانحرافات للفروق بين المعالم الحقيقة والمقدرة (RMSE). أظهرت نتائج الدراسة وجود أثر ذي دلالة إحصائية لكل من طول الاختبار وحجم العينة والتفاعل بينهما في دقة تقديرات معالم الفقرات (التمييز، الصعوبة، التخمين) عند استخدام طريقة الأرجحية العظمى في معايرة الفقرات، وكانت تزداد دقة تقديرات معالم الفقرات بزيادة طول الاختبار وحجم العينة. وبينت كذلك النتائج بأنه عند معايرة قدرة الأفراد بطرق التقدير الثلاثة المستخدمة في برنامج بايلوج عن وجود أثر ذي دلالة إحصائية لكل من طول الاختبار وحجم العينة وطريقة التقدير في دقة تقديرات معلمة القدرة للفرد، حيث تفوقت طريقة توقع الاقتران (EAP) على بقية الطرق في دقتها لتقدير معلمة القدرة ، وأظهرت النتائج بشكل عام أن طرق التقدير المستخدمة في برنامج بايلوج أنتجت تقديرات دقيقة لمعالم الفقرات وقدرات الأفراد عندما كان طول الاختبار (٣٠) فقرة وأعلى وحجم العينة (٥٠٠) فأعلى عند استخدام النموذج اللوجستي ثلاثي المعلمة.

الكلمات المفتاحية: برنامج بايلوج، النموذج الثلاثي المعلمة، حجم العينة، طول الاختبار، طرق التقدير.

١* أستاذ مشارك في القياس والتقويم، جامعة اليرموك، الأردن.

Int. J. Res. Edu. Psy. 5, No. 2 (Oct. 2017)

Investigating te effect of test length and sample size on the accuracy methods forestimating items parameters and persons' abilities in the Bilog program

Zaid Saleh Bani Ata.

Faculty of Education, Yarmouk University,

Irbid - Jordan

Abstract

This study aimed at investigating the effect of test length and sample size on the accuracy methods for estimating item parameters and persons abilities in the program Bilog. To achieve this goals 100 replicated binary responses through three- parameter model were generated using WINGEN program on four test lengths (10, 30 ,60, 80) and four sample sizes (250, 500, 1000, 1500). And using estimation methods (MML, EAP, MAP) used in Bilog program the generated data were analyzed by Bilog program to estimate the item parameter(discrimination, difficulty, guessing), the abilities of persons, and the computed values of RMSE. The results revealed significant differences due to sample size and test length and their interaction in the accuracy estimates of item parameter when using marginal maximum likelihood (MML), and the accuracy of item parameter increased when increasing test length and sample size. The results also revealed that when calibrating the person ability using three estimation methods used in Bilog program there is significant differences due to sample size, test length and estimation method and their interaction in the accuracy estimates of person ability, also the method (EAP) is better accurate than other estimation methods of person ability. In general the results revealed the accuracy item parameters and abilities of persons increased as the test length exceeded 30 items and the sample size exceeded 500 subjects.

Keywords: Bilog program, Three- Parameter Logistic Model, Sample size, Test length, Estimation methods.

تقصي أثر طول الاختبار وحجم العينة على دقة طرق تقدير معالم الفقرات وقدرات الأفراد في برنامج بايلوج

زايد صالح بني عطا

قسم علم النفس الإرشادي والتربوي

كلية التربية – جامعة اليرموك

إربد – الأردن

خلفية الدراسة

تعد نظرية الاستجابة للفقرة (Item Response Theory) بمثابة الثورة والمستقبل الزاهر للقياس النفسي والتربوي (Anstasi & Urbena, 2005)، حيث قدمت إطارا مرجعيا شاملا لبناء المقاييس النفسية والتربوية، وطريقة تفسير الدرجات على هذه المقاييس ومؤشرات إحصائية عن الدقة والموضوعية في القياس من خلال تقدير الخطأ المعياري للقياس لكل مستوى من مستويات السمة المستهدفة بعملية القياس مقارنة بما قدمته النظرية التقليدية في القياس (Mislevy & Bock, 1990; Van der Linden, 2010).

وقد انبثق عن نظرية الاستجابة للفقرة مجموعة من النماذج الاحتمالية القائمة على الاقتران اللوغارتيمي التي تحدد العلاقة بين أداء الفرد على الفقرة والقدرة التي تكمن وراء هذا الأداء، وأن العلاقة بين أداء الفرد على الفقرة وقدرته يمكن أن تحدد من خلال ما يسمى منحنى خاصية الفقرة (Item Characteristic Curve)، وتفترض كذلك أن مقدار الاحتمال يكون دالة متزايدة وتيريا لموقع الفرد على متصل القدرة، مما يعني أن احتمال الإجابة الصحيحة يزداد بزيادة قدرة الفرد (Hambleton, 1994; Henard, 2000). وتصنف هذه النماذج حسب مستوى الاستجابة إلى نماذج ثنائية التدريج ونماذج متعددة التدريج (De Gruijter & Van der Kamp, 2005). وتعد النماذج ثنائية التدريج من أشهر النماذج أحادية البعد استخداما في بناء الاختبارات والمقاييس النفسية والتربوية (De Ayala, 2009; Embreston & Rise, 2000).

وقد بين بيكر (Baker, 2001) أن النماذج ثنائية التدريج تختلف باختلاف معالم الفقرة المراد تقديرها، فالنموذج اللوغارتيمي ذو المعلمة الواحدة (One- Parameter Logistic Model)، أو المشهور باسم نموذج راش (Rasch Model) يقدر معلمة الصعوبة والقدرة، أما النموذج الثاني فهو النموذج اللوغارتيمي ثنائي المعلمة (Two-Parameter Logistic Model) الذي يقدر معلمتي الصعوبة والتمييز للفقرة وقدرة الفرد، والثالث من بين هذه النماذج وهو النموذج اللوجستي ثلاثي المعلمة (Three –Parameter Logistic Model) الذي يقدر معالم الفقرات الصعوبة، والتمييز والتخمين بالإضافة إلى قدرة الفرد.

إن النجاح في استخدام نظرية الاستجابة للفقرة والاستفادة من مزاياها مقارنة مع نظرية القياس التقليدية في بناء وتطوير المقاييس النفسية والتربوية وتحليل فقراتها يعتمد بالدرجة الأولى على تقدير معالم النموذج المستخدم، التي تعتمد على أساليب التحليل العددي (Numerical Analysis)، حيث اعتبر لورد ونوفيك (Lord &

(Novick, 1968) عملية التقدير الإحصائي للعلاقة بين احتمال الاستجابة الصحيحة عن فقرة من فقرات الاختبار والقدرة التي يقيسها المشكلة الرئيسة لمستخدم هذه النظرية؛ مما جعل البحث السيكومتري يهتم بالبحث عن أفضل أساليب التقدير الإحصائي لمعالم الفقرات وقدرات الأفراد، بالإضافة إلى ذلك تطوير النماذج الاحتمالية للوصول إلى أفضل التقديرات، وتطوير البرامج الحاسوبية لإجراء عمليات التقدير بيسر وسهولة.

ولتقدير معالم النموذج المستخدم فقد أوجد البحث السيكومتري طرقًا مختلفة تعتمد في غالبيتها على أسلوبين؛ إما أن يتم تقدير معالم الفقرة وقدرات الأفراد معا بشكل مشترك، وهذا الأسلوب يعتمد على تقديرات الأرجحية القصوى، وإما أن تتم عملية التقدير بشكل هامشي؛ أي أنه يتم تقدير معالم الفقرات أولا بعد حذف معالم قدرات الأفراد عن طريق تكامل اقتران الاحتمالية الهامشي، ثم إيجاد قدرات الأفراد ويعتمد هذا الأسلوب على منهج بييز (De Ayala, 2009; Fox, 2010; Hambleton, 1989).

ولأن هذه الأساليب تتطلب عمليات رياضية متقدمة ومعقدة، فقد تركزت جهود الباحثين والشركات العالمية في تطوير البرمجيات الحاسوبية المتنوعة لإيجاد تقديرات معالم النموذج المستخدم، ومن أشهر الشركات العالمية التي صممت برامج متنوعة للتعامل مع نماذج نظرية الاستجابة للفقرة شركة البرامج العلمية الدولية (Scientific Software International, Inc)، حيث طورت أكثر من برنامج، ومن هذه البرامج برنامج بايلوج (Bilog-MG3).

ويعد برنامج Bilog-MG من أكثر البرامج شهرة في استخدامه لتحليل فقرات الاختبارات باستخدام أحد نماذج نظرية الاستجابة ثنائية التدريج) أحادي المعلمة، ثنائي المعلمة وثلاثي المعلمة)، وذلك لاعتماده على طريقة الأرجحية العظمى ومنهج بيز في تقدير معالم النموذج، واحتوائه كذلك على ميزات أكثر من البرامج الأخرى، حيث يمكن من خلاله قراءة الملفات والبيانات من برامج مختلفة وبصيغ متنوعة ويمكن التعامل مع مجموعات مختلفة وعينات مختلفة (Bock & Zimowski, 1996).

ويعتمد برنامج بايلوج (Bilog-Mg3) لتقدير معالم الفقرة وقدرة الأفراد لنماذج نظرية الاستجابة للفقرة ثنائية التدريج (أحادي المعلمة، ثنائي المعلمة ، ثلاثي المعلمة) الطرق التالية: (Rupp, 2003; Zimowski, Muraki, Mislevy & Bock, 2003)

أولا: طريقة الأرجحية العظمى الهامشية (Marginal Maximum likelihood: MML)

طورت هذه الطريقة من قبل بوك واتيكن (Bock & Aticken, 1981) حيث استخدمت هذه الطريقة لتقدير معالم الفقرات وقدرات الأفراد في برنامج بايلوج، وتجري عملية تقدير معالم الفقرات في هذه الطريقة بإيجاد اقتران الاحتمالية الهامشي (Marginal Likelihood Function) من خلال تكامل اقتران الكثافة الاحتمالية على معلم القدرة للفرد لإيجاد معالم الفقرات وفق العلاقة الرياضية التالية (Du Toti, 2003):

$$L(a,b,c) = \prod_{-\infty}^{\infty} \int_{a=1}^{\infty} g(\theta_a) L(\theta_a; a,b,c) d\theta_a \quad \ldots\ldots 1$$

وبعد معرفة معالم الفقرات يتم تقدير معلمة القدرة للفرد من خلال تعظيم دالة الارجحية العظمى التي تعطى بالعلاقة التالية:

$$Log_e L_i(\theta) = \sum_{j=1}^{n} \left\{ x_{ij} Log_e p_j(\theta) + (1 - x_{ij}) Log_e \left(1 - p_j(\theta)\right) \right\} \quad \ldots 2$$

حيث يشير P_j إلى احتمالية إجابة الفرد ذو القدرة (θ) عن الفقرة (i) إجابة صحيحة، وبالتالي فان إيجاد قيمة معلمة القدرة للفرد تتم من خلال إيجاد المشتقة الأولى لدالة الترجيح الواردة في المعادلة رقم (٢) ومساواتها بالصفر. وبعد ذلك يتم القيام بالتكرار المتعاقب وفق طريقة نيوتن رافسون للحصول على تقديرات ثابتة لمعلمة القدرة للفرد من خلال أسلوب فيشر (Fisher-

Scoring Solution)، ويتم تقدير الخطأ المعياري لتقدير معلمة القدرة للفرد وفق هذه الطريقة من خلال إيجاد الجذر التربيعي لمعكوس دالة المعلومات (Zimowski, et al., 2003)

وتعد هذه الطريقة من الطرق المشهورة في تقدير معالم الفقرات وقدرات الأفراد لفاعليتها في التقدير سواء كان عدد فقرات الاختبار قليلا أو كثيرا، بالإضافة إلى ذلك القيم التقديرية للأخطاء المعيارية الناتجة منها تتميز بالدقة نتيجة الإعادة المتعاقبة لعمليات التقدير، ويمكن من خلالها الحصول على قيم تقديرية لمعلمة قدرات الأفراد الذين أجابوا إجابة صحيحة أو إجابة خاطئة عن جميع فقرات الاختبار، وبالتالي لا يوجد هدر للمعلومات ناتج عن حذف استجابة بعض الأفراد (Zimowski, et al., 2003). وتجدر الإشارة كذلك بأن برنامج بايلوج يستخدم هذه الطريقة فقط في تقدير معالم الفقرات للنماذج الأحادية ثنائية التدريج بالإضافة إلى معلمة القدرة للفرد، ولا يستخدم الطرق الأخرى في تقدير معالم الفقرات لنفس النماذج (Du Toti, 2003).

ولكن يشوب هذه الطريقة بعض القصور المتمثل بحاجتها إلى توزيع القدرة مسبقا، قبل البدء بعملية تقدير معالم الفقرات، مما يجعل عملية التقدير معتمدا على ملائمة مستويات القدرة الافتراضية، بالإضافة إلى ذلك التقدير غير الدقيق لمعلمة التخمين عند استخدام النموذج اللوجستي ثلاثي المعلمة مما يؤثر سلبا على دقة تقديرات معالم الفقرة وقدرات الأفراد. ومن المآخذ الأخرى التي وجهت لهذه الطريقة بأنها تحتاج إلى عينات كبيرة للوصول إلى دقة في تقدير معالم الفقرات من أجل الاقتراب من توزيع مستويات القدرة التي تحتاجها لتقدير الارجحية العظمى لمعالم الفقرات (علام، ٢٠٠٥؛ Hambeleton & Swaminthan, 1985).

ثانيا: أسلوب بييز (Bayesian Estimation)

يعتمد برنامج بايلوج هذا الأسلوب في تقدير معلمة القدرة للفرد وليس في تقدير معالم الفقرات كما أشير سابقا، ويعود أسلوب بييز في الاستدلال الإحصائي للعالم البريطاني توماس بييز (Thomas Bayes) حيث يمكن من خلاله تقدير معالم الفقرات ومعلمة القدرة للفرد (Bellhouse, 2004). ويتم تقدير معلمة القدرة للفرد وفق هذا الأسلوب في برنامج بايلوج بطريقتين: يطلق على الأولى توقع الاقتران البعدي (Expected A Posteriori: EAP)، أما الطريقة الثانية وهي تعظيم الاقتران البعدي (Maximum A Posteriori: MAP).

وتستند طريقة توقع الاقتران البعدي (EAP) في إيجاد تقديرات معلمة القدرة للفرد في برنامج بايلوج إلى تقديرات معالم الفقرات من خلال طريقة الارجحية العظمى (MML) حيث يحسب توقع الاقتران البعدي (EAP) من خلال الوسط الحسابي للتوزيع البعدي لمعلمة القدرة (θ) دون الحاجة إلى تقريب متتابع كما في طريقة الأرجحية العظمى، وتقوم فكرة تقدير معلمة القدرة للفرد وفقا لهذه الطريقة على تقسيم متصل القدرة إلى (٦١) نقطة بطول (٠,١) حيث تسمى كل نقطة تربيع ((Qr ويحدد لكل نقطة تربيع وزن ((W(Qr)، وفي هذه الطريقة يتم تقدير معلمة القدرة للفرد بدون إجراء عمليات تقريب متتابع من خلال المعادلة الرياضية التالية (Du Toti, 2003):

$$\theta = \sum_{r=1}^{61} \frac{\left[Q_r \times L(Q_r) \times w(Q_r)\right]}{\left[L(Q_r) \times w(Q_r)\right]} \quad \text{.....3}$$

حيث تشير $L(Q_r)$ إلى دالة الأرجحية العظمى، وتمتاز هذه الطريقة بأنها غير متكررة وتعطي تقديرات لقدرات الأفراد (θ) لجميع أنماط الاستجابة، وتقاس درجة الدقة من خلال الخطأ المعياري المحسوب من الانحراف المعياري للتوزيع البعدي لمعلمة قدرات الأفراد (θ). ولكن يؤخذ عليها بأن التقدير يكون متحيزا نحو وسط التوزيع عندما يكون عدد فقرات الاختبار قليلا وذلك لأنها تعتمد على معلومات أولية عن وسط المجتمع وانحرافه المعياري.

ويمكن كذلك تقدير معلمة القدرة للفرد من خلال الطريقة الثانية، طريقة تعظيم الاقتران

(MAP)، وإجراءات هذه الطريقة مماثلة لإجراءات طريقة الأرجحية العظمى في تقدير معلمة القدرة للفرد، حيث يتم في هذه الطريقة تقدير معلمة القدرة للفرد من خلال معادلة الترجيح التالية:

$$p(\theta/x) = \sum_{j=1}^{n} (x_{ij} \operatorname{Log}_e(\theta) + (1-x_{ij}) \operatorname{Log}_e(1-P_j(\theta)) + \operatorname{Log}_e(g(\theta)) \quad \text{.....4}$$

ويشيرgo إلى اقتران الكثافة الاحتمالية لمعلمة القدرة، ويتم إيجاد قيمة معلمة القدرة للفرد من خلال تفاضل المعادلة رقم (٤) ومساواتها بالصفر، وهذه تشبه طريقة الأرجحية العظمى، إذ يتم تقديرها بعد ذلك عن طريق معادلة فيشر. ومما يؤخذ على هذه الطريقة في تقدير معلمة القدرة للفرد أنها تعطي أخطاء معيارية أعلى من طريقة توقع الاقتران البعدي (EAP) خصوصا عندما يكون عدد فقرات الاختبار أقل من (٢٠) فقرة (Zimowski et al, 2003).

وعلى الرغم من الانتقادات التي وجهت لأسلوب بييز في إيجاد تقديرات القدرة للأفراد والمتعلقة بطبيعة المعلومات الأولية (التوزيع القبلي)، فقد أشار جووجين (Gao & Chen, 2005) إلى أن تقديرات بييز تعطي نفس نتيجة تقديرات الأرجحية العظمى إذا كان التوزيع القبلي غير غني بالمعلومات، أما إذا كان التوزيع القبلي طبيعيا فإن تقديرات بييز أكثر دقة من طريقة الأرجحية العظمى.

ولأن الخطأ المعياري يعد أحد أهم المؤشرات الإحصائية للحكم على دقة تقديرات معالم الفقرات وقدرات الأفراد، فقد وفرت طرق تقدير معالم الفقرات وقدرات الأفراد المستخدمة في برنامج بايلوج هذا المؤشر الإحصائي، حيث أشار ثيسين ووينر ودراسقو (Drasgow, 1989; Thissen & Wainer, 1982) إلى أهمية تحديد مقدار الخطأ المعياري في تقدير معالم الفقرات وقدرات الأفراد كمؤشر على دقة القياس التي تساعد في اتخاذ القرارات الصائبة في المجالات التربوية والنفسية المختلفة في ضوء نتائج المقاييس والاختبارات.

ونقطة أخرى جديرة بالذكر وهي أن هذا المؤشر يلعب دورا مهما في دالة معلومات الاختبار(Test Information Function) التي من خلالها يستدل منها على ثبات الاختبار في نظرية الاستجابة للفقرة (Embreston & Rise, 2000).

ونظرا للاستخدام المتزايد لنماذج نظرية الاستجابة للفقرة والاستفادة من مزاياها في التقويم النفسي والتربوي، اهتم البحث السيكومتري في البحث عن العوامل التي تؤثر في دقة تقديرات معالم الفقرات وقدرات الأفراد، ومن أهم هذه العوامل التي اهتمّ بها البحث السيكومتري طريقة التقدير المستخدمة في تقدير معالم النموذج المستخدم، بالإضافة إلى طول الاختبار وحجم عينة المفحوصين، حيث تباينت وجهة نظر الباحثين حول الطريقة الأفضل في تقدير معالم النموذج، فقد بينت نتائج الدراسات التي قام بها سواميناثان وجيفورد (Swaminathan & Gifford, 1982, 1985, 1986) حول المفاضلة بين طريقة بييز والأرجحية القصوى بأن طريقة بيز أنتجت تقديرات لقدرات الأفراد أكثر دقة من طريقة الأرجحية القصوى بالاعتماد على مؤشر الجذر التربيعي لمتوسط مربعات الانحرافات للفروق بين المعالم الحقيقية والمقدرة RMSE (Root Mean Standard Error)) من خلال توليد بيانات باستخدام النماذج ثنائية التدريج .

وقام وانغ وفيسبول (Wang & Vispoel, 1998) بدراسة هدفت إلى تقويم فاعلية أربع طرق في تقدير معلمة القدرة للأفراد هي: طريقة الأرجحية العظمى، وثلاث طرق مرتبطة بطريقة بييز، هي: طريقة أوينز (Owen's method)، وطريقة تعظيم التوزيع البعدي (MAP) وطريقة توقع التوزيع البعدي (EAP). ولتحقيق الهدف من الدراسة تم استخدام أسلوب المحاكاة لاختبار تكيفي. أشارت نتائج الدراسة إلى وجود اختلافات واضحة بين طريقة الأرجحية العظمى وطريقة بييز في دقة تقديرات معلمة القدرة للأفراد حيث كانت طريقة الأرجحية العظمى تنتج أخطاء عالية في التقدير

مقارنة مع أسلوب بييز، وفيما يتعلق بطرق بييز كانت طريقة أوينز أقل الطرق في دقة تقدير معلمة القدرة.

وهدفت دراسة جاو وجن (Gau & Chen, 2005) إلى المقارنة بين طريقتي الأرجحية العظمى الهامشية وطريقة بيز الهامشية باستخدام أسلوب تعظيم الاقتران في دقة تقدير معالم الفقرات. ولتحقيق الهدف من الدراسة تم توليد بيانات بأحجام عينات (٥٠٠، ١٠٠٠، ٢٠٠٠) واختبار بأطوال (٦٠، ٣٠، ١٠). أظهرت النتائج من خلال استخدام النموذج ثلاثي المعلمة بأن دقة تقديرات الأرجحية العظمى وطريقة بييز عند أحجام العينات (٢٠٠٠، ٥٠٠) كانت متشابهة من حيث الدقة بالاعتماد على مؤشر (RMSD) بغض النظر عن طول الاختبار، في حين كانت طريقة بيز أفضل عندما كان حجم العينة ١٠٠.

وأجرى غاري وفيرمونت (Garre & Vermunt, 2006) دراسة هدفت إلى تفادي تقديرات أطراف المتصل في نظرية الاستجابة للفقرة باستخدام التوزيع القبلي البيزي. ولتحقيق الهدف من الدراسة تم الاعتماد على بيانات حقيقية من اختبارين: الأول مكون من (٥) فقرات، والثاني مؤلف من (٩) فقرات على عينات مكونة من (١٠٠٠ ، ١٠٠)، حيث تم استخدام طريقة الأرجحية العظمى وطريقة بييز في تقدير معالم الفقرات وقدرة الأفراد. بينت نتائج الدراسة بأن تقديرات بييز أكثر ثباتا في تقدير معالم القدرة من طريقة الأرجحية العظمى خصوصا عند أطراف المتصل (أي ذوي القدرات العالية، وذوي القدرات المنخفضة).

وقام العبابنه (٢٠٠٦) بدراسة هدفت إلى مقارنة الأرجحية العظمى وطريقة بيز في تقدير معلمة القدرة. ولتحقيق الهدف من الدراسة تم استخدام اختبار قدرة عقلية مطور من قبل الباحث. أظهرت نتائج الدراسة أن دقة تقديرات معلمة القدرة تزداد في حالة عينة ذوي القدرة العالية والمتدنية عند استخدام طريقة بييز مقارنة مع الأرجحية العظمى بغض النظر عن عدد الفقرات المستخدمة، وبينت النتائج زيادة دقة تقديرات معلمة القدرة عند استخدام طريقة الأرجحية العظمى مقارنة بدقة تقديرات معلمة القدرة عند استخدام طريقة بييز وذلك عند المعايرة بجميع فقرات المقياس، بينما يحدث العكس في حالة الاقتصار على عينة فقرات من المقياس، وفي حال وجود تباين بين متوسطات قدرات المفحوصين ومتوسط صعوبة الفقرة فإن دقة تقديرات طريقة الأرجحية العظمى لمعلمة القدرة تكون أعلى منها باستخدام طريقة بييز.

كما أجرى الشريفين (٢٠١٢) دراسة هدفت إلى الكشف عن أثر طريقة تقدير معالم الفقرات وقدرة الأفراد على قيم معالم الفقرة، والخصائص السيكوترية للاختبار في ضوء تغير حجم العينة. ولتحقيق الهدف من الدراسة تم تطبيق اختبار فيزياء مكون من (٢٣) فقرة على (١٠٠٠) طالب وطالبة من طلبة الصف الثاني عشر. أظهرت نتائج الدراسة وجود فروق دالة إحصائيا بين متوسطات الأخطاء المعيارية لتقديرات قدرات الأفراد يعزى للتفاعل بين طريقة التقدير وحجم العينة، لصالح طريقة بييز في التقدير، وخاصة عند العينات الصغيرة، في حين لم تظهر فروق ذات دلالة إحصائية تعزى لمتغير حجم العينة أو طريقة التقدير.

وعلى صعيد الدراسات التي اهتمت بمقارنة فعالية بعض البرامج فقد أشارت نتائج الدراسات (Mislevey & Stocking, 1989; Qualls & Ansley, 1985; Yen, 1987) التي فاضلت بين برنامج Bilog وبرنامج Logist تحت شروط حجم العينة وطول الاختبار بأن دقة التقديرات الناتجة من استخدام برنامج Bilog بشكل عام كانت أدق من تقديرات برنامج Logist.

وهدفت دراسة باتسولا وجيسرو (Patsula & Gessoroh, 1995) إلى مقارنة أثر طول الاختبار وحجم العينة على دقة تقديرات معالم الفقرة باستخدام برنامجي Bilog , Testgraf من خلال

المحاكاة. بينت النتائج بأن تقديرات معالم الفقرات كانت متماثلة للبرنامجين كليهما في حين كان برنامج Testgraf أكثر دقة في تقدير معلمة التخمين عند جميع أطوال الاختبار المستخدم (٢٠، ٤٠)، وبشكل عام أظهرت النتائج كذلك أن دقة تقديرات معالم الفقرة تتماثل لكلا البرنامجين بزيادة حجم العينة.

وأجرى هيونغ ولوهسس ولين وشين (Huang, Lohss, Lin & Shin, 2001) دراسة هدفت إلى مقارنة فعالية بعض البرامج وهي (Bilog, Bilog-Mg, Pic) في معايرة اختبار إجازة باختلاف حجم العينة من خلال توليد ٣٦٠ فقرة باستخدام النموذج ثلاثي المعلمة، وأحجام عينات تتراوح من ٢٥٠ مفحوص إلى ١٥٠٠، وذلك من خلال تشكيل ١٤ عينة. أظهرت النتائج أن أخطاء التقدير لمعلمة الصعوبة ومعلمة التمييز تكون أكبر عندما يقل حجم العينة في حال استخدام برنامج (Bilog).

يتضح بشكل عام من الأدب النظري بأن هناك توجها عاما نحو البحث عن دقة تقديرات معالم الفقرات وقدرات الأفراد خصوصا عند استخدام نظرية الاستجابة للفقرة؛ لما لهذه الدقة للتقديرات من أثر في التطبيقات العملية للاختبارات والمقاييس النفسية والتربوية في المواقف المختلفة، حيث أجمعت الدراسات على وجود عوامل كثيرة تؤثر في دقة تقديرات معالم الفقرات وقدرات الأفراد من أهمها طول الاختبار، وحجم العينة، وطرق تقديرها المتمثلة بتعدد البرامج الحاسوبية، ويتضح كذلك وجود التباين في نتائج الدراسات من حيث تحديد الطريقة الأفضل في التقدير.

ولأن برنامج Bilog-MG يعد من أكثر البرامج العالمية شهرة في استخدامه لتحليل فقرات الاختبارات ثنائية التدريج باستخدام أحد نماذج نظرية الاستجابة (أحادي المعلمة، ثنائي المعلمة، ثلاثي المعلمة) وذلك لاعتماده على طريقة الأرجحية العظمى وأسلوب بييز في تقدير معالم النموذج، واحتوائه على ميزات كثيرة وسهولة التعامل معه (Bock & Zimowski, 1996). لذا جاءت هذه الدراسة لتقصي دقة طرق تقدير معالم الفقرات وقدرات الأفراد المستخدمة في برنامج Bilog-Mg تحت شروط حجم العينة، وطول الاختبار باستخدام النموذج اللوجستي ثلاثي المعلمة كمحاولة لتقييم دقة وكفاءة طرق التقدير في برنامج Bilog-Mg في تقدير خصائص الفقرات وأخطاء القياس كمؤشر على دقة القياس من خلال استخدام أسلوب المحاكاة. ونقطة أخرى جديرة بالذكر هي أن أحد أهم الأسباب التي تجعل هناك حاجة ماسة لدراسة دقة التقديرات من خلال الأخطاء المعيارية هو أن مطوري الاختبارات والمقاييس النفسية والتربوية لا يستخدمون نفس الطريقة ونفس البرنامج في تقدير معالم الفقرات وقدرات الأفراد وعند استخدام برامج متنوعة في تقدير معالم النموذج لا يمكن الحصول على نفس المعالم. وقد تم اختيار النموذج اللوجستي ثلاثي المعلمة، لأن هذا النموذج يعد النموذج العام للنماذج ثنائية التدريج، والأقل تشددا، ولأن هذا النموذج يفترض تأثر الإجابات بعامل التخمين الذي تميز به والذي يعد أحد العوامل المؤثرة في أداء الاختبار(De Ayala, 2009)، وفي الوقت نفسه يلاحظ ندرة الدراسات العربية على المستوى العربي والمحلي التي اهتمت بتقييم دقة طرق تقدير معالم الفقرات وقدرة الأفراد المستخدمة في برنامج Bilog-Mg ، فلم يجد الباحث أي دراسة تناولت موضوع الدراسة الحالية.

مشكلة الدراسة وأسئلتها

إن الركيزة الأساسية التي يتوقف عليها استخدام نظرية الاستجابة للفقرة هي قضية التقدير الإحصائي لمعالم الفقرات وقدرة الأفراد، حيث تعتمد دقة هذا التقدير على كثير من العوامل الذي اهتم البحث السيكومتري بدراستها من أهمها طرق التقدير، ولأن هذه الطرق جميعها تعتمد التحليل العددي، فقد تم تصميم برامج حاسوبية متنوعة لإيجاد مثل هذه التقديرات، حيث تباينت وجهة نظر الباحثين حول الطريقة الأفضل في دقة التقديرات وبنفس الوقت دراسة العوامل الأخرى المؤثرة على دقة التقديرات،

منها حجم العينة وطول الاختبار، بالإضافة إلى انتهاك افتراضات النموذج المستخدم. ويتضح بأن غالبية الدراسات التي اهتمت بتقييم دقة التقديرات اعتمدت على مؤشرات إحصائية وصفية، حيث بات من الضروري معرفة دلالة الفروق للخصائص السيكومترية لكل من معالم الفقرات وقدرات الأفراد باختلاف العوامل المؤثرة على دقة التقديرات، وبناء عليه فإن هذه الدراسة حاولت تقصي دقة طرق تقديرات معالم الفقرات وقدرات الأفراد في برنامج بايلوج تحت شروط مختلفة (طول الاختبار، حجم العينة، وطريقة التقدير المستخدمة في برنامج بايلوج) من خلال محاكاة الواقع للمواقف الاختبارية المختلفة؛ من أجل تقديم أدلة إمبريقية عن حجم العينة وطول الاختبار والطريقة المناسبة للوصول إلى أدق التقديرات عند استخدام برنامج Bilog-Mg في تحليل بيانات الاختبارات والمقاييس النفسية والتربوية . لذا حاولت هذه الدراسة على وجه التحديد الإجابة عن السؤالين الآتين :-

١. هل تختلف دقة تقديرات معالم الفقرات (الصعوبة، والتمييز، والتخمين) المعايرة بطريقة الأرجحية العظمى باختلاف طول الاختبار وحجم العينة والتفاعل بينهما ؟

٢. هل تختلف دقة تقديرات معلمة القدرة للفرد باختلاف طول الاختبار وحجم العينة وطريقة التقدير المستخدمة (الأرجحية العظمى، وتوقع الاقتران البعدي وتعظيم الاقتران البعدي) والتفاعل بينها؟

أهمية الدراسة

الأهمية النظرية: تكمن الأهمية النظرية للدراسة في محاولتها الكشف عن دقة تقديرات طرق التقدير المستخدمة في برنامج بايلوج لمعالم الفقرات وقدرات الأفراد في ضوء تغير طول الاختبار وحجم العينة والتفاعل بينهما، وذلك بالاعتماد على التصاميم التجريبية من خلال أسلوب المحاكاة من أجل تحديد طريقة التقدير المستخدمة في برنامج بايلوج والتي تؤدي إلى دقة عالية في تقديرات معالم الفقرات وقدرات الأفراد تحت شروط تغير طول الاختبار وحجم العينة نظرا لشيوع استخدام البرنامج عالميا في تقدير معالم الفقرات وقدرات الأفراد. إضافة إلى ذلك قلة الدراسات العربية بشكل عام – في حدود علم الباحث- التي اهتمت بدراسة طرق التقدير المستخدمة في برامج نظرية الاستجابة للفقرة، مما قد يسهم في الإضافةإلى الأدب النظري في هذا المجال عربيا.

الأهمية العملية: تنبثق الأهمية العملية للدراسة من خلال مساهمتها في توفير أدلة إمبريقية من خلال أساليب الإحصاء الاستدلالي للوصول إلى الطريقة الأفضل في تقدير معالم النموذج المستخدم عند استخدام إحدى طرق التقدير السائدة في برنامج Bilog-Mg التي يمكن استخدامها من قبل الباحثين في الدراسات والأبحاث خصوصا التي تهتم ببناء وتطوير الاختبارات والمقاييس التربوية والنفسية، كذلك يمكن أن تخدم نتائج الدراسة مراكز القياس والتقويم التربوي في المؤسسات الحكومية والخاصة التي تهتم بتحليل بيانات الاختبارات والمقاييس النفسية والتربوية.

مصطلحات الدراسة

معالم الفقرات: وهي معالم الصعوبة، والتمييز، والتخمين المنبثقة عن النموذج اللوجستي ثلاثي المعلمة.

دقة التقدير: وهو مصطلح يشير إلى جودة التقدير وذلك بالاعتماد على الجذر التربيعي لمتوسطات مربعات الانحرافات للفروق بين المعالم المقدرة والمعالم الحقيقية.

طريقة الأرجحية العظمى: هي إحدى الطرق المستخدمة في برنامج Bilog-Mg لتقدير معالم الفقرة (الصعوبة، التمييز، التخمين) وقدرة الأفراد التي تعمل على تعظيم اقتران الاحتمالية.

طريقة بييز: وهي إحدى الطرق المستخدمة في برنامج Bilog-Mg لتقدير معلمة القدرة للفرد،

وتتم وفق طريقتين هما: طريقة توقع الاقتران البعدي وطريقة تعظيم الاقتران البعدي.

تصميم الدراسة

لتحقيق أهداف الدراسة الحالية في الكشف عن دقة تقديرات معالم الفقرات وتقديرات القدرة للأفراد في برنامج Bilog-MG باختلاف طول الاختبار وحجم العينة وطريقة التقدير المستخدمة، اتبعت الدراسة الحالية التصميم التجريبي العاملي (X 4 X 3 4) بالاعتماد على المحاكاة، حيث كانت المتغيرات المستقلة في التصميم كلا من حجم العينة، وقد تم استخدام أربعة مستويات من حجم العينة (٢٥٠، ٥٠٠، ١٠٠٠، ١٥٠٠)، ويعد مثل هذا الحجم من العينات مناسبا للظروف التطبيقية للاختبارات في الواقع العملي، حيث تدرجت من العينات الصغيرة إلى المتوسطة وأخيرا كبيرة الحجم. وفي الوقت نفسه تم استخدام أربعة مستويات من طول الاختبار (١٠، ٣٠، ٦٠، ٨٠) متغيرا مستقلا ثان، وتعد مثل هذه الأطوال مناسبة لأطوال الاختبارات المستخدمة في الواقع العملي سواء كانت قصيرة أم متوسطة أم طويلة. أما المتغير المستقل الثالث فكان طريقة التقدير المستخدمة في برنامج بايلوج وله ثلاث مستويات (الارجحية العظمى، توقع الاقتران البعدي، تعظيم الاقتران البعدي)

وللحكم على دقة التقدير، فقد اعتمد الباحث محك الجذر التربيعي لمتوسط مربعات الانحرافات للفروق بين المعالم الحقيقية والمقدرة متغيرا تابعا Root Mean Standard Error (RMSE) الذي يعطى بالعلاقة الآتية:

$$RMSE = \sqrt{\frac{\sum_{I=1}^{K}(\pi_i - \hat{\pi}_i)^2}{K}} \quad \text{.....5}$$

حيث تشير π_i إلى المعلمة الحقيقة (الصعوبة، والتمييز، والتخمين، والقدرة) والرمز $\hat{\pi}_i$ يشير إلى المعالم المقدرة (الصعوبة، والتمييز، والتخمين، ومعلمة القدرة) من النموذج اللوجستي ثلاثي المعلمة باستخدام طرق التقدير المستخدمة ببرنامج بايلوج، ويشير الرمز K إلى عدد مرات التكرار في توليد البيانات. ويعتبر مؤشر RMSE من أهم المؤشرات الإحصائية للتحقق من مدى انسجام البيانات مع النموذج الرياضي المستخدم (Hambleton, Swaminathan & Rogers, 1991)، ويمكن كذلك استخدامه كمؤشر للوقوف على مدى تباين المعالم المقدرة عن المعالم الحقيقة، وهو من المؤشرات الإحصائية الهامة التي أوصى بها هارويل (Harwell, 1997) باعتباره وحدة معيارية يعكس مدى ابتعاد المعالم الحقيقة عن المعالم المقدرة، حيث يمكن الحكم على دقة التقديرات من خلال قيمته فكلما كانت منخفضة دل ذلك على دقة عالية من التقدير؛ بسبب التجانس العالي بين هذه القيم.

توليد البيانات

تم توليد البيانات باستخدام طريقة المحاكاة التي تعد من الناحية العملية هامة، حيث تتيح المجال للحصول على بيانات من مواقف إختبارية مختلفة وتحت شروط مختلفة لإيجاد حلول لمشكلات إحصائية سيكومترية يصعب الوصول إليها بالمواقف العملية المختلفة أثناء تطبيق الاختبارات والمقاييس النفسية والتربوية، بالإضافة إلى ذلك يمكن الحصول على بيانات من توزيعات طبيعية من أجل الوصول إلى معلومات غير متحيزة، وضبط الخطأ العيني (Lord, 1975; Wilcox, 1988). وسيتم توليد البيانات باستخدام برنامج WINGEN، وهو من تصميم وإنتاج هان وهامبيلتون (Han & Hambleton, 2007). وقد تمت عملية توليد البيانات وفق المراحل الآتية:

- توليد معلمة الصعوبة للفقرات وفق التوزيع الطبيعي (Normal~ 0 , 1)، وهذا ينتج فقرات متباينة في الصعوبة تتراوح ما بين (-٣ و٣) وهذا يتطابق مع التوزيع القبلي لمعلمة القدرة المستخدم في برنامج بايلوج (Harwell & Baker, 1991).
- تم توليد معلمة التخمين للفقرات وفق توزيع بيتا

(Beta ~ 8, 32) وهذا التوزيع يحاكي التوزيع المسبق لمعلمة التخمين (Baker & Kim, 2004) وينتج قيم لمعلمة التخمين تماثل قيم التخمين للاختبار الاختيار من متعدد المؤلف من خمسة بدائل.

المرحلة الثانية: توليد استجابات المفحوصين وفق التوزيع الطبيعي (Normal ~ 0 , 1) بواقع ١٠٠ مرة لكل خلية من خلايا التصميم التجريبي باستخدام القيم نفسها للمعالم الحقيقة للفقرات التي تم توليدها في المرحلة الأولى ليصبح عدد البيانات التي تم توليدها (4X4X3X100)، وقد تم اختيار عدد مرات التكرار (number of replication) ١٠٠ حيث أشار هارويل وزملاؤه (Harewell, Stone, Hsu & Kirsci, 1996) بأن دقة التقديرات تزداد بزيادة عدد مرات التكرار بحيث تصبح دقة التقديرات أكثر استقراراً، وهو ما أكده كماتو (Kamata, 1998) من خلال مراجعته لدراسات المحاكاة، حيث وجد بأن عدد مرات التكرار التي استخدمت في ذلك النوع من الدراسات كانت تتراوح ما بين ٥ إلى ١٠٠ مرة، وأن التقديرات كانت تستقر عندما يزيد عدد مرات التكرار عن ٥٠.

تحليل البيانات

من أجل تحقيق الهدف من الدراسة والإجابة عن أسئلة الدراسة، قام الباحث بتحليل البيانات التي تم توليدها في المرحلة الثانية باستخدام برنامج Bilog-Mg3 لإيجاد معالم الفقرات المقدرة (الصعوبة، التمييز، التخمين) باستخدام طريقة الأرجحية العظمى، ويعود السبب في ذلك لكون البرنامج لا يستخدم إلا هذه الطريقة في تقدير معالم الفقرات. أما بالنسبة إلى تقدير معلمة القدرة للفرد فقد تم تقديرها باستخدام طرق التقدير الثلاث المستخدمة في برنامج بايلوج (الأرجحية العظمى، توقع الاقتران البعدي وتعظيم الاقتران البعدي) (Zimowski, Muraki, ; Mislevey & bock, 200 Rupp, 2003).

وتم حساب قيم RMSE لتقديرات كل من معالم الفقرات حسب طريقة الأرجحية العظمى، وكذلك بالنسبة إلى قدرات الأفراد ولكل طريقة من طرق التقدير المستخدمة لتقدير معلمة القدرة للفرد في برنامج بايلوج من خلال تصميم ملف تنفيذي باستخدام برنامج WINGEN يربط بينه وبين برنامج Bilog-Mg 3 لتحليل البيانات المولدة، وقد تم تخزين قيم RMSE لكل من معالم الفقرات وقدرة الأفراد في ملف خاص واستخدام برنامج SPSS لإيجاد النتائج من خلال استخدام تحليل التباين.

نتائج الدراسة

للإجابة عن السؤال الأول للدراسة تم استخدام طريقة الأرجحية العظمى الهامشية (MML) لتقدير معالم الفقرات (الصعوبة، التمييز، التخمين) لكون البرنامج لا يستخدم إلا هذه الطريقة في تقدير معالم الفقرات، وبنفس الوقت تم إيجاد الجذر التربيعي لمتوسط مربعات الانحرافات للفروق بين المعالم الحقيقية والمقدرة (RMSE) لجميع البيانات المولدة. ويظهر الجدول (١) الوسط الحسابي والانحراف المعياري لقيم RMSE لمعالم الفقرات، باختلاف طول الاختبار وحجم العينة.

جدول ١: الوسط الحسابي والانحراف المعياري لقيم RMSE لمعالم الفقرات باختلاف طول الاختبار وحجم العينة

طول الاختبار	حجم العينة	معالم الفقرات					
		التمييز		الصعوبة		التخمين	
		الوسط الحسابي	الانحراف المعياري	الوسط الحسابي	الانحراف المعياري	الوسط الحسابي	الانحراف المعياري
١٠	٢٥٠	٠,٣١٩	٠,١٠٦	٠,١٨١	٠,٠٤٢	٠,٠٤٣	٠,٠٠٧
	٥٠٠	٠,٢٧٠	٠,٠٦٢	٠,١٣٨	٠,٠٤١	٠,٠٤٤	٠,٠١٠
	١٠٠٠	٠,٢٣٤	٠,٠٥٢	٠,١٣٠	٠,٠٣٧	٠,٠٣٩	٠,٠١٠
	١٥٠٠	٠,٢٢٠	٠,٠٥٣	٠,١٢١	٠,٠٤٠	٠,٠٣٧	٠,٠١١
٣٠	٢٥٠	٠,٣٠٧	٠,٠٤١	٠,١٩٣	٠,٠٣٥	٠,٠٥٦	٠,٠٠٥
	٥٠٠	٠,٢٤٥	٠,٠٣٧	٠,١٤٧	٠,٠٢٣	٠,٠٥٨	٠,٠٠٦
	١٠٠٠	٠,١٩٨	٠,٠٣١	٠,١١٥	٠,٠١٥	٠,٠٥٠	٠,٠٠٥
	١٥٠٠	٠,١٥٩	٠,٠٢٤	٠,٠٩٩	٠,٠١٥	٠,٠٤٥	٠,٠٠٥
٦٠	٢٥٠	٠,٣٢١	٠,٠٧٦	٠,٢٠٥	٠,٠٢٥	٠,٠٥٥	٠,٠٠٥
	٥٠٠	٠,٢٢٤	٠,٠٢٥	٠,١٧٩	٠,٠٢٥	٠,٠٥٢	٠,٠٠٥
	١٠٠٠	٠,١٧٨	٠,٠٢٠	٠,١٤٩	٠,٠١٧	٠,٠٤٨	٠,٠٠٥
	١٥٠٠	٠,١٤٨	٠,٠١٥	٠,١٢٩	٠,٠١٧	٠,٠٤٤	٠,٠٠٤
٨٠	٢٥٠	٠,٢٢٥	٠,٠٣٠	٠,٢٤٤	٠,٠٢٥	٠,٠٥٧	٠,٠٠٤
	٥٠٠	٠,٢٠٨	٠,٠٣٢	٠,٢٢٦	٠,٠٢١	٠,٠٥٦	٠,٠٠٥
	١٠٠٠	٠,١٥٤	٠,٠١٦	٠,١٦٢	٠,٠٢٠	٠,٠٥٢	٠,٠٠٥
	١٥٠٠	٠,١٣٦	٠,٠١٤	٠,١٤٩	٠,٠١٧	٠,٠٤٨	٠,٠٠٤

تشير النتائج الواردة في الجدول (١) إلى أن هناك تباينا ملحوظا في الأوساط الحسابية لقيم RMSE لمعالم الفقرات (التمييز، الصعوبة، التخمين) المعايرة بطريقة الأرجحية العظمى الهامشية (MML) تبعا لطول الاختبار وحجم العينة، حيث أظهرت النتائج بأن أعلى وسط حسابي لقيمة RMSE لمعلمة التمييز كانت (٠,٣٢١) عندما كان حجم العينة يساوي (٢٥٠) وطول الاختبار (٦٠) فقرة، في حين كانت أقل قيمة له (٠,١٣٦) عندما كان طول الاختبار (٨٠) فقرة وحجم العينة (١٥٠٠). أما بالنسبة لمعلمة الصعوبة فقد بينت النتائج الواردة في الجدول نفسه بأن أعلى وسط حسابي لقيمة RMSE بلغت (٠,٣١٩) عندما كان طول الاختبار (١٠) فقرات وحجم العينة (٢٥٠)، وأصبحت اقل قيمة له (٠,٠٩٩) عندما كان طول الاختبار (٨٠) فقرة وحجم العينة (١٥٠٠). وبالنظر إلى قيم الأوساط الحسابية لـ RMSE لمعلمة التخمين يتضح بأن هناك تقاربا في قيم الأوساط الحسابية حيث كانت أعلى قيمة للوسط الحسابي (٠,٠٥٨) عندما كان طول الاختبار (٣٠) فقرة وحجم العينة (٥٠٠)، بينما كانت اقل قيمة له (.٠٣٧) عندما كان طول الاختبار (١٠) فقرات وحجم العينة (١٥٠٠).

وللوقوف على دلالات الفروق بين الأوساط الحسابية لقيم RMSE لمعالم الفقرات المعايرة بطريقة الأرجحية العظمى الهامشية تبعا لطول الاختبار وحجم العينة، فقد تم استخدام تحليل التباين الثنائي لكل معلمة على حدة. ويظهر الجدول (٢) نتائج هذا التحليل.

جدول ٢: نتائج تحليل التباين الثنائي للكشف عن دلالة الفروق بين الأوساط الحسابية لقيم RMSE لمعالم الفقرات (التمييز، الصعوبة ، التخمين) تبعا لطول الاختبار وحجم العينة

المعلمة	مصدر التباين	مجموع المربعات	درجات الحرية	وسط المربعات	قيمة F	الدلالة الإحصائية	الدلالة العملية η^2
التمييز	طول الاختبار	١,٠٦٢	٣	٠,٣٥٤	١٧٧	٠,٠٠٠	٠,٢٣٧
	حجم العينة	٤,٢١١	٣	١,٤٠٤	٧٠٢	٠,٠٠٠	٠,٥٥٢
	التفاعل	٠,١٩٩	٩	٠,٠٢٢	١١,٥	٠,٠٠٠	٠,٠٥٥
	الخطأ	٣,٤١٦	١٥٨٤	٠,٠٠٢			
	الكلي	٨,٨٨٨	١٥٩٩				
الصعوبة	طول الاختبار	٠,٨١٨	٣	٠,٢٧٣	٢٧٣	٠,٠٠٠	٠,٤٠٣.
	حجم العينة	١,٥٧٠	٣	٠,٥٢٣	٥٢٣	٠,٠٠٠	٠,٥٦٤
	التفاعل	٠,١٤٠	٩	٠,٠١٦	١٦	٠,٠٠٠	٠,١٠٣
	الخطأ	١,٢١٥	١٥٨٤	٠,٠٠١			
	الكلي	٣,٧٤٣	١٥٩٩				
التخمين	طول الاختبار	٠,٠٤٠	٣	٠,٠١٣	٣٢٥	٠,٠٠٠	٠,٣٧٩
	حجم العينة	٠,٠٢٥	٣	٠,٠٠٨	٢٠٠	٠,٠٠٠	٠,٢٧٦
	التفاعل	٠,٠٠٢	٩	٠,٠٠٠٢	٥	٠,٠٠٠	٠,٠٢٧
	الخطأ	٠,٠٦٥	١٥٨٤	٠,٠٠٠٠٤			
	الكلي	٠,١٣٠	١٥٩٩				

أظهرت النتائج الواردة في الجدول (٢) أن هناك فروقا ذات دلالة إحصائية عند مستوى الدلالة (α = ٠,٠٥) بين الأوساط الحسابية لقيم الجذر التربيعي لمتوسط مربعات الانحرافات (RMSE) للفروق بين المعالم الحقيقية والمقدرة بطريقة الأرجحية العظمى الهامشية لمعالم الفقرات (التمييز، الصعوبة، التخمين) تعزى لكل من طول الاختبار وحجم العينة والتفاعل بينهما. وبينت النتائج كذلك بأن حجم العينة كان أكثر إسهاما في تباين قيم الجذر التربيعي لمتوسط مربعات الانحرافات (RMSE) للفروق بين المعالم الحقيقة والمقدرة بطريقة الأرجحية العظمى الهامشية لمعالم الفقرات (التمييز، الصعوبة) من طول الاختبار، وذلك من خلال النظر إلى قيم الدلالة العملية (مربع إيتا) حيث أسهم ب (٥٥٪، ٥٦٪) في تباين دقة تقدير معالم الفقرات التمييز، والصعوبة على الترتيب، في حين كان إسهام عامل الطول في تباين دقة تقدير معلمة التخمين أكثر من عامل حجم العينة حيث أسهم بنسبة (٣٨٪) .

وللوقوف على مواقع الفروق بين الأوساط الحسابية لقيم (RMSE) في تقدير معالم الفقرات العائد لعامل طول الاختبار، فقد استخدم اختبار شافيه للكشف عن تلك الفروق لكل معلمة من معالم الفقرات، ويوضح الجدول (٣) هذه الفروق.

جدول ٣: نتائج المقارنات الثنائية بين الأوساط الحسابية لقيم RMSE لمعالم الفقرات حسب عامل طول الاختبار

المعلمة	طول الاختبار	الوسط الحسابي	طول الاختبار			
			١٠	٣٠	٦٠	٨٠
التمييز	١٠	٠,٢٦٠		٠٣٣,٠*	٠٤٢,٠*	٠,٠٧٢*
	٣٠	٠,٢٢٧			٠٠٩,٠*	٠٣٩,٠*
	٦٠	٠,٢١٨				٠٣٠,٠*
	٨٠	٠,١٨٨				
الصعوبة	١٠	٠,١٤٣		٠,٠٠٥	٠,٠٢٢*	٠,٠٥٢*
	٣٠	٠,١٣٨			٠,٠٢٧*	٠,٠٥٧*
	٦٠	٠,١٦٥				٠,٠٣٠*
	٨٠	٠,١٩٥				
التخمين	١٠	٠,٠٤١		٠,٠١١*	٠,٠٠٩*	٠,٠١٢*
	٣٠	٠,٠٥٢			٠,٠٠٢*	٠,٠٠١*
	٦٠	٠,٠٥٠				٠,٠٠٣*
	٨٠	٠,٠٥٣				

* دال إحصائيا عند مستوى الدلالة ٠,٠٥ = α

يتضح من النتائج الواردة في الجدول (٣) أن الفروق بين الأوساط الحسابية لقيم RMSE لمعلمة التمييز كانت دالة إحصائيا ولجميع أطوال الاختبار، وكانت جميعها لصالح الاختبار الأطول، حيث يتضح بأن قيمة الوسط الحسابي لدقة تقدير معلمة التمييز (RMSE) كانت الأعلى عندما كان طول الاختبار (٨٠) فقرة، وكانت تقل هذه الدقة بنقصان طول الاختبار لتصبح الأقل عندما كان طول الاختبار (١٠) فقرات. وعلى الرغم من دلالة الفروق بين الأوساط الحسابية لدقة تقدير معلمة التمييز عند الأطوال (٣٠، ٦٠، ٨٠) فقرة إلا أن استخدام طريقة (MML) في تقدير معلمة التمييز للفقرات قد عكست دقة عالية عند هذه الأطوال. وتشير كذلك نتائج المقارنات الواردة في الجدول (٣) بين الأوساط الحسابية لدقة تقدير معلمة الصعوبة بأن الفروق كانت دالة إحصائيا عند أطوال الاختبارات (٣٠، ٦٠، ٨٠) فقرة، وكانت دقة التقدير في معلمة الصعوبة لصالح الاختبار الذي طوله (٣٠) فقرة وأصبحت الأعلى عندما كان طول الاختبار (٨٠) فقرة، بينما لم تكن الفروق دالة إحصائيا بين الاختبار الذي طوله (١٠) فقرات والاختبار الذي طوله (٣٠) فقرة. وبينت النتائج الواردة في الجدول نفسه بأن الفروق بين الأوساط الحسابية في دقة تقدير معلمة التخمين كانت جميعها دالة إحصائيا وجاءت جميعها لصالح الاختبار الذي طوله (١٠) فقرات حيث كان الأعلى في دقة تقدير معلمة التخمين في حين كانت هذه الدقة تتناقص بزيادة طول الاختبار لتصبح الأقل عندما كان طول الاختبار (٨٠) فقرة.

أما بالنسبة إلى الكشف عن الفروق بين الأوساط الحسابية لدقة تقدير معالم الفقرات والعائد لعامل حجم العينة فقد تم استخدام أسلوب شافيه للكشف عن دلالة تلك الفروق، ويظهر الجدول (٤) هذه الدلالات.

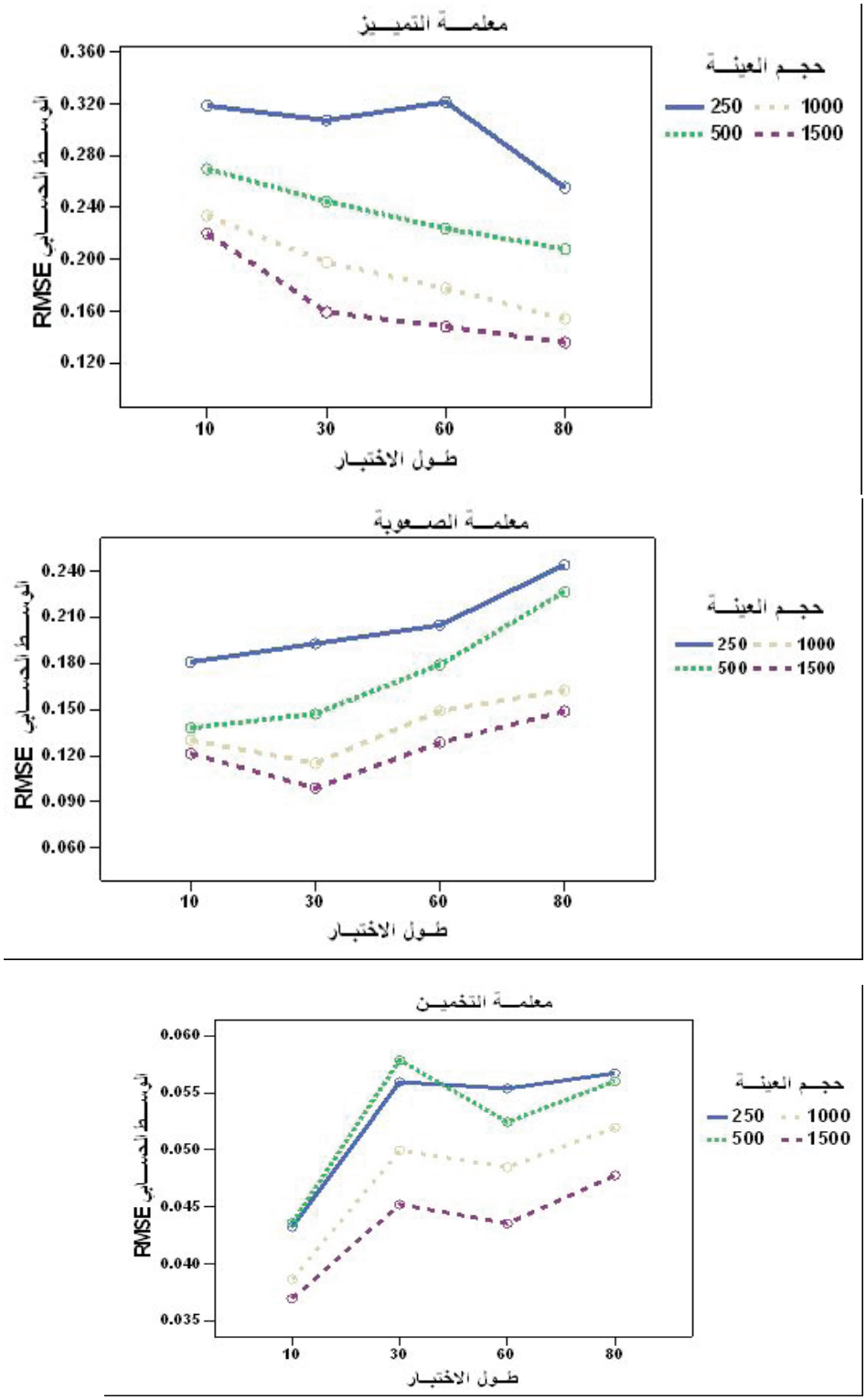
جدول ٤: نتائج المقارنات الثنائية بين الأوساط الحسابية لقيم RMSE لمعالم الفقرات حسب عامل حجم العينة

المعلمة	حجم العينة	الوسط الحسابي	حجم العينة			
			٢٥٠	٥٠٠	١٠٠٠	١٥٠٠
التمييز	٢٥٠	٠,٣٠١		٠٦٥,٠*	١١٠,٠*	٠,١٣٥*
	٥٠٠	٠,٢٣٦			٠٤٥,٠*	٠٧٠,٠*
	١٠٠٠	٠,١٩١				٠٢٥,٠*
	١٥٠٠	٠,١٦٦				
الصعوبة	٢٥٠	٠,٢٠٥		٠,٠٣٢*	٠,٠٦٦*	٠,٠٨١*
	٥٠٠	٠,١٧٣			٠,٠٣٤*	٠,٠٤٩*
	١٠٠٠	٠,١٣٩				٠,٠١٥*
	١٥٠٠	٠,١٢٤				
التخمين	٢٥٠	٠,٠٥٣		٠,٠٠١	٠,٠٠٦*	٠,٠١٠*
	٥٠٠	٠,٠٥٢			٠,٠٠٥*	٠,٠٠٩*
	١٠٠٠	٠,٠٤٧				٠,٠٠٤*
	١٥٠٠	٠,٠٤٣				

* دال إحصائيا عند مستوى الدلالة ٠,٠٥ = α

تكشف نتائج المقارنات الثنائية الواردة في الجدول (٤) بين الأوساط الحسابية لدقة تقدير معلمتي التمييز والصعوبة بأن الفروق بين هذه الأوساط جاءت جميعها دالة إحصائيا ولصالح حجم العينة (١٥٠٠)، حيث كانت الأعلى في دقة التقدير، وتناقصت دقة تقدير معلمتي التمييز والصعوبة بتناقص حجم العينة لتصبح الأقل دقة عندما كان حجم العينة (٢٥٠). ويتضح كذلك بأن أعلى دقة تقدير لمعالم الفقرات (التمييز، الصعوبة، التخمين) المعايرة بطريقة (MML) كانت الأعلى عندما كان حجم العينة أكبر من (١٠٠٠) وبالتالي يشير إلى مناسبة مثل هذه الحجم للعينات لمعايرة فقرات الاختبار باستخدام طريقة (MML) للوصول إلى دقة أعلى في تقدير معالم الفقرات خصوصا عند استخدام النموذج اللوجستي تلاثي المعلمة. وفي الوقت نفسه يتضح كذلك من النتائج الواردة في الجدول نفسه أن دقة تقدير معلمة التخمين كانت تزداد بزيادة حجم العينة، وهي بذلك جاءت مماثلة لمعلمتي التمييز والصعوبة، إلا أن الفروق بين الأوساط الحسابية لدقة تقدير معلمة التخمين لم تكن دالة إحصائيا عندما كان حجم العينات (٢٥٠، ٥٠٠)، في حين كانت دالة إحصائيا عند بقية الأحجام من العينات.

وتجدر الإشارة كذلك إلى أن نتائج تحليل التباين الثنائي الواردة في الجدول (٢) أشارت إلى وجود اثر لتفاعل كل من طول الاختبار وحجم العينة في دقة تقدير معالم الفقرات (التمييز، والصعوبة، والتخمين) بمعنى أن هناك تأثيرا مشتركا لكل من العاملين، وتم تمثيل هذا التفاعل بيانيا لكل معلمة على حدة كما هو موضح في الشكل رقم (١)



الشكل ١: التمثيل البياني للتفاعل بين طول الاختبار وحجم العينة في دقة تقدير معالم الفقرات (التمييز، والصعوبة، والتخمين)

يتضح من الشكل (١) بشكل عام بأن دقة تقدير معلمة التمييز كانت تزداد بزيادة حجم العينة وطول الاختبار بشكل رتبي، ويتضح كذلك بأن هناك تفاوتا واضحا في دقة تقدير معلمة التمييز للفقرة عندما كان حجم العينة (٢٥٠) حيث كانت الأقل دقة مقارنة مع المستويات الأخرى لحجم العينة. ويتضح من الشكل (١) كذلك أن دقة تقدير معلمة الصعوبة كانت تزداد بزيادة حجم العينة عندما تكون أطوال الاختبارات (٣٠، ٦٠) فقرة بشكل رتبي، وكانت الأعلى عندما يكون الاختبار طوله (٣٠) فقرة. أما بالنسبة إلى دقة تقديرات معلمة التخمين فقد تبين من الشكل (١) بأن أعلى دقة تقدير كانت عندما كان الاختبار طوله (١٠) فقرات يليه الاختبار الذي طوله (٦٠) فقرة. وتجدر الإشارة كذلك بأن طريقة الأرجحية العظمى الهامشية (MML) المستخدمة في برنامج بايلوج كانت أكثر دقة في تقدير معالم الفقرات عندما يكون حجم العينة (٥٠٠) فأكثر حيث تتأثر دقة تقديرات معالم الفقرات بحجم العينة أكثر من طول الاختبار، وهذا ما أظهرته نتائج تحليل التباين الثنائي الواردة في الجدول (٢) من خلال مؤشر الدلالة العملية مربع إيتا.

وللإجابة عن السؤال الثاني لمعرفة أثر كل من طول الاختبار وحجم العينة، وطريقة التقدير المستخدمة في برنامج بايلوج على دقة تقدير معلمة القدرة للفرد (θ) والتفاعل بينها، فقد تم أيضا إيجاد الأوساط الحسابية والانحرافات المعيارية لدقة تقدير معلمة القدرة للفرد (RMSE) لكل طول من أطوال الاختبار ولكل مستوى من حجم العينة حسب طريقة التقدير المستخدمة في برنامج بايلوج، ويوضح الجدول (٥) هذه الأوساط الحسابية والانحرافات المعيارية.

جدول ٥: الوسط الحسابي والانحراف المعياري لدقة تقدير معلمة القدرة للفرد (θ) باختلاف طول الاختبار وحجم العينة وطريقة التقدير المستخدمة في برنامج بايلوج

طول الاختبار	حجم العينة	طريقة التقدير: الأرجحية العظمى MML		توقع الاقتران EAP		تعظيم الاقتران MAP	
		الوسط الحسابي	الانحراف المعياري	الوسط الحسابي	الانحراف المعياري	الوسط الحسابي	الانحراف المعياري
١٠	٢٥٠	٠,٦٤٨	٠,٠٦٠	٠,٤٨٩	٠,٠٢٠	٠,٤٩١	٠,٠٢١
	٥٠٠	٠,٦٢٥	٠,٠٣٥	٠,٥٠٢	٠,٠١٢	٠,٤٩٩	٠,٠١٥
	١٠٠٠	٠,٦١٧	٠,٠٢٦	٠,٥١٨	٠,٠٠٩	٠,٥٠٢	٠,٠١١
	١٥٠٠	٠,٦٢١	٠,٠٢٥	٠,٥١٤	٠,٠٠٩	٠,٥٠٠	٠,٠٠٩
٣٠	٢٥٠	٠,٤٣١	٠,٠٢٦	٠,٣٢٢	٠,٠١٥	٠,٣٤١	٠,٠١٧
	٥٠٠	٠,٤٢٠	٠,٠١٧	٠,٣٤٦	٠,٠٠٩	٠,٣٦٣	٠,٠١١
	١٠٠٠	٠,٤٠٠	٠,٠١٣	٠,٣٢٣	٠,٠٠٧	٠,٣٤٢	٠,٠٠٨
	١٥٠٠	٠,٣٨٦	٠,٠١١	٠,٣٣٤	٠,٠٠٦	٠,٣٤٤	٠,٠٠٥
٦٠	٢٥٠	٠,٣٢٤	٠,٠٣٠	٠,٢٦٥	٠,٠٠٩	٠,٣١٣	٠,٠١١
	٥٠٠	٠,٣٣٧	٠,١١٤	٠,٢٤٣	٠,٠٠٨	٠,٢٨٧	٠,٠٠٩
	١٠٠٠	٠,٣٢٤	٠,٠١٧	٠,٢٤٢	٠,٠٠٦	٠,٢٩٢	٠,٠٠٦
	١٥٠٠	٠,٣١٥	٠,٠١٣	٠,٢٣٨	٠,٠٠٥	٠,٢٨٧	٠,٠٠٥
٨٠	٢٥٠	٠,٢٨١	٠,٠٢١	٠,٢٣٦	٠,٠١٠	٠,٢٨٨	٠,١١
	٥٠٠	٠,٢٨٥	٠,٠١٥	٠,٢٤٢	٠,٠٠٧	٠,٢٩٠	٠,٠٠٨
	١٠٠٠	٠,٢٦٦	٠,٠١٠	٠,٢٢٣	٠,٠٠٦	٠,٢٨٢	٠,٠٠٦
	١٥٠٠	٠,٢٦١	٠,٠٠٩	٠,٢٣٠	٠,٠٠٤	٠,٢٨٣	٠,٠٠٤

تشير النتائج الواردة في الجدول (٥) بأن هناك تباينا ملحوظا بين الأوساط الحسابية لدقة تقدير معلمة القدرة (RMSE) (θ) باختلاف طول الاختبار وحجم العينة وطريقة التقدير المستخدمة في تقدير معلمة القدرة للفرد، حيث يتضح بشكل عام بأن دقة تقدير معلمة القدرة للفرد (θ) كانت تزداد بزيادة طول الاختبار وحجم العينة من خلال استخدام جميع طرق التقدير في برنامج بايلوج، وفي الوقت نفسه كانت أعلى دقة تقدير لمعلمة القدرة (θ) عند استخدام طريقة بييز (طريقة توقع الاقتران EAP)، وللكشف عن دلالة الفروق بين الأوساط الحسابية لدقة تقدير معلمة القدرة للفرد باختلاف طول الاختبار وحجم العينة وطريقة التقدير المستخدمة في برنامج بايلوج والتفاعل بينها، تم استخدام تحليل التباين الثلاثي، حيث يبين الجدول (٦) نتائج هذا التحليل.

جدول ٦: نتائج تحليل التباين الثلاثي للكشف عن دلالة الفروق بين الأوساط الحسابية لدقة تقدير معلمة القدرة (θ) حسب طول الاختبار وحجم العينة وطريقة التقدير المستخدمة في برنامج بايلوج

مصدر التباين	مجموع المربعات	درجات الحرية	وسط المربعات	قيمة F	الدلالة الإحصائية	الدلالة العملية η^2
حجم العينة	٠,١٠٣	٣	٠,٠٣٤	٣٤	٠,٠٠٠	٠,٠٣٩
طول الاختبار	٥٧,٥٤٥	٣	١٩,١٨٢	١٩١٨٢	٠,٠٠٠	٠,٩٥٧
طريقة التقدير	٥,٢٤١	٢	٢,٦٢١	٢٦٢١	٠,٠٠٠	٠,٦٧٢
حجم العينة X طول الاختبار	٠,١٢٢	٩	٠,٠١٤	١٤	٠,٠٠٠	٠,٠٤٢
حجم العينة X طريقة التقدير	٠,٠٨٢	٦	٠,٠١٤	١٤	٠,٠٠٠	٠,٠٣١
طول الاختبار X طريقة التقدير	٢,٢١٤	٦	٠٣٦٩	٣٦٩	٠,٠٠٠	٠,٤٦٣
حجم العينة X طول الاختبار X طريقة التقدير	٠,١٨٣	١٨	٠,٠١٠	١٠	٠,٠٠٠	٠,٠٦٧
الخطأ	٢,٥٦٤	٤٧٥٢	٠,٠٠١			
الكلي	٦٨,٠٥٤	٤٧٩٩				

تكشف النتائج الواردة في الجدول (٦) أن هناك فروقا ذات دلالة إحصائية عند مستوى الدلالة (α = ٠,٠٥) بين الأوساط الحسابية لدقة تقدير معلمة قدرة الفرد (θ) تعزى لكل من طول الاختبار وحجم العينة وطريقة التقدير والتفاعل بينها. وتظهر النتائج كذلك بأن طول الاختبار وطريقة التقدير قد أسهما في تباين قيم دقة تقدير معلمة القدرة للفرد (θ) أكثر من حجم العينة، حيث كانت قيم الدلالة العملية (مربع إيتا) لكل منهما على التوالي (٦٧,٢٪، ٩٥,٧٪)، والذي يؤكد أهمية عامل طول الاختبار، بالإضافة إلى الطريقة المستخدمة في دقة تقدير معلمة القدرة للفرد. وتجدر الإشارة كذلك إلى أنه تم استخدام اختبار شيفيه للكشف عن مواقع الفروق بين الأوساط الحسابية لدقة تقدير معلمة القدرة للفرد لكل عامل من العوامل الثلاثة على حدة، ويظهر الجدول (٧) مواقع تلك الفروق.

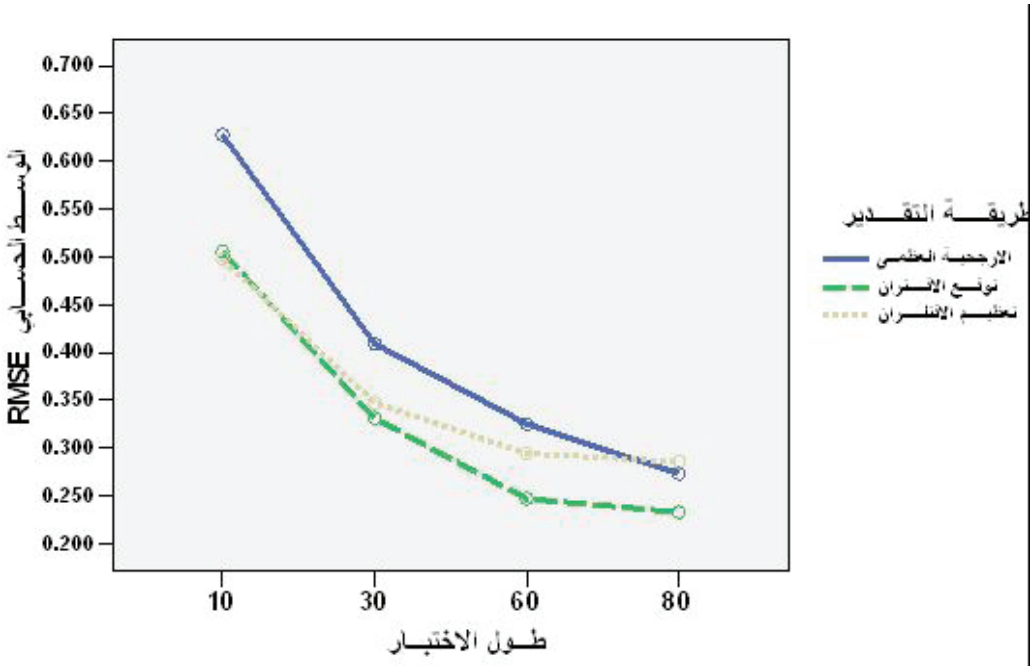
جدول ٧: نتائج المقارنات الثنائية بين الأوساط الحسابية لقيم دقة تقدير معلمة قدرة الفرد حسب طول الاختبار و حجم العينة وطريقة التقدير

العامل	طول الاختبار	الوسط الحسابي	طول الاختبار ١٠	٣٠	٦٠	٨٠
طول الاختبار	١٠	٠,٥٤٤		٠,١٨١*	٠,٢٥٥*	٠,٢٨٠*
	٣٠	٠,٣٦٣			٠,٠٧٤*	٠,٠٩٩*
	٦٠	٠,٢٨٩				٠,٠٢٥*
	٨٠	٠,٢٦٤				
حجم العينة	حجم العينة	الوسط الحسابي	حجم العينة ٢٥٠	٥٠٠	١٠٠٠	١٥٠٠
	٢٥٠	٠,٣٦٩		٠,٠٠١	٠,٠٠٨*	٠,٠١٠*
	٥٠٠	٠,٣٧٠			٠,٠٠٩*	٠,٠١١*
	١٠٠٠	٠,٣٦١				٠,٠٠٢
	١٥٠٠	٠,٣٥٩				
طريقة التقدير	الطريقة	الوسط الحسابي	طريقة التقدير MML	EAP	MAP	
	MML	٠,٤٠٩		٠,٠٨٠*	٠,٠٥٣*	
	EAP	٠,٣٢٩			٠,٠٢٧*	
	MAP	٠,٣٥٦				

* دال إحصائيا عند مستوى الدلالة ٠,٠٥ = α

أظهرت نتائج المقارنات الثنائية الواردة في الجدول (٧) بين الأوساط الحسابية لدقة تقدير معلمة القدرة للفرد والمتعلقة بعامل طول الاختبار بأن هناك فروقا دالة إحصائيا بين الوسط الحسابي لدقة تقدير معلمة قدرة الفرد عند مستوى طول الاختبار (١٠) فقرات، وبين الأوساط الحسابية لأطوال الاختبار (٣٠، ٦٠، ٨٠، فقرة) حيث جاءت هذه الفروق لصالح الاختبار المكون من (٨٠) فقرة بوسط حسابي (٠,٢٦٤)، وأشارت أيضا نتائج المقارنات الثنائية الواردة في الجدول نفسه إلى أن الفروق بين الأوساط الحسابية لدقة تقدير معلمة القدرة للفرد كانت دالة إحصائيا بين مستويين لطول الاختبار، ودائما كان الفرق لصالح الاختبار الأطول. وكشفت نتائج المقارنات الثنائية بين الأوساط الحسابية لدقة تقدير معلمة القدرة للفرد والمتعلقة بعامل حجم العينة بأن الفروق لم تكن دالة إحصائيا عندما كان حجم العينة (٢٥٠، ٥٠٠)، في حين كانت الفروق دالة إحصائيا بين كل مستوى من مستويات حجم العينة (٢٥٠، ٥٠٠)، وبقية المستويات لحجم العينة (١٠٠٠، ١٥٠٠)، وجاءت هذه الفروق لصالح حجم العينة (١٥٠٠) ولم تكن الفروق دالة إحصائيا بين مستويات حجم العينة (١٠٠٠) و(١٥٠٠). أما بالنسبة لطريقة التقدير المستخدمة في تقدير معلمة القدرة للفرد، فقد أظهرت نتائج المقارنات الثنائية بأن الفروق بين الأوساط الحسابية جاءت جميعها دالة إحصائيا بين الطرق الثلاث، ولصالح طريقة بييز (توقع الاقتران EAP).

وبالرجوع إلى نتائج الجدول (٦) فقد أظهرت النتائج بأن التفاعل بين العوامل الثلاثة كان دالاً إحصائياً، وبينت النتائج كذلك بأن التفاعل ما بين طول الاختبار وطريقة التقدير كان له إسهاماً كبيراً بالنسبة لدقة تقدير معلمة القدرة للفرد، حيث بلغت قيمة مؤشر الدلالة العملية (مربع إيتا) للتفاعل بينهما (٤٦,٣٪). ولأن التفاعل بين طول الاختبار وطريقة التقدير كان له الإسهام الأكبر في دقة تقدير معلمة القدرة للفرد فقد تم تمثيل التفاعل بينهما، ويظهر الشكل (٢) التفاعل بين عامل الطول وطريقة التقدير.



الشكل ٢: التمثيل البياني للتفاعل بين طول الاختبار وطريقة التقدير في دقة تقدير معلمة قدرة الفرد

يتضح من الشكل (٢) بأن دقة تقدير معلمة القدرة للفرد كانت الأعلى عندما استخدمت طريقة توقع الاقتران (EAP) ولجميع مستويات أطوال الاختبار في تقدير معلمة القدرة للفرد، وكانت تزداد هذه الدقة بزيادة طول الاختبار، وخصوصا عندما يزيد طول الاختبار عن (٣٠) فقرة، ويتضح كذلك من الشكل نفسه بأن دقة تقدير معلمة القدرة للفرد باستخدام طريقة الأرجحية العظمى كانت دائما الأقل مقارنة مع طرق التقدير المستخدمة في برنامج بايلوج لمعلمة القدرة للفرد وعند جميع مستويات الاختبار، إلا أنها ازدادت دقة تقديرها لمعلمة القدرة للفرد عندما أصبح طول الاختبار (٦٠) فقرة فأعلى، ومثل ذلك يشير بشكل عام بأن الطرق التي تستخدم أسلوب بييز في برنامج بايلوج في تقدير معلمة القدرة للفرد كانت تنتج تقديرات أعلى دقة من طريقة الأرجحية العظمى بشكل عام.

مناقشة النتائج

أشارت النتائج المتعلقة بالسؤال الأول إلى أن التغير في طول الاختبار وحجم العينة يؤثران في دقة تقدير معالم الفقرات (التمييز، والصعوبة، والتخمين) المعبر عنه بالجذر التربيعي لمتوسط مربعات الانحرافات للفروق بين معالم الفقرات الحقيقة والمقدرة (RMSE) بطريقة الأرجحية العظمى الهامشية (MML) المستخدمة في برنامج بايلوج، حيث لوحظ التحسن الملحوظ في دقة تقديرات معالم الفقرات (التمييز، الصعوبة، التخمين) بزيادة طول الاختبار وحجم العينة لكل من معلمة التمييز ومعلمة الصعوبة من خلال تتبع قيم الأوساط الحسابية لمؤشر دقة التقديرات (RMSE) الواردة في الجدول (١) وذلك بتناقص قيمها. أما بالنسبة إلى دقة تقدير معلمة التخمين فقد كانت الأعلى عندما كان الاختبار قصيرا (١٠ فقرات) عند مستويات حجم العينة (١٠٠٠، ١٥٠٠) واستقرت عندما أصبح طول الاختبار (٣٠) فقرة فأعلى، ومثل هذه النتيجة تؤكد فعالية طريقة الأرجحية العظمى المستخدمة في برنامج بايلوج في تقدير معالم الفقرات عند استخدام

النموذج اللوجستي ثلاثي المعلمة، وجاءت متفقة مع نتائج الدراسات المماثلة (الشريفين، ٢٠١٢ Huang et al.، 2001; Gau & Chen, 2005) وكذلك جاءت متفقة مع نتائج دراسات اهتمت بدراسة أثر طول الاختبار وحجم العينة على دقة تقديرات معالم الفقرات مثل (Hulin, Lisak & Drasgow, 1982; Baur & Lukes, 2009).

وتوصلت الدراسة الحالية كذلك إلى وجود الأثر المشترك لكل من طول الاختبار وحجم العينة في دقة تقدير معالم الفقرات باستخدام طريقة الأرجحية العظمى كما هو وارد في الشكل (١)، حيث كانت تقديرات معالم الفقرات تميل إلى الاستقرار عندما يزيد طول الاختبار عن (٣٠) فقرة وحجم العينة عن (٥٠٠)، ويعود السبب في تأثير هذين العاملين على دقة تقديرات معالم الفقرات عند استخدام برنامج بايلوج الذي يعتمد طريقة الأرجحية العظمى إلى الأسلوب الرياضي الذي تعتمده هذه الطريقة، إذ تستخدم هذه الطريقة في تقدير معالم الفقرات التكامل العددي (Numerical Integration) الذي يعرف بتربيع جاوس (Gaussian Quadrature) من أجل تقريب وتسهيل عملية التكامل، ولتحقيق ذلك فإن برمجية بايلوج واعتمادا على نقاط تربيع جاوس تستخدم أسلوب العرض البياني من خلال المدرج التكراري؛ مما يجعل طول الاختبار وحجم العينة عاملين مهمين في تقديرات معالم الفقرات، ولأن القيمة الافتراضية في برمجية بايلوج لعدد نقاط تربيع جاوس هي (١٥) التي اعتمدها الباحث لتثبيت هذا العامل، والذي بدوره أصبح محددا من محددات تعميم نتائج الدراسة الحالية، تصبح الحاجة ماسة لإجراء دراسة مستقبلية بإضافة عامل عدد نقاط التربيع بالإضافة إلى طول الاختبار وحجم العينة وانعكاس ذلك على دقة تقديرات معالم الفقرة عند استخدام برمجية بايلوج في معايرة الفقرات.

ومن الجدير بالذكر كذلك توصل الدراسة الحالية إلى أهمية عامل حجم العينة في دقة تقدير معالم الفقرات عند استخدام برمجية بايلوج من خلال طريقة الأرجحية العظمى، حيث يبدو ذلك واضحا من خلال مؤشر الدلالة العملية (مربع إيتا) التي بلغت قيمه لمعلمة التمييز، والصعوبة والتخمين (٠,٢٧٦، ٠,٥٦٤، ٠,٥٥٢) على التوالي، وربما يكون السبب في ذلك اعتماد طريقة الأرجحية العظمى المستخدمة في برمجية بايلوج على التوزيع القبلي لقدرات الأفراد، والتوزيع الافتراضي لقدرات الأفراد في برمجية بايلوج هو التوزيع الطبيعي، وكما هو معلوم بأن زيادة حجم العينة تصبح أكثر تمثيلا لخصائص المجتمع المستهدف، الذي بدوره ينعكس على دقة القياس خصوصا في الاختبارات والمقاييس النفسية والتربوية بافتراض أن الخصائص النفسية والتربوية التي تقيسها مثل تلك الاختبارات والمقاييس النفسية والتربوية تتوزع توزيعا طبيعيا بين الأفراد، وبالتالي فإن زيادة حجم العينة يترتب عليه اقتراب توزيع الخاصية المستهدفة من التوزيع الطبيعي، وعندما يتماثل توزيع قدرات الأفراد مع التوزيع القبلي لقدرات الأفراد في برمجية بايلوج تصبح التقديرات الناتجة من طريقة الأرجحية أكثر دقة، لأن تقدير معالم الفقرات يعتمد على مدى ملاءمة التوزيع الفعلي لقدرات الأفراد مع التوزيع القبلي الافتراضي، وتتأثر بمستويات قدرات الأفراد، وهذا بدوره جعل هناك استقرارا في تقدير معالم الفقرات عندما كان حجم العينة (٥٠٠) فأكثر.

وبشكل عام، يتضح من نتائج الدراسة الحالية أنه عند استخدام برمجية بايلوج لمعايرة فقرات الاختبار بطريقة الأرجحية العظمى باستخدام النموذج اللوجستي ثلاثي المعلمة لا بد من الانتباه إلى طول الاختبار وحجم العينة باعتبارهما عاملين مهمين في التأثير على دقة القياس، مما يجعل الحاجة للوصول إلى تقديرات دقيقة لمعالم الفقرات لأن يكون طول الاختبار (٣٠) فقرة فأعلى وحجم العينة أكثر من (٥٠٠)، وكما هو معلوم بأن دقة التقديرات لمعالم الفقرات والمعبر عنها بالأخطاء المعيارية أو دقة القياس مهمة في

سياق نظرية الاستجابة للفقرة، وذلك لتأثير هذه التقديرات على دالة معلومات الاختبار والفقرة التي تعد إحدى ركائز نظرية الاستجابة للفقرة عند توظيفها في التطبيقات العملية للاختبارات والمقاييس النفسية والتربوية.

أما فيما يتعلق بنتائج السؤال الثاني والمتعلق بدقة تقديرات معلمة القدرة للفرد (θ)، فقد أظهرت نتائج تحليل التباين الثلاثي الواردة في الجدول (٦) وجود أثر لكل من طول الاختبار، وحجم العينة، وطريقة التقدير المستخدمة والتفاعل بينها في دقة تقدير معلمة القدرة للفرد، وبتتبع قيم الأوساط الحسابية لدقة تقدير معلمة القدرة للفرد الواردة في الجدول (٥) يتضح بأن هناك نزعة عامة في دقة تقدير معلمة القدرة للفرد، حيث كانت تزداد هذه الدقة بزيادة طول الاختبار وحجم العينة ودائما كانت دقة تقديرات معلمة القدرة للأفراد لصالح الطرق التي تعتمد أسلوب بييز خصوصا طريقة توقع الاقتران (EAP)، وقد جاءت هذه النتيجة متفقة مع نتائج الدراسات (الشريفين، ٢٠١٢؛ العبابنه، ٢٠٠٧ Swaminathan & Gifford, 1982, 1985, 1986; Wang & Vispoel, 1998). وتجدر الإشارة كذلك بأن دقة تقديرات معلمة القدرة للفرد كانت الأفضل عندما كان طول الاختبار (٦٠) فقرة فأعلى وحجم العينة (١٠٠٠) فأعلى لجميع طرق معايرة معلمة القدرة للفرد المستخدمة في برنامج بايلوج، وقد تفوقت طريقة توقع الاقتران البعدي (EAP) على جميع طرق التقدير المستخدمة في برنامج بايلوج في دقتها لتقدير معلمة القدرة للفرد، والذي يترتب عليه أنه عند استخدام برنامج بايلوج في معايرة قدرات الأفراد يفضل استخدام طريقة توقع الاقتران البعدي (EAP) للحصول على أخطاء معيارية منخفضة في تقدير معلمة القدرة للفرد كون البرنامج يستخدم طريقة الارجحية العظمى كطريقة افتراضية في تقدير معلمة القدرة للفرد.

إن السبب الذي أعطى أفضلية لطرق بييز في دقة تقدير معلمة القدرة للفرد على طريقة الارجحية العظمى يعزى إلى أن طريقة الأرجحية العظمى تغالي في تقدير قيمة الخطأ المعياري في تقدير قدرة الأفراد الذين أجابوا إجابة صحيحة أو إجابة خاطئة عن جميع فقرات الاختبار، مما جعل قيم الخطأ المعياري لقدرات الأفراد لهذه الطريقة أعلى من قيم الخطأ المعياري لقدرات الأفراد من طرق بييز. وتجدر الإشارة كذلك إلى أن السبب الذي جعل طريقة توقع الاقتران (EAP) تتفوق على طريقة الارجحية العظمى وطريقة تعظيم الاقتران (MAP) في دقة تقدير معلمة القدرة يكمن في أن كلتا الطريقتين (MML, MAP) تستخدمان التقريب المتتابع (Iteration) حسب طريقة نيوتن رافسون (Newton- Raphson Method) من أجل الحصول على تقديرات ثابتة لمعلمة القدرة، في حين تختلف الإجراءات الرياضية لطريقة توقع الاقتران (EAP) في تقديرها لمعلمة القدرة عن الطريقتين السابقتين في عدم استخدامها التقريب المتتابع، حيث تقوم بتقسيم متصل السمة الذي تفترض التوزيع الطبيعي له إلى (٦١) نقطة تسمى كل واحدة منها نقطة تربيع بأطوال متساوية، حيث يتم تحديد القيمة العددية لاقتران الأرجحية عند كل نقطة تربيع بدون إجراءات عمليات التقريب المتتابع، وهو ما أعطى هذه الطريقة أفضلية على الطرق الأخرى في دقة تقديرات معلمة القدرة للفرد، مما انعكس على قيمة الخطأ المعياري لتقدير معلمة القدرة للفرد.

ومن النتائج المهمة التي توصلت إليها الدراسة الحالية مدى إسهام طول الاختبار وطريقة التقدير والتفاعل بينهما في تباين دقة التقدير لمعلمة القدرة، وذلك من خلال مؤشر الدلالة العملية (مربع إيتا) التي بلغت قيمه لكل من طول الاختبار، وطريقة التقدير والتفاعل بينهما على التوالي (٠,٤٦٣، ٠,٦٧٢، ٠,٩٥٧) حيث يظهر من الشكل (٢) بأن أعلى دقة تقدير لمعلمة القدرة للفرد كانت عندما كان طول الاختبار (٦٠، ٨٠) فقرة ولجميع طرق التقدير المستخدمة في برنامج بايلوج، وربما يكون السبب في ذلك هو أن زيادة

عدد فقرات الاختبار يقلل من الأخطاء المعيارية لها، وهذا ما أكدته نتائج السؤال الأول، وبالتالي ينعكس ذلك على الأخطاء المعيارية لمعلمة القدرة للفرد، مما جعلها تنخفض بزيادة عدد فقرات الاختبار، ومثل ذلك جاء متفقا مع ما أشارت إليه بعض الدراسات مثل (Gau & Chen, 2005; Hulin, et al., 1982) التي أشارت إلى أن طول الاختبار يعد أهم عامل في تقدير معلمة القدرة للفرد وبالتحديد عند استخدام النموذج اللوجستي ثلاثي المعلمة لكونه النموذج الأقل تشددا من بين النماذج الأخرى ثنائية التدريج (أحادي المعلمة وثنائي المعلمة).

وجاءت نتائج الدراسة الحالية بشكل عام منسجمة مع نتائج الدراسات التي تناولت دقة التقديرات الناتجة من استخدام برنامج بايلوج (Huang, et al., 2001; Mislevey & Stocking, 1989; Qualls & Ansley, 1985; Yen, 1987) بأنه ينتج تقديرات دقيقة تحت شروط حجم العينة وطول الاختبار وطريقة التقدير، حيث كانت تزداد دقة التقدير لمعالم الفقرات وقدرات الأفراد بزيادة طول الاختبار وحجم العينة، مما جعل من هذا البرنامج شهرة عالمية في استخدامه في معايرة فقرات الاختبارات ثنائية التدريج خصوصا وأنه يستخدم طرق تقدير تعتمد على دالة الأرجحية ومنهج بييز. ومن الجدير بالذكر وفي ضوء نتائج الدراسة الحالية أنه عند استخدام النموذج اللوجستي ثلاثي المعلمة، وللوصول إلى دقة عالية في تقديرات معالم الفقرات وقدرات الأفراد لا بد أن لا يقل طول الاختبار عن ٣٠ فقرة وحجم العينة أعلى من (٥٠٠) فرد، وهذا ما أكده رب (Rupp, 2003) الذي أشار بأنه عند استخدام النموذج اللوجستي ثلاثي المعلمة لابد أن يكون طول الاختبار أعلى من (١٥) فقرة وأن لا يقل حجم العينة عن (٥٠٠).

الاستنتاجات والتوصيات

إن النتائج التي توصلت إليها الدراسة الحالية أشارت عموما إلى كفاءة طريقة الأرجحية العظمى المستخدمة في برنامج بايلوج في تقدير معالم فقرات الاختبار وبالتحديد عند استخدام النموذج اللوجستي ثلاثي المعلمة في إنتاج تقديرات دقيقة لمعالم فقرات الاختبار خاصة عند استخدام اختبارات تزيد على (٣٠) فقرة مع عينات تزيد على (٥٠٠) فردا، وفي نفس الوقت أشارت النتائج بأنه عند استخدام برمجية بايلوج في معايرة قدرات الأفراد للوصول إلى أفضل التقديرات لمعلمة القدرة للفرد ينصح باستخدام طريقة توقع الاقتران (EAP) مع اختبارات تزيد أطوالها على (٣٠) فقرة، وعلى الرغم من تباين دقة تقديرات معلمة القدرة للفرد باستخدام طرق التقدير الثلاثة المستخدمة في برنامج بايلوج، إلا أنها أشارت بشكل عام إلى كفاءة هذه الطرق في دقة تقدير معلمة القدرة للفرد خصوصا مع الاختبارات الطويلة نسبيا والذي بدوره انعكس على فعالية برنامج بايلوج وشهرته في معايرة الاختبارات ثنائية التدريج كبرنامج يستخدم طريقة الأرجحية العظمى ومنهج بييز في معايرة فقرات الاختبارات والمقاييس النفسية والتربوية ومعلمة القدرة للفرد.

وعلى الرغم من أهمية النتائج التي توصلت إليها الدراسة الحالية، إلا أن بعض القيود على هذه الدراسة تحد من تعميم النتائج بشكل عام، وذلك لوجود عوامل أخرى قد تلعب دورا في تأثيرها في دقة تقديرات معالم الفقرات وقدرات الأفراد، منها عدد نقاط التربيع المستخدمة في البرنامج حيث يفترض البرنامج القيمة الافتراضية (١٥)؛ مما يجعل الحاجة ماسة لإجراء دراسة مماثلة تأخذ بالحسبان عدد نقاط التربيع كمتغير آخر بالإضافة إلى متغيرات الدراسة ودراسة أثرها على دقة التقديرات، كذلك تم استخدام النموذج اللوجستي ثلاثي المعلمة حيث يمكن إعادة الدراسة على النماذج الثنائية الأخرى (أحادي المعلمة، وثنائي المعلمة) تحت نفس الشروط. ومن الجدير بالذكر كذلك بأنه تم في هذه الدراسة استخدام طريقة الأرجحية العظمى في تقدير معالم الفقرات لكون البرنامج يستخدم هذه الطريقة فقط، وهو ما يستدعي الحاجة لإجراء دراسة تتناول المقارنة

بين دقة تقديرات معالم الفقرات بطريقة الأرجحية العظمى المستخدمة في برنامج بايلوج مع برامج أخرى تستخدم طريقة بييز تحت نفس الشروط.

المراجع

أولا : المراجع العربية

الشريفين، نضال (٢٠١٢). اثر طريقة تقدير معالم الفقرة وقدرات الأفراد على قيم معالم الفقرة والخصائص السيكومترية للاختبار، في ضوء تغير حجم العينة. المجلة التربوية، ٢٦ .٢٣٨-١٧٧.

عبابنه، عماد (٢٠٠٥). مقارنة فاعلية طريقة الأرجحية العظمى وطريقة بييز في تقدير معلمة القدرة عند استخدام النموذج اللوغارتيمي الثلاثي. مجلة الاكاديمية العربية المفتوحة، ٣، ٥-٢٢.

علام، صلاح الدين (٢٠٠٥). نماذج الاستجابة للمفردات الاختبارية أحادية البعد ومتعددة الأبعاد وتطبيقاتها في القياس النفسي والتربوي(ط١). القاهرة: دار الفكر العربي.

المراجع الاجنبية

Anstasi, A. & Urbina, S. (2005). Psychological Testing (7 th ed.). New Jersey: Prentic-Hall, Inc.

Baker, F. B. (2001). The basics of item response theory (2nd ed). College Park, MD: ERIC Clearing House on Assessment and Evaluation

Baker, F. B., & Kim, S. H. (2004). Item response theory: Parameter estimation techniques (2nd ed.). New York: Marcel Dekker, Inc.

Baur, T., and Lukes, D .(2009). An evaluation of the IRT models through monto Carlo simulation. UW-L Journal of Undergraduate Research, XII, 1-7.

Bellhouse, D. (2004). The revered Tomes Bayes, FRS: A biography to celebrate the tercentenary of his birth. Statistical Science, 1, 3-43.

Bock, R. D., and Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. Psychometrika 46, 443-460.

Bock, R. D & Zimowski, M. F. (1996). Multiple group IRT, In W. J. Van Der Linden and R.K. Hambleton (Eds). Handbook of modern item response theory pp, 433 – 448. New York: Springer Verlag.

de Ayala, R. J. (2009). The theory and practice of item response theory. New York, NY: Guliford

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. Applied Psychological Measurement, 13, 77-90.

du Toit, M. (2003). IRT from SSI. BILOG-MG, MULTILOG, PARSCALE, TESTFACT. Lincolnwood, IL: Scientific Software International.

Embretson, S. E. & Reise, S. P. (2000). Item Response Theory for Psychologists. New Jersey: Lawrence Erlbaum Associates, Publishers.

Garre. G. & Vermunt. K. (2006). Avoiding Boundary Estimation in Latent Class

Harwell, M. R. (1997). Analyzing the result of Monto Carlo Studies in item response theory. Educational and Psychology Measurements, 57, 260-279.

Harwell, M. R., & Baker, F. B. (1991). The use of prior distributions in marginalized Bayesian item parameter estimation: A didactic. Applied Psychological Measurement, 15, 375-389.

Harwell, M., Stone, C. A., Hsu,T., Kirisci, L. (1996). Monte Carlo studies in item response theory. Applied Psychological Measurement, 20, 101-125.

Henard, D. H.(2000).Item response theory. In L. Grimm & Yarnold (Ed), Reading and understanding more multivariate statistics.(pp 67-97).Washington DC: American Psychological Association.

Huang, C. Y, Lohss, W. E, Lin , & Shin, D. (2001). Item calibration of license test with multiple specialty components. Submitted to Division DI: Educational Measurement, Psychometrics and assessment, Enabled Tiger, a web-based manuscripts processing system, Michigan State University.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameterlogistic item characteristic curves: A monte carlo study. Applied Psychological Measurement, 6, 249-260.

Analysis by Bayesian Posterior Mode Estimation. Behaviormentrika, 33, 256-271.

Gau, F., & Chen, L .(2005). Bayesian or non-bayesian : A comparison study of item parameter estimation in three-parameter logistic model. Applied Measurement in Education, 18. 351-380.

Hambleton, R .(1989).Principles and selected application of item response theory. In Linn, Robert, l .(Ed). Educational Measurement (3rd ed). New York: American Council on Education, Macmillan Publishing Company.

Hambleton, R ., K . (1994). Item Response Theory: A broad psychometric frame work for measurement advances. Psicothema, 3, 535-556.

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

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park CA: Sage Publication.

Han, K.T. and Hambleton , R .K(2007). User's Manual for WinGen: Windows Software that Generated IRT Model Parameter and Item Response. Center for Educational Assessment Research Report No 642, Amherst, MA: University of Massachusetts Center for Educational Assessment.

Qualls, A, L ., & Ansley, T, N. (1985). A comparison of item and ability parameter estimates derived from Logist and Bilog. Paper presented at the meeting of the National Council on Measurement in Education. Chicago, IL.

Rupp, A. A. (2003). Item response modeling with BILOG-MG and MULTILOG for windows. International Journal of Testing, 3, 365-384.

Seong, T. J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. Applied Psychological Measurement, 14,299-311

Stone, C. A. (1992). Recovery of Marginal Maximum Likelihood estimates in thetwo-parameter logistic response model: An evaluation of MULTILOG. Applied Psychological Measurement, 16, 1-16.

Swaminathan, H., & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. Journal of Educational Statistics, 7, 175-191.

Swaminathan, H., & Gifford, J. A.(1985). Bayesian estimation in two-parameter logistic model. Psychometrika, 50, 349-364.

Swaminathan, H., & Gifford, J. A. (1986). Bayesian estimation in three-parameter logistic model. Psychometrika, 51, 589-601.

Kamata. A. (1998). Some generalization of the Rasch model: An application of the hierarchical generalized linear model. Unpublished doctoral dissertation Michgan State University, Ann Arbor.

Fox, J. (2010). Bayesian item response modeling: Theory and application. New York: Springer.

Lord, F. M, (1975). Evaluation with artificial data of a procedure of estimating ability and item characteristic curve parameters. Princeton, Nj: Educational Testing Serves.

Lord, F .M, & Novick, M. R. (1968). Statistical Theories of mental test scores. London, Addison, Wesley: Publishing Company.

Mislevy, R. J. and Bock, R. D.(1990). BILOG3: Item analysis and test scoring with binary logistic models (2nd ed). Scientific software, Inc.

Mislevy, R. J. & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. Applied Psychological Measurement, 13, 57-75.

Patsula, L. N., & Gessaroli, M. E. (1995). A comparison of item parameters estimates and ICCs produced with TESTGRAF and BILOG under different test lengths and sample sizes. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Keeves (Ed). Educational Research, Methodology and Measurement: An International Handbook (pp 134-138). New York: Pergamen Press

Yen, W. (1987). A comparison of the efficiency and accuracy of Bilog and Logist. Pschometrika, 2, 275-291.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). Bilog Mg3 [Computer Software]. In M.du Toit (ed), IRT from SSI: Bilog-MG, Multilog, Parscale, Testfact.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. Pschometrika, 47, 397-412.

Van der Linden, W. J. (2010). Item Response Theory. International Encyclopedia of Education, 4, 81-88.

Wang, W. and Chen, C. (2005) Item parameter recovery, standard error estimates, and fit statistics of the Winsteps program for the family of Rasch models. Educational and Psychological Measurement 65, 376-404.

Wang, T, Vispoel. W. (1998). Properties of Ability Estimation Methods in Computerized Adaptive Testing. Journal of Educational Measurement, 35, 109-135.

Wilcox, R. R. (1988). Simulation as a research technique. In J. P.