مجلة جامعة الشارقة للعلوم الإنسانية والاجتماعية

جامعة الشارقة
ملتقى الحضارات

مجلة جامعة الشارقة

دورية علمية محكمة

للعلوم الإنسانية والاجتماعية

المجلد 11 ، العدد 2
صفر 1346 هـ / ديسمبر 2014 م

الترقيم الدولي المعياري للدوريات 1996-2339

تقصي دقة تقدير النموذج اللوجستي ثلاثي المعلمة لمعالم الفقرة وقدرة الأفراد في ضوء تغير طول الاختبار وحجم العينة: دراسة محاكاة

زايد صالح بني عطا

ملية التربية - جامعة اليرموك
إربد - الأردن

تاريخ الاستلام 2012-05-22 **تاريخ القبول 2013-03-25**

ملخص البحث

هدفت الدراسة إلى تقصي دقة تقدير النموذج اللوجستي ثلاثي المعلمة لمعالم الفقرة وقدرة الأفراد، في ضوء تغير طول الاختبار وحجم العينة. ولتحقيق الهدف من الدراسة وُلِّدت بيانات ثنائية الاستجابة بواقع 50 مرة لستة مستويات من طول الاختبار (10، 25، 50، 75، 100، 300 فقرة) وست مستويات لحجم العينة (100، 250، 500، 1000، 2000، 4000) من خلال استخدام برنامج (WINGEN). وباستخدام برنامج (Bilog- Mg3) حُلِّلت البيانات المولدة لكل خلية من تقاطع مستوى طول الاختبار ومستوى حجم العينة من أجل تقدير معالم الفقرات وقدرة الأفراد وإيجاد قيم RMSE و BIAS لمعالم الفقرات وقدرة الأفراد. كشفت نتائج الدراسة عن وجود أثر ذي دلالة إحصائية لكل من طول الاختبار وحجم العينة والتفاعل بينهما في دقة تقديرات معالم الفقرات وقدرة الأفراد. وكشفت النتائج أيضا أن الوسط الحسابي لقيم RMSE لمعالم الفقرات وقدرة الأفراد أخذ بالتناقص عندما زاد طول الاختبار على 50 فقرة وزاد حجم العينة على 2000. وقد أيدت قيم معاملات الارتباط بين المعالم المقدرة والحقيقة هذه النتيجة، حيث وصلت إلى الارتباط شبه التام، بالإضافة إلى ذلك كانت قيم التحيز في التقديرات قريبة من الصفر.

الكلمات المفتاحية : المحاكاة ، النموذج الثلاثي المعلمة ، حجم العينة، طول الاختبار ، دقة التقدير

خلفية الدراسة:

شهد منتصف القرن العشرين تطورات جوهرية في منهجيات القياس النفسي وطرق تصميم الاختبارات والمقاييس وتقنيات تحليل البيانات المستمدة منها، من خلال ظهور ما يسمى بنظرية الاستجابة للفقرة (Item Response Theory)، التي اعتبرت بمثابة الثورة والمستقبل الزاهر للقياس النفسي والتربوي (Anstasi & Urbena, 2005)، حيث قدمت إطارا مرجعيًّا لبناء المقاييس النفسية والتربوية، وطريقة تفسير الدرجات على هذه الاختبارات مقارنة بما قدمته النظرية الكلاسيكية في القياس (Mislevy & Bock, 1990; Van der Linden, 2010).

وتستند نظرية الاستجابة للفقرة على افتراضين أساسين هما: افتراض أحادية البعد (Uindimensionality) ويعني أن هناك قدرة واحدة أو سمة واحدة تفسر أداء المفحوص على الاختبار أو المقياس، ونظرا لوجود عوامل شخصية وعقلية مثل: القلق، والدافعية، ومفهوم الذات وعوامل أخرى متعلقة بظروف تطبيق الاختبار فقد أشار «هامبلتون» (Hambleton, 1994) أن مثل هذا الافتراض لا يتحقق بشكل مؤكد، ومع ذلك لابد من وجود عامل أساس في الاختبار، وهو ما يشار إليه بالقدرة التي يقيسها الاختبار. أما الافتراض الثاني فهو الاستقلال الموضعي (Local Independence) ويقصد به أن تكون استجابات المفحوصين على الفقرات المختلفة مستقلة إحصائيا عند مستوى قدرة معين، وحتى يكون هذا الافتراض صحيحًا يجب أن لا تؤثرَ استجابة المفحوص عن فقرة ما على استجابته على الفقرات الأخرى في الاختبار(Baker, 2001).

وتعتمد نظرية الاستجابة للفقرة على مجموعة من النماذج الاحتمالية القائمة على الاقتران اللوغاريتمي التي تحدد العلاقة بين أداء الفرد على الفقرة والقدرة أو السمة 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 وتيريًّا لموقع الفرد على متصل السمة، مما يعني أن احتمال الإجابة الصحيحة يزداد بزيادة قدرة الفرد (Hambleton, 1994; Henard, 2000). وتصنف هذه النماذج حسب مستوى الاستجابة إلى نماذج ثنائية التدريج أو نماذج متعددة التدريج (De Gruijter & Van der Kamp,2005).

وتعتبر النماذج ثنائية التدريج من أشهر النماذج أحادية البعد استخدامًا في بناء الاختبارات والمقاييس (De Gruijter & Van Der Kamp, 2005; Embreston & Rise, 2000)، إذ يوجد منها ثلاثة نماذج، أولها: النموذج اللوغاريتمي ذو المعلمة الواحدة (One- Parameter Logistic Model)، أو المشهور باسم نموذج راش (Rasch Model)، ويعتبر من أكثر النماذج الثنائية استخداما، حيث لا يتطلب عمليات حسابية معقدة على الرغم من تشدد افتراضاته، وهي أن الفقرات متساوية في تمييزها ولا يوجد فيها تخمين (Harris, 1989).أما النموذج الثاني فهو النموذج اللوغاريتمي ثنائي المعلمة (Two- Parameter Logistic Model)، وهو يختلف عن النموذج الأحادي بأنه يتيح

تقدير معلمة التمييز للفقرة إضافة إلى معلمة الصعوبة، ويعد النموذج ثنائي المعلمة أقل عمومية من النموذج الثالث والأخير من بين هذه النماذج وهو النموذج اللوجستي ثلاثي المعلمة (Three –Parameter Logistic Model) (علام، 2005).

ويعد النموذج اللوجستي ثلاثي المعلمة أقل النماذج ثنائية التدريج تشددا، إذ يسمح بأن تختلف فقرات الاختبار في صعوبتها وتمييزها، حيث يصعب الحصول على مجموعة من الفقرات تميز بدرجة واحدة بين مستويات السمة التي يقيسها الاختبار، وفي الوقت نفسه أضاف معلما ثالثا يمثل احتمال توصل الفرد إلى الإجابة الصحيحة على الفقرة عن طريق التخمين الذي تميز به عن النموذج اللوجستي ثنائي المعلمة، وأطلق عليه «لورد» معلم التخمين أو معلم الخط التقاربي الأدنى (lower Asymptote Line)، والصيغة الرياضية التي تصف النموذج اللوجستي ثلاثي المعلمة هي: (Lord, 1980)

$$P_i(\theta) = C_i + \frac{1 - C_i}{1 + e^{Da_i(\theta - b_i)}} (i = 1,2,....,n)$$

وإن أحد أهم القضايا الرئيسة عند استخدام أحد نماذج نظرية الاستجابة للفقرة تكمن في تقدير معالم النموذج، التي تعتمد على أساليب التحليل العددي من خلال استخدام البرامج الحاسوبية المختلفة. فقد عدّ «لورد ونوفيك» (Lord & Novick, 1968) التقدير الإحصائي للعلاقة بين احتمال الاستجابة الصحيحة عن فقرة من فقرات الاختبار والقدرة التي يقيسها الاختبار المشكلة الرئيسة لمستخدم هذه النظرية، حيث بيّنَت «ستوكنغ» (Stocking, 1990) أن تقدير معالم الفقرة يعد من أهم القضايا التي يعتمد عليها نجاح نظرية الاستجابة للفقرة، خصوصا في التطبيقات التي تعتمد كثيرًا على تلك المعالم، 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.

ويعتبر الخطأ المعياري مؤشرا إحصائيا يعتمد عليه الباحثون للحكم على دقة تقدير العينة للمجتمع (Agresti & Finnlay, 2009)، وهذا يتماثل مع الاختبارات فهناك تغاير في العلامات أو في تقدير القدرة من موقف اختباري لموقف آخر، وأن الباحث الذي يستخدم نظرية الاستجابة للفقرة يقوم بالحصول على الخطأ المعياري لكل من القدرة ومعالم الفقرات، فقد أشار ثيسين ووينر(Thissen & Wainer, 1982) إلى أهمية تحديد مقدار الخطأ في تقدير المعلمات الذي يعبر عن الدقة في تقدير معلمات النموذج المستخدم باعتبار أنه إذا حُصِل على مقدار قليل للخطأ المعياري فإنه مؤشر على دقة القياس، خصوصا وأن الخطأ المعياري يؤدي دورا مهما في دالة معلومات الاختبار(Test Information Function) التي تعد مؤشرا يستدل منها على ثبات الاختبار في نظرية الاستجابة للفقرة (Embreston & Rise, 200).

فالدقة تشير إلى المدى الذي يتوافق فيه القرار المستند إلى درجات الاختبار مع القرار

الذي يمكن اتخاذه فيما لو كانت الدرجات لا تتضمن أية أخطاء قياس، ومن هذا المنطلق اهتم البحث السيكومتري في مجال استخدام نظرية الاستجابة للفقرة في البحث عن العوامل التي تؤثر في دقة التقديرات لمعالم الفقرات وقدرات الأفراد، نظرا للزيادة المضطردة في استخدامها لغايات التقويم النفسي والتربوي، ومن العوامل التي اهتم البحث السيكومتري بدراستها طول الاختبار، وحجم عينة المفحوصين. فقد أشارت نتائج بعض الدراسات (Barnes & Wise, 1991; Wainer & Mislevy, 1990) بأن الحد الأدنى للمفحوصين لإنتاج تقديرات دقيقة تختلف باختلاف مواقف الاختبار والنموذج المستخدم، وهذا ما أكده هامبيلتون وجونز (Hambleton & Jones, 1994)بأن هناك عاملين يؤثران في دقة تقدير معالم الفقرات ومن ثَمَّ على دالة المعلومات للاختبار، الأمر الذي يؤثر في دقة القياس، وهما طول الاختبار، وحجم عينة المفحوصين.

فعلى صعيد الدراسات التي اهتمت بدراسة أثر حجم العينة وطول الاختبار على دقة تقديرات معالم الفقرات وقدرة الأفراد عند استخدام النماذج ثنائية التدريج وخصوصا النموذج اللوجستي ثلاثي المعلمة ، فقد أجرى هامبلتون وتروب (Hambleton & Ttrub,1973) دراسة هدفت إلى مقارنة النموذج اللوجستي أحادي المعلمة مع النموذج اللوجستي ثنائي وثلاثي المعلمة باستخدام بيانات وُلِّدَتْ بالمحاكاة. ولتحقيق الهدف من الدراسة اعتُمِدَ منحنى خاصية الفقرة لتحديد دقة التقدير. أشارت نتائج الدراسة أن النموذج اللوجستي ثلاثي المعلمة كان الأفضل في تقدير قدرة الفرد. وفي دراسة أخرى قام بها هامبلتون وكوك (Hambleton & Cook, 1980) بدراسة هدفت إلى معرفة اثر حجم العينة وطول الاختبار في دقة تقدير قدرة الفرد وفق النموذج الثلاثي، ولتحقيق الهدف من الدراسة دُرِسَت ثلاثة مستويات لحجم العينة (50، 200، 1000)، وثلاثة مستويات لطول الاختبار(10، 20، 80). بيَّنت نتائج الدراسة أن أكبر قيمة للخطأ المعياري للتنبؤ كانت 2.19 عندما يكون طول الاختبار 10 فقرات، وحجم العينة 50 فردا، بينما تراوحت قيمة الخطأ المعياري بين 0.88 و 1.50 عندما كان طول الاختبار 20 فقرة وحجم العينة 200 فردا. وأشارت نتائج الدراسة كذلك بأن قيمة الخطأ المعياري للتنبؤ تنقص عندما كان طول الاختبار 80 فقرة وحجم العينة 1000 فردا، وبشكل عام بينت النتائج بأن دقة التقدير تزداد بزيادة حجم العينة وطول الاختبار.

وهدفت دراسة هيلين وليساك ودراسجو (Hulin, Lisak & Drasgow, 1982) إلى دراسة أثر حجم العينة وطول الاختبار في دقة تقدير معالم الفقرات ومعلمة القدرة للأفراد باستخدام النموذج اللوجستي ثنائي المعلمة وثلاثي المعلمة. ولتحقيق الهدف من الدراسة استُخدمت أربعة مستويات لحجم العينة (200 ، 500 ، 1000، 2000) وثلاثة مستويات لطول الاختبار (15 ، 30، 60). بينت نتائج الدراسة أنه عند استخدام النموذج اللوجستي ثلاثي المعلمة كان معامل الارتباط بين معلمة التمييز الحقيقية والمقدرة 0.36 عندما كان طول الاختبار 15 فقرة وحجم العينة 200 فرد، في حين كان معامل الارتباط يساوي 0.84 عندما كان طول الاختبار 60 فقرة وحجم العينة 2000 فرد. كما أشارت النتائج إلى أن معامل الارتباط بين معلمة الصعوبة الحقيقية والمقدرة بلغ 0.72 عندما كان طول الاختبار 15 فقرة وحجم العينة 200 فرد، وازداد هذا المعامل ليصبح 0.84

عند استخدام حجم عينة 2000 فرد وطول اختبار 60 فقرة. وفيما يتعلق بقدرة الفرد فقد أشارت نتائج الدراسة إلى أن قيمة معامل الارتباط بين القدرة الحقيقة والمقدرة تزداد بزيادة حجم العينة وطول الاختبار، حيث بلغت أكبر قيمة له 0.96 عندما كان طول الاختبار 60 فقرة وحجم العينة 2000 فرد.

وأجرى فاريش (Farish, 1984) دراسة هدفت إلى الكشف عن أثر اختلاف حجم العينة على تقدير معلمة الصعوبة باستخدام نموذج راش. ولتحقيق الهدف من الدراسة طُبِّق اختبار تحصيلي في الرياضيات على عينة بلغ حجمها الكلي (2000) طالب. كشفت نتائج الدراسة أن زيادة حجم العينة يزيد من مطابقة الاختبار لنموذج راش.

وأجرى الدرابيع (2001) دراسة هدفت إلى التحقق من فعالية النموذج اللوغاريتمي ذي المعلمة الواحدة نموذج راش، في دقة تقدير قدرة الفرد ومعلمة صعوبة الفقرة عند استخدام حجم عينة (50، 100، 500)، وعدد فقرات الاختبار(25، 50 ، 300). بينت نتائج الدراسة وجود فروق جوهرية لتفاعل كل من متغير حجم العينة وطول الاختبار في دقة تقدير قدرة الفرد، ووجود فروق جوهرية في دقة تقدير قدرة الفرد تعزى إلى متغير طول الاختبار وحده، وليس هناك دلالة لحجم العينة على دقة تقدير قدرة الفرد، وفيما يتعلق بدقة تقدير معلمة الصعوبة، فقد بينت النتائج بأن هناك فروقا جوهرية تعزى لتفاعل كل من حجم العينة وطول الاختبار، وفي الوقت نفسه كان ثمة فروق جوهرية لكل من طول الاختبار وحجم العينة كل على حدة.

كما قام ستون ويوموثو (Stone & Yumoto, 2004) بدراسة أثر حجم العينة في تقدير معالم الفقرات ثنائية التدريج باستخدام نموذج راش ونماذج نظرية الاستجابة للفقرة. ولتحقيق الهدف من الدراسة تم سحب 30 عينة عشوائية من نتائج تطبيق اختبار (Knox's Cube Test Revised). بينت نتائج الدراسة بأن نموذج راش يعطي أقل تقدير لمعلمة الصعوبة، وأن العينات قليلة الحجم كانت الأقل في مطابقة للنموذج المستخدم.

ومن جهة أخرى أجرى جلاس (Glass, 2005) دراسة هدفت إلى معرفة أثر حجم العينة وعدد الفقرات في دقة تقدير معلمة القدرة وفق طريقة «بيز» التي تعتمد أسلوب تعظيم الاقتران، ولتحقيق الهدف من الدراسة وُلِّدَتْ بيانات ثنائية التدريج بأحجام عينات (500 ، 1000، 2000) وفقرات بعدد (200 ، 440)، و اعتُمِدَ الوسط الحسابي للأخطاء المعيارية لمعلمة القدرة عند مستويات مختلفة مختارة من القدرة موزعة على أطراف متصل القدرة (2، 1، 0، 1 -، 2 -) عند أحجام العينات المختلفة والفقرات المختلفة. أظهرت نتائج الدراسة أنه بزيادة عدد الفقرات عند أحجام العينات (500 ، 2000) تقل الأخطاء المعيارية، وهذا بدوره يؤدي إلى زيادة دقة التقدير، أما عند زيادة عدد الفقرات عند حجم العينة (1000) تقل دقة التقدير.

وهدفت دراسة لوكس وباور (Baur & Lukes, 2009) تقييم نماذج نظرية الاستجابة للفقرة من خلال استخدام أسلوب «مونتو كارلو» للمحاكاة (Monto Carlo Simulation). ولتحقيق الهدف من الدراسة استُخدِمت خمسة مستويات لحجم العينة (100، 250، 500،

1000، 2000) وخمسة مستويات لطول الاختبار(5، 10، 15، 20، 30)، وقد حللت البيانات باستخدام النماذج: أحادي، وثنائي وثلاثي المعلمة. أشارت نتائج الدراسة أنه عند استخدام النموذج ثنائي المعلمة أعطى تقديرات دقيقة أكثر من النموذج الأحادي والنموذج ثلاثي المعلمة، حيث كانت معاملات الارتباط عالية بين المعالم الحقيقة للفقرة والمعالم المقدرة، وأشارت النتائج كذلك أنه عند استخدام النموذج ثلاثي المعلم اظهر تحيزا في تقديره لمعالم الفقرات، وبشكل عام أشارت النتائج إلى أن معاملات الارتباط كانت متقاربة وجيدة عند استخدام النموذج أحادي المعلمة والنموذج ثنائي المعلمة بين صعوبة الفقرة وقدرات الأفراد.

وفي دراسة الثوابية (2010) التي هدفت إلى استقصاء أثر حجم العينة في تقدير معلمة صعوبة الفقرة والخطأ المعياري في تقديرها باستخدام نظرية الاستجابة للفقرة. ولتحقيق الهدف من الدراسة تم تقدير معلمة الصعوبة، والخطأ المعياري في تقديرها باستخدام اختبار في الرياضيات للصف العاشر الأساسي مكون من 80 فقرة وطبق على عينات عشوائية تراوح حجمها ما بين 200 طالب وطالبة إلى 11292 طالبا وطالبة.أشارت نتائج الدراسة بأن قيمة معلمة الصعوبة تزداد بزيادة حجم العينة، حيث بلغ متوسط صعوبة فقرات الاختبار 0.31 لوجيت عندما كان حجم العينة 200 وازداد بحيث أصبح 1.1 عندما كان حجم العينة 11292 طالبا وطالبة. وأشارت النتائج كذلك بان الخطأ المعياري في تقدير معلمة الصعوبة يتناقص بزيادة حجم العينة ، حيث بلغ متوسط الأخطاء المعيارية في التقدير 0.32 عندما كان حجم العينة 200، وتناقص بحيث أصبح 0.07 عندما كان حجم العينة 11292 طالبا وطالبة.

إن المتصفح لنتائج الدراسات السابقة يجد بأن هناك توجها عاما بأن الحد الأدنى للمفحوصين لإنتاج تقديرات دقيقة تختلف من موقف الاختبار والنموذج المستخدم، وعندما يتعلق الأمر باستخدام النموذج اللوجستي ثلاثي المعلمة، يتضح بأن هناك تباينا عاما في نتائج الدراسات حول الحد الأدنى لحجم العينة وطول الاختبار، كذلك لم تتناول الدراسات السابقة التفاعل بين هذين العاملين باستثناء دراسة الدرابيع (2001) التي استخدمت النموذج اللوغاريتمي ذي المعلمة الواحدة « نموذج راش». لذا جاءت هذه الدراسة لتقصي دقة تقدير النموذج اللوجستي ثلاثي المعلمة لمعالم الفقرة ومعلمة قدرة الفرد في ضوء تغير طول الاختبار وحجم العينة والتفاعل بينهما، كمحاولة لتقييم دقة وكفاءة النموذج اللوجستي ثلاثي المعلمة في تقدير خصائص الفقرات وأخطاء القياس كمؤشر على دقة القياس من خلال استخدام أسلوب المحاكاة للواقع التطبيقي لهذا النموذج، وللدور المهم لطول الاختبار باعتباره عاملا يؤثر في المواقف الاختبارية من حيث الجهد والوقت والإعداد والتطبيق وتفسير النتائج خصوصا في ظل الاستخدام المتزايد لنماذج نظرية الاستجابة للفقرة في التقويم التربوي والنفسي سواء في تطوير المقاييس أو معادلة الاختبارات والاختبارات المحوسبة في الوسط العربي والمحلي. وتجدر الإشارة كذلك بأن هذه الدراسة اختلفت عن الدراسات السابقة بأخذها مستويات متباينة من أطوال الاختبارات تراوحت من الاختبارات القصيرة وحتى الاختبارات الطويلة على غرار بنوك الأسئلة، وبالمثل تناولها أيضا مستويات متباينة في حجم العينة، من العينات الصغيرة

وحتى العينات الكبيرة التي تناسب تقنين الاختبارات والمقاييس.

وقد اخْتِير النموذج اللوجستي ثلاثي المعلمة هدفا للدراسة الحالية؛ لأن هذا النموذج يعد النموذج العام للنماذج ثنائية التدريج، حيث أشار هامبلتون وتروب (Hambleton & Traub, 1971) بأن النموذج اللوجستي ثلاثي المعلمة هو الأفضل، والأقل تشددا؛ ولأن هذا النموذج يفترض تأثر الإجابات بعامل التخمين الذي تميز به والذي يعد أحد العوامل المؤثرة في أداء الاختبار(McDonald, 1989)، وفي الوقت نفسه يلاحظ ندرة الدراسات العربية على المستوى العربي والمحلي التي اهتمت بتقييم فاعلية النموذج ثلاثي المعلمة في تقدير معالم الفقرات وقدرة الأفراد، فلم يجد الباحث أي دراسة تناولت موضوع الدراسة الحالية.

مشكلة الدراسة وأسئلتها:

جاءت نظرية الاستجابة للفقرة استجابة للتطور البحثي السيكومتري، للتخلص من العيوب التي شابت استخدام النظرية التقليدية في التقويم التربوي والنفسي، إلا أن 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. فقد تباينت وجهة نظر الباحثين حول العوامل المؤثرة في دقة التقديرات، فمنهم من يرى أن للبرنامج المستخدم في تحليل البيانات أثرًا في دقة التقدير، ومنهم من يرى أن الطريقة المستخدمة في تقدير معالم الفقرات وقدرة الأفراد تؤدي دورا في دقة التقدير، أو انتهاك افتراضات النظرية، وكذلك لحجم العينة وطول الاختبار أثر في دقة التقدير، حيث يتضح بأن العينات المختلفة قد تولد تقديرات مختلفة. وقد استندت الدراسات السابقة في تقييم دقة التقديرات على الطرق الإحصائية الوصفية؛ لذا بات من الضروري معرفة دلالة الفروق للخصائص السيكومترية لكل من الفقرات والأفراد باختلاف العوامل التطبيقية للاختبار كاختلاف حجم العينة وطول الاختبار والتفاعل بينهما باعتبارها عوامل مهمة في دقة تقديرات معالم الفقرات وقدرة الأفراد المنبثقة من استخدام النموذج اللوجستي ثلاثي المعلمة من خلال محاكاة الواقع للمواقف الاختيارية المختلفة؛ من أجل تقديم أدلة إمبريقية وقواعد إيهامية عن حجم العينة وطول الاختبار المناسبين للوصول إلى أدق التقديرات؛ إذ إن استخدام المحاكاة في توليد البيانات بناء على عدد المفحوصين وطول الاختبار يسهم في دراسة متعمقة لأثر هذه المتغيرات على دقة التقدير في ظل ظروف تطبيقية مكررة ضمن موقف تطبيقي موحد. لذا حاولت هذه الدراسة على وجه التحديد الإجابة عن السؤالين الآتيين :-

1. هل تختلف دقة تقديرات معالم الفقرات (الصعوبة، والتمييز، والتخمين) المعايرة باختلاف طول الاختبار وحجم العينة ؟
2. هل تختلف دقة تقديرات معلمة القدرة للفرد باختلاف طول الاختبار وحجم العينة ؟

أهمية الدراسة:

تكمن أهمية الدراسة الحالية في الكشف عن دقة التقديرات المنبثقة من استخدام النموذج اللوجستي ثلاثي المعلمة لمعالم الفقرات وقدرة الأفراد في ضوء تغير طول الاختبار وحجم العينة والتفاعل بينهما، وذلك بالاعتماد على التصاميم التجريبية من خلال أسلوب المحاكاة بدلا من الاستناد على الطرق الوصفية في تقييم النتائج. وتكمن أهمية الدراسة كذلك في توفير أدلة إمبريقية من خلال أساليب الإحصاء الاستدلالي للوصول إلى قواعد إيهام للباحثين عند استخدام النموذج اللوجستي ثلاثي المعلمة بأفضل طول للاختبار وحجم للعينة لإعطاء تقديرات دقيقة لمعالم الفقرات وقدرة الأفراد. وفي الوقت نفسه تفتح الأبواب للدراسات العربية التي تستخدم المحاكاة في التحقق من القضايا السيكومترية المتعلقة باستخدام نظرية الاستجابة للفقرة والعوامل التي تؤثر في دقة التقديرات لندرتها في الأدب السيكومتري العربي – في حدود علم الباحث-.

مصطلحات الدراسة:

النموذج ثلاثي المعلمة: هو أحد نماذج نظرية الاستجابة للفقرة ثنائية التدريج، حيث يمكن لهذا النموذج تقدير أربع معلمات، هي: قدرة الفرد، ومعامل صعوبة الفقرة، ومعامل التمييز، ومعامل تخمين الفقرة.

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

دقة التقدير: وهو مصطلح يشير إلى جودة التقدير التي يميزها الاحتمالية الكبيرة في أن التقدير قريب من المعالم الحقيقة، وذلك بالاعتماد على الجذر التربيعي لمتوسطات مربعات الانحرافات للفروق بين المعالم المقدرة والمعالم الحقيقية، ومعاملات الارتباط بينهما.

تصميم الدراسة:

لتحقيق أهداف الدراسة الحالية في الكشف عن دقة تقدير النموذج اللوجستي ثلاثي المعلمة لمعالم الفقرات وقدرة الأفراد باختلاف طول الاختبار وحجم العينة، اتبعت الدراسة الحالية التصميم التجريبي العاملي (6X6) بالاعتماد على المحاكاة، حيث كان المتغيران المستقلان في التصميم طول الاختبار وحجم العينة، وقد استُخدِمت ستة مستويات من حجم العينة (100، 250 ، 500 ، 1000 ، 2000 ، 4000)، 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 استُخدِمت ستة مستويات من طول الاختبار (10 ، 25 ، 50 ، 75 ، 100 ، 300)، وتعد مثل هذه الأطوال مناسبة لأطوال الاختبارات المستخدمة في الواقع العملي سواء كانت قصيرة أو متوسطة أو

طويلـة. ويظهـر الجـدول 1 تصميـم الدراسـة.

الجدول 1

التصميم العاملي للدراسة

طول الاختبار / حجم العينة	10	25	50	75	100	300
100	x*	x*	x*	x*	x*	x*
250	x*	x*	x*	x*	x*	x*
500	x*	x*	x*	x*	x*	x*
1000	x*	x*	x*	x*	x*	x*
2000	x*	x*	x*	x*	x*	x*
4000	x*	x*	x*	x*	x*	x*
	x*	x*	x*	x*	x*	x*

* : تشير إلى الخلية التي وُلِّدَتْ فيها الاستجابات

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

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

حيث تشير π_i إلى المعلمة الحقيقة (الصعوبة، والتمييز، والتخمين، والقدرة) والرمز $\hat{\pi}_i$ يشير إلى المعالم المقدرة من النموذج اللوجستي ثلاثي المعلمة، وتشير K إلى عدد مرات التكرار. ويعد مؤشر RMSE مؤشرا إحصائيا مهمًّا للتحقق من مدى انسجام البيانات مع النموذج المستخدم، وهو من المؤشرات التي أوصى بها هارويل (Harwell, 1997) باعتباره مؤشرا إحصائيا يعتمد عليه في تقييم دقة التقدير، حيث يمكن أن ينظر له على بأنه وحدة معيارية تمثل مدى ابتعاد القيم الحقيقة عن القيم المقدرة، فكلما كانت قيمته منخفضة دل ذلك على دقة عالية من التقدير؛ بسبب التجانس العالي بين هذه القيم.

ومن المؤشرات الإحصائية التي اعتمدت في هذه الدراسة كأساس نظري للحكم على كفاءة التقدير ودقته مؤشر التحيز(BIAS)، الذي يعد مؤشرا للكشف عن نوع الخطأ هل هو منتظم أو غير منتظم بين التقديرات المقدرة والحقيقة، بالإضافة إلى ذلك نوع التحيز

في التقديرات هل هو موجب أو سالب، وإذا كانت قيمته قريبة من الصفر، فإن ذلك يشير إلى أن الخطأ المنتظم قليل (Pelton, 2002)، ويحسب التحيز بالقانون الآتي:

$$BIAS=\frac{\sum_{i=1}^{K}(\pi_i-\hat{\pi}_i)}{K}$$

والرموز الواردة في المعادلة هي نفسها التي أشير إليها في معادلة RMSE. وتجدر الإشارة كذلك أنه اعتُمِد معامل ارتباط بيرسون بين المعالم المقدرة والمعالم الحقيقة مؤشرًا على دقة التقديرات الناتجة من النموذج المستخدم.

توليد البيانات:

وُلِّدَت البيانات باستخدام طريقة المحاكاة؛ إذ أشار «ويلكوس» (Wilcox, 1988) إلى أهمية دراسات المحاكاة؛ لأنها تتيح المجال للحصول على بيانات من مواقف مختلفة وتحت شروط مختلفة لإيجاد حلول لمشكلات إحصائية لا يمكن الوصول إليها بالمواقف العملية. وقد وُلِّدَت البيانات باستخدام برنامج WINGEN وهو من تصميم وإنتاج هان وهامبيلتون (Han & Hambleton, 2007)، حيث يمكن من خلال هذا البرنامج توليد استجابات لاختبارات أحادية البعد سواء كانت ثنائية التدريج أم متعددة التدريج، بالإضافة إلى استجابات متعددة الأبعاد. وتجدر الإشارة كذلك إلى أن البرنامج يؤهل المستخدم لتوليد استجابات لأكثر من مجموعة من المفحوصين بأنواع مختلفة من التوزيعات، ويمكن من خلاله تحليل الاستجابات الثنائية المولدة لإيجاد المعالم المقدرة لكل من الفرد والفقرة من خلال استخدام احد البرامج التي تستند على نظرية الاستجابة للفقرة. وقد وُلِّدَت عملية البيانات وفق المراحل الآتية:

المرحلة الأولى: وُلِّدَت ستة مستويات من الاختبارات كما ورد في تصميم الدراسة، حيث وُلِّدَت معالم الفقرات وفق معادلة النموذج اللوجستي ثلاثي المعلمة باستخدام برنامج WINGENضمن الشروط الآتية :-

- وُلِّدَت معلمة التمييز للفقرات وفق توزيع منتظم (2 , 0.4) ~ Uniform وهذا المدى يعطي معاملات تمييز جيدة، وتعد قيم معلمة التمييز الحقيقية التي وُلِّدَت مماثلة للقيم الحقيقية التي استخدمها هامبلتون وسوامينيثان (Hambleton & Swaminthan, 1985) حيث أكدا على أن تكون قيم معلمة التمييز تتراوح ما بين 0 و 2 لوجيت.
- وُلِّدَت معلمة الصعوبة للفقرات وفق التوزيع الطبيعي (1 , 0) ~Normal، وهذا ينتج فقرات متباينة في الصعوبة تتراوح ما بين (3- و3).
- وُلِّدَت معلمة التخمين للفقرات وفق توزيع التوزيع الطبيعي (0.05 , 0.17) ~ Normal

وهذا التوزيع ينتج قيم لمعلمة التخمين تماثل قيم التخمين للاختبار الاختيار من متعدد المؤلف من خمسة بدائل.

المرحلة الثانية: وُلِّدَت استجابات المفحوصين وفق التوزيع الطبيعي (0 , 1) Normal ~ بواقع 50 مرة لكل خلية من خلايا التصميم التجريبي الوارد في الجدول 1 باستخدام القيم نفسها للمعالم الحقيقة للفقرات التي وُلِّدَت في المرحلة الأولى، وبذلك يكون عدد البيانات التي وُلِّدَت يساوي (6 X 50 X 6). وقد اختِيرت عدد مرات التوليد (number of replication) 50 مرة؛ لأن عدد مرات التوليد التي استخدمت في دراسات المحاكاة كانت تتراوح مابين 3 إلى 100 مرة، وأن التقديرات كانت تستقر عندما يكون عدد مرات التوليد 50 فما فوق (Kamata, 1998).

تحليل البيانات:

لتحقيق الهدف من الدراسة، قام الباحث بتحليل البيانات التي وُلِّدَت في المرحلة الثانية باستخدام برنامج Bilog-Mg 3 النسخة المطورة من خلال شركة البرامج العالمية الدولية (Scintific Software International,Inc) لإيجاد معالم الفقرات المقدرة: (الصعوبة، والتمييز، والتخمين) ومعلمة قدرة الفرد المقدرة من النموذج اللوجستي ثلاثي المعلمة، باستخدام طريقة الأرجحية القصوى (Zimowski, Muraki, Mislevey & bock, 2003) حيث تتميز هذه الطريقة بالفاعلية والقيم التقديرية للأخطاء المعيارية الناتجة تتميز بالدقة، ويمكننا كذلك الحصول على قيم تقديرية لمعالم الأفراد الذين أجابوا إجابة صحيحة أو إجابة خاطئة عن جميع الفقرات (علام، 2005). ولقد اختِير برنامج Bilog-Mg 3 بالاستناد إلى نتائج بعض الدراسات (Kirici, Hsu & Yu, 2001; Toland, 2008; Yen, 1987) التي أشارت إلى أفضلية هذا البرنامج في إعطاء تقديرات أدق مقارنة مع غيره من البرامج المستخدمة لنفس الغرض. وقد تم الحفاظ على طريقة التقدير، وهي طريقة الأرجحية العظمى لما لطرق التقدير من أثر على دقة التقديرات بالاستناد على نتائج الدراسات التي تناولت المقارنة بين طرق التقدير وأثرها في دقة تقديرات معالم الفقرات وقدرة الأفراد (Gau & Chen, 2005; Lord, 1986 ; Rondall, 2007 ; Swaminathan & Gifford, 1986).

ولإيجاد قيم RMSE والتحيز في التقديرات فقد تم ذلك من خلال الملف التنفيذي باستخدام برنامج WINGEN للربط بينه وبين برنامج Bilog-Mg 3 لتحليل البيانات المولدة، وحساب قيم RMSE وBIAS ومعاملات الارتباط لكل من معالم الفقرات وقدرة الأفراد وتخزينها في ملف خاص؛ حيث أصبح 1800 قيمة من RMSEلكل من معالم الفقرات وقدرة الأفراد (الصعوبة، التمييز، التخمين ، قدرة الفرد) و 1800 قيمة أيضا للتحيز لكل من معالم الفقرات وقدرة الأفراد، وقد استُخدم برنامج spss لاستخراج النتائج.

نتائج الدراسة:

للإجابة عن السؤال الأول للدراسة فقد قدِّرَت قيم معالم الفقرات (الصعوبة، والتمييز، والتخمين) لكل خلية من خلايا التصميم التجريبي وإيجاد قيم RMSE لكل معلمة من معالم الفقرات تبعا لمتغيري الدراسة طول الاختبار وحجم العينة، ويوضح الجدول 2 الوسط الحسابي والانحراف المعياري لقيم RMSE لمعالم الفقرات، باختلاف طول الاختبار وحجم العينة.

الجدول 2

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
10	100	0.385	0.082	0.278	0.096	0.047	0.007
	250	0.305	0.062	0.189	0.057	0.050	0.013
	500	0.272	0.059	0.168	0.038	0.045	0.009
	1000	0.212	0.059	0.127	0.031	0.046	0.011
	2000	0.187	0.047	0.116	0.032	0.041	0.011
	4000	0.145	0.047	0.094	0.029	0.038	0.008
25	100	0.361	0.104	0.240	0.049	0.043	0.005
	250	0.325	0.099	0.142	0.019	0.038	0.005
	500	0.236	0.032	0.126	0.024	0.039	0.006
	1000	0.203	0.029	0.108	0.015	0.033	0.006
	2000	0.159	0.031	0.070	0.011	0.028	0.005
	4000	0.115	0.021	0.056	0.010	0.023	0.004

0.005	0.052	0.022	0.209	0.134	0.429	**100**	**50**
0.004	0.049	0.032	0.198	0.047	0.321	**250**	
0.005	0.041	0.018	0.156	0.022	0.262	**500**	
0.003	0.034	0.011	0.087	0.020	0.190	**1000**	
0.004	0.030	0.008	0.070	0.017	0.142	**2000**	
0.003	0.023	0.006	0.059	0.016	0.107	**4000**	
0.003	0.055	1.570	1.266	0.129	0.419	**100**	**75**
0.003	0.048	0.026	0.182	0.038	0.316	**250**	
0.003	0.044	0.015	0.141	0.031	0.236	**500**	
0.003	0.037	0.014	0.108	0.019	0.175	**1000**	
0.003	0.033	0.008	0.089	0.015	0.133	**2000**	
0.003	0.026	0.008	0.072	0.010	0.110	**4000**	
0.004	0.058	0.455	0.385	0.116	0.453	**100**	**100**
0.003	0.057	0.019	0.177	0.027	0.289	**250**	
0.003	0.050	0.011	0.135	0.023	0.249	**500**	
0.002	0.043	0.008	0.111	0.014	0.178	**1000**	
0.003	0.036	0.008	0.080	0.011	0.128	**2000**	
0.003	0.032	0.006	0.065	0.008	0.097	**4000**	
0.002	0.058	0.700	1.257	0.135	0.486	**100**	**300**
0.002	0.060	0.285	0.321	0.028	0.303	**250**	
0.002	0.049	0.012	0.139	0.016	0.232	**500**	
0.002	0.044	0.007	0.111	0.013	0.177	**1000**	
0.002	0.034	0.004	0.083	0.010	0.132	**2000**	
0.001	0.027	0.004	0.069	0.012	0.117	**4000**	

يتضح من النتائج الواردة في الجدول (2) تباينا ملحوظًا في الأوساط الحسابية لقيم RMSE لمعلمة التمييز تبعا لطول الاختبار وحجم العينة، حيث كان أعلى وسط حسابي له عندما كان حجم العينة يساوي 100 وطول الاختبار 300 فقرة، بينما كانت أقل قيمة للوسط الحسابي عندما كان طول الاختبار 100 فقرة وحجم العينة 4000. أما بالنسبة لدقة تقدير معلمة الصعوبة فقد أظهرت النتائج الواردة في الجدول نفسه بأن هناك تباينا في الأوساط الحسابية لقيم RMSE، حيث كانت أعلى قيمة له عندما كان طول الاختبار 75 و 300 فقرة وحجم العينة 100، بينما كانت اقل قيمة له عندما كان طول الاختبار

100 فقرة وحجم العينة 4000، وهي مماثلة لنتيجة معلمة التمييز. وتجدر الإشارة كذلك بأن النتائج الواردة في الجدول 2 قد أظهرت تقاربا قريبا في الأوساط الحسابية لقيم RMSE المتعلقة بمعلمة التخمين.

وللتعرف إلى دلالات الفروق بين متوسطات قيم RMSE لتقدير معلمة التمييز، والصعوبة، والتخمين باختلاف طول الاختبار وحجم العينة، فقد استُخدم تحليل التباين الثنائي لكل معلمة على حدة. ويظهر الجدول 3 نتائج هذا التحليل.

الجدول 3

نتائج تحليل التباين الثنائي للكشف عن دلالة الفروق بين متوسطات دقة تقدير معالم الفقرات (التمييز، والصعوبة ، والتخمين) حسب متغيري طول الاختبار وحجم العينة

المعلمة	مصدر التباين	مجموع المربعات	درجات الحرية	وسط المربعات	قيمة F	الدلالة الإحصائية	الدلالة العملية η^2
التمييز	طول الاختبار	0.081	5	0.016	4.776	0.000	0.013
	حجم العينة	19.686	5	3.937	1166.033	0.000	0.768
	التفاعل	0.819	25	0.033	9.704	0.000	0.121
	الخطأ	5.956	1764	0.003			
	الكلي	26.542	1799				
الصعوبة	طول الاختبار	12.835	5	2.567	28.281	0.000	0.074
	حجم العينة	62.032	5	12.406	136.683	0.000	0.279
	التفاعل	53.661	25	2.146	23.648	0.000	0.251
	الخطأ	160.114	1764	0.091			
	الكلي	288.642	1799				
التخمين	طول الاختبار	0.034	5	0.007	252.153	0.000	0.417
	حجم العينة	0.132	5	0.026	981.965	0.000	0.736
	التفاعل	0.017	25	0.001	24.952	0.000	0.261
	الخطأ	0.048	1764	2.70E-5			
	الكلي	0.231	1799				

تشير النتائج الواردة في الجدول (3) بأن هناك فروقا ذات دلالة إحصائية عند مستوى الدلالة (α = 0.05) بين المتوسطات الحسابية لقيم RMSE لمعالم الفقرات (التمييز، والصعوبة، والتخمين) تعزى لكل من طول الاختبار وحجم العينة والتفاعل بينهما، وتشير النتائج كذلك بأن حجم العينة كان أكثر إسهاما في تباين قيم RMSE لمعالم الفقرات من طول الاختبار، وذلك من خلال النظر إلى قيم الدلالة العملية (مربع ايتا) حيث أسهم بـ (77%، 28%، 74%) في تباين دقة تقدير معالم الفقرات التمييز، والصعوبة والتخمين على الترتيب.

وللكشف عن مواقع الفروق بين المتوسطات الحسابية لقيم RMSE في تقدير معالم الفقرات العائد لحجم العينة، فقد استُخدم اختبار «شافيه» للكشف عن تلك الفروق لكل معلمة من معالم الفقرات، ويوضح الجدول 4 هذه الفروق.

الجدول 4

نتائج المقارنات الثنائية بين الأوساط الحسابية لقيم RMSE لمعالم الفقرات حسب حجم العينة

المعلمة	حجم العينة	الوسط الحسابي	حجم العينة					
			100	250	500	1000	2000	4000
التمييز	100	0.435		0.114*	0.176*	0.235*	0.277*	0.309*
	250	0.314			0.062*	0.121*	0.163*	0.195*
	500	0.245				0.059*	0.101*	0.133*
	1000	0.191					0.042*	0.074*
	2000	0.144						0.032*
	4000	0.107						
الصعوبة	100	1.050		0.404*	0.462*	0.497*	0.521*	0.537*
	250	0.364			0.058	0.093*	0.117*	0.133*
	500	0.203				0.035	0.059	0.075
	1000	0.138					0.024	0.040
	2000	0.112						0.016
	4000	0.092						

التخمين								
100	0.053		0.002*	0.007*	0.013*	0.018*	0.024*
250	0.050			0.005*	0.011*	0.016*	0.022*
500	0.048				0.005*	0.011*	0.017*
1000	0.043					0.006*	0.011*
2000	0.038						0.006*
4000	0.034						

* دال إحصائيا عند مستوى الدلالة $\alpha = 0.05$

يكشف الجدول (4) أن هناك فروقا ذات دلالة إحصائية ($\alpha = 0.05$) بين الأوساط الحسابية لدقة تقدير معلمة التمييز المقدرة تبعا لكل مستوى من مستويات حجم العينة، حيث كان أعلى فرق عندما كان مستوى حجم العينة 100 و 4000 ولصالح مستوى حجم العينة 4000، حيث كان الوسط الحسابي لدقة تقدير معلمة التمييز للفقرة يساوي 0.115، مما يعكس فعالية أكبر لدقة تقدير معلمة التمييز، بينما كان الوسط الحسابي لدقة تقدير معلمة التمييز عند مستوى حجم العينة 100 هو الأعلى، حيث بلغت قيمته 0.424 ومن ثَمَّ فإن مثل هذا الحجم من العينة عكس فعالية قليلة لدقة تقدير معلمة التمييز.

أما بالنسبة لمعلمة الصعوبة، فقد أظهرت نتائج المقارنات الثنائية الواردة في الجدول (4) بأن هناك فروقا دالة إحصائيا بين الأوساط الحسابية لدقة تقدير معلمة الصعوبة عندما كان مستوى حجم العينة 100مع بقية المستويات (250، 500، 1000، 2000، 4000). وتظهر النتائج كذلك بأن هناك فروقا دالة إحصائيا بين الأوساط الحسابية لدقة تقدير معلمة الصعوبة عندما كان حجم العينة 250 وبين مستويات حجم العينة (1000، 2000، 4000)، في حين لم يكن هناك فروق دالة إحصائيا بين الأوساط الحسابية لدقة تقدير معلمة الصعوبة عندما كان مستوى حجم العينة (500، 2000، 4000) مما عكس دقة في تقدير معلمة الصعوبة كلما زاد حجم العينة عن 500 فأكثر. وأخيرا تظهر النتائج المتعلقة بدقة تقدير معلمة التخمين بأن جميع الفروق جاءت دالة إحصائيا بين كل مستوى من مستويات حجم العينة، ولكن ازدادت دقة التقدير بزيادة حجم العينة.

وتجدر الإشارة كذلك أنه استخدم اختبار شافيه للكشف عن مواقع الفروق بين الأوساط الحسابية لدقة تقدير معالم الفقرات العائد لطول الاختبار، ويبين الجدول(5) تلك الفروق.

الجدول 5

نتائج المقارنات الثنائية بين الأوساط الحسابية لقيم RMSE لمعالم الفقرات حسب طول الاختبار

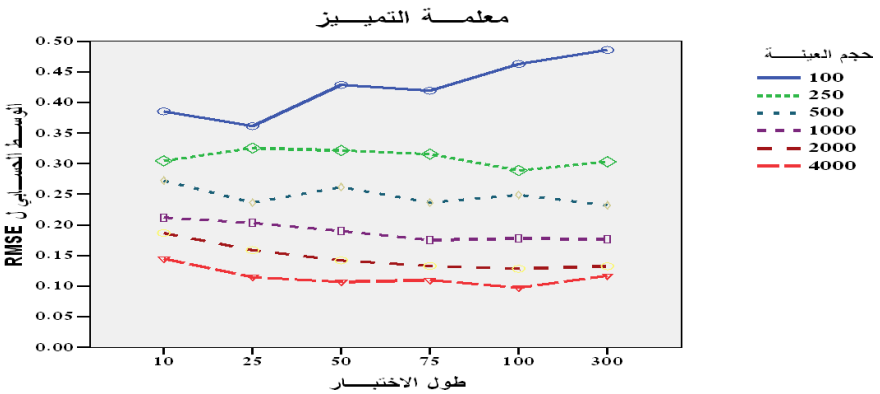
المعلمة	طول الاختبار	الوسط الحسابي	طول الاختبار					
			10	25	50	75	100	300
التمييز	10	0.279		0.018*	0.009	0.019*	0.017*	0.010
	25	0.227			0.009	0.002	0.001	0.008
	50	0.212				0.010	0.008	0.001
	75	0.241					0.002	0.010
	100	0.253						0.007
	300	0.224						
الصعوبة	10	0.234		0.038	0.032	0.148*	0.003	0.168*
	25	0.156			0.006	0.186*	0.035	0.206*
	50	0.375				0.180*	0.029	0.200*
	75	0.265					0.151*	0.021
	100	0.421						0.171*
	300	0.508						
التخمين	10	0.037		0.011*	0.007*	0.004*	0.001	0.000
	25	0.043			0.004*	0.007*	0.012*	0.011*
	50	0.048				0.002*	0.008*	0.007*
	75	0.048					0.006*	0.005*
	100	0.042						0.001
	300	0.049						

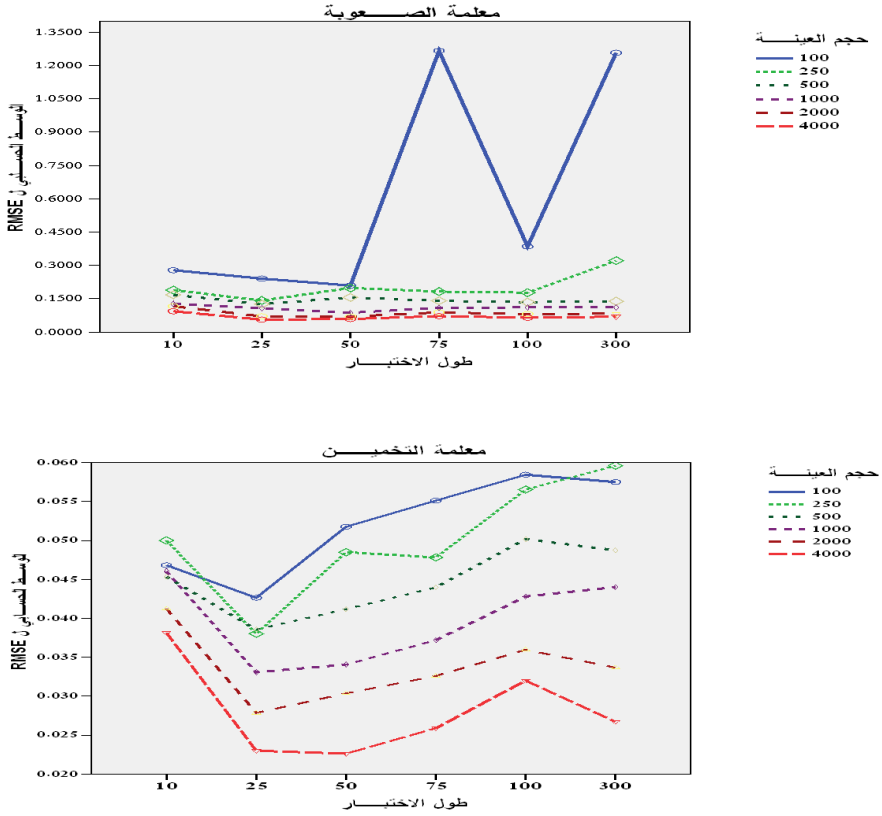
تشير النتائج الواردة في الجدول (5) أن هناك فروقا ذات دلالة إحصائية بين الوسط الحسابي لدقة تقدير معلمة التمييز عند مستوى طول للاختبار 10 فقرات وبين الوسط الحسابي لدقة تقدير معلمة التمييز عندما كان طول الاختبار (75، 100 فقرة) ولصالح الاختبار 100 فقرة، حيث كان الأعلى دقة في تقدير معلمة التمييز، إذ بلغت قيمة الوسط الحسابي لدقة تقدير معلمة التمييز عند هذا الطول 0.097، في حين لم تكن هناك فروق ذات دلالة إحصائية بين متوسطات الحسابية عندما كان طول الاختبار (25، 50، 75، 100، 300) ومن ثَمَّ عكست مثل هذه الأطوال من الاختبارات دقة عالية في تقدير معلمة التمييز. أما بالنسبة لدقة تقدير معلمة الصعوبة فقد أظهرت نتائج المقارنات بأن

هناك فروقا ذات دلالة إحصائية بين الوسط الحسابي لدقة تقدير معلمة الصعوبة عندما كان طول الاختبار 10 فقرات وبين الوسط الحسابي للمعامل نفسه عندما كان طول الاختبار 75 و 300 فقرة. وتشير النتائج كذلك وجود فروق دالة بين الوسط الحسابي لدقة تقدير معلمة الصعوبة عندما كان طول الاختبار 25 فقرة والوسط الحسابي عندما كان طول الاختبار 75 و 300 فقرة وبالمثل عندما كان طول الاختبار 50 فقرة، وأيضا كان هناك فرق دال إحصائيا بين الوسط الحسابي لدقة تقدير معلمة الصعوبة عندما كان طول الاختبار 75 فقرة والوسط الحسابي للاختبار المكون من 100 فقرة، وكذلك وجود فرق دال إحصائيا بين الوسط الحسابي لدقة تقدير معلمة الصعوبة عندما كان طول الاختبار 100 فقرة والوسط الحسابي للمعامل نفسه عندما كان طول الاختبار 300 فقرة، ولكن بشكل عام كان أكبر فرق بين الأوساط الحسابية لدقة تقدير معلمة الصعوبة عند الأطوال 25 فقرة، 50 فقرة و300 فقرة لصالح الاختبار 25 فقرة حيث كان الوسط الحسابي لدقة تقدير معلمة الصعوبة عند هذا الطول 0.124 يليه الاختبار المكون من 50 فقرة، حيث بلغت قيمة الوسط الحسابي لدقة تقدير معلمة الصعوبة 0.130

أما بالنسبة للفروق بين الأوساط الحسابية لدقة تقدير معلمة التخمين، فقد أظهرت نتائج الجدول (5) بأن هناك فروقا دالة إحصائيا بين الأوساط الحسابية لدقة تقدير معلمة الفقرة لكل أطوال الاختبار باستثناء الفرق بين الوسط الحسابي لدقة تقدير معلمة التخمين عند طول 25 فقرة، والوسطين الحسابيين عند طول 100 و 300 فقرة، وكذلك لم يكن الفرق بين الوسط الحسابي لدقة تقدير معلمة التخمين عند مستوى طول100 فقرة والوسط الحسابي للمعامل نفسه عند طول 300 فقرة، وقد كان الاختبار المكون من 25 فقرة الأكثر دقة في تقدير معلمة التخمين إذ بلغت قيمة الوسط الحسابي لدقة تقدير معلمة التخمين عند هذا الطول 0.034

وتجدر الإشارة كذلك بأن نتائج تحليل التباين الثنائي الواردة في الجدول (3)، أظهرت أن هناك فروقا دالة إحصائيا لتفاعل كل من طول الاختبار وحجم العينة فى دقة تقدير معالم





الشكل 1

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

يلاحظ من الشكل (1) بشكل عام بأن أعلى دقة لتقدير معالم الفقرات(التمييز، والصعوبة، والتخمين) لجميع أطوال الاختبار كانت عند مستوى حجم العينة 4000، في حين تراجعت دقة تقدير معالم الفقرات لتصبح الأقل عندما أصبح مستوى حجم العينة 100، والذي يشير إلى أن مثل هذا الحجم من العينات لا يناسب استخدام النموذج اللوجستي ثلاثي المعلمة لتدريج الفقرات. وتجدر الإشارة كذلك بأن أفضل التقديرات لمعالم الفقرات كانت لدى الأطوال (25، 50، 75، 100 فقرة) عند مستويات حجم العينة (1000، 2000، 4000)، ونقطة أخرى جديرة بالذكر من خلال الشكل (1) بأنه إذا استُخدم النموذج ثلاثي المعلمة مع مستويات أطوال كبيرة للاختبارات على غرار بنوك الأسئلة فهذا يتطلب

أن يكون حجم العينة كبير أعلى من 2000 فرد للوصول إلى أفضل التقديرات لمعالم الفقرات.

وللإجابة عن السؤال الثاني للدراسة لمعرفة أثر كل من طول الاختبار وحجم على العينة في دقة تقدير معلمة الفرد (θ) والتفاعل بينهما، فقد تم أيضا إيجاد الأوساط الحسابية والانحرافات المعيارية لدقة تقدير معلمة قدرة الفرد(RMSE) لكل طول من أطوال الاختبار ولكل مستوى من حجم العينة، حيث يوضح الجدول (6) هذه الأوساط.

الجدول 6

الوسط الحسابي والانحراف المعياري لدقة تقدير معلمة قدرة الفرد باختلاف طول الاختبار وحجم العينة

حجم العينة	طول الاختبار											
	10		25		50		75		100		300	
	الوسط الحسابي	الانحراف المعياري	الوسط الحسابي	الانحراف المعياري	الوسط الحسابي	الانحراف المعياري	الوسط الحسابي	الانحراف المعياري	الوسط الحسابي	الانحراف المعياري	الوسط الحسابي	الانحراف المعياري
100	0.513	0.039	0.389	0.026	0.320	0.017	0.657	0.485	0.349	0.224	0.427	0.369
250	0.523	0.026	0.377	0.013	0.305	0.011	0.313	0.012	0.277	0.008	0.299	0.008
500	0.527	0.018	0.374	0.011	0.297	0.008	0.293	0.009	0.280	0.006	0.270	0.005
1000	0.514	0.011	0.370	0.008	0.295	0.007	0.283	0.006	0.270	0.004	0.267	0.003
2000	0.515	0.007	0.370	0.007	0.297	0.005	0.284	0.005	0.265	0.003	0.260	0.002
4000	0.518	0.006	0.376	0.005	0.297	0.003	0.284	0.003	0.269	0.002	0.262	0.002

يتضح من النتائج الواردة في الجدول(6)بأن هناك اختلافات قليلة بين المتوسطات الحسابية وتقاربا في الانحرافات المعيارية لدقة تقدير معلمة القدرة، باستثناء الوسط الحسابي لدقة تقدير معلمة الفرد عندما كان الاختبار مؤلفا من 10 فقرات لجميع مستويات حجم العينة. وتشير النتائج بشكل عام بأن دقة تقدير معلمة القدرة للفرد تزاد بزيادة طول الاختبار وحجم العينة، حيث كانت أعلى دقة للتقديرات في معلمة القدرة عندما كان الاختبار مكونا من 300 فقرة عند مستوى حجم العينة (2000، 4000).

ولمعرفة دلالة الفروق بين المتوسطات الحسابية لدقة تقدير معلمة القدرة للفرد باختلاف طول الاختبار وحجم العينة والتفاعل بينهما، استُخدم تحليل التباين الثنائي، ويبين الجدول 7 نتائج هذا التحليل.

الجدول 7

نتائج تحليل التباين الثنائي للكشف عن دلالة الفروق بين متوسطات دقة تقدير معلمة القدرة للفرد حسب متغيري حجم العينة وطول الاختبار

مصدر التباين	مجموع المربعات	درجات الحرية	وسط المربعات	قيمة F	الدلالة الإحصائية	الدلالة العملية η^2
طول الاختبار	11.447	5	2.289	193.564	0.000	0.354
حجم العينة	2.788	5	0.558	47.145	0.000	0.118
التفاعل	4.068	25	0.163	13.758	0.000	0.163
الخطأ	20.864	1764	0.012			
الكلي	39.167	1799				

يلاحظ من النتائج الواردة في الجدول(7) بأن هناك فروقا ذات دلالة إحصائية عند مستوى الدلالة ($\alpha = 0.05$) بين المتوسطات الحسابية لدقة تقدير معلمة قدرة الفرد تعزى لكل من طول الاختبار وحجم العينة والتفاعل بينهما. وتشير النتائج كذلك بأن طول الاختبار قد أسهم في تباين قيم RMSE لقدرة الفرد أكثر من حجم العينة والتفاعل بينهما، حيث كانت قيم مربع ايتا لكل منهم على التوالي (35%، 12%، 16%). وللكشف عن مواقع الفروق بن المتوسطات الحسابية لدقة تقدير معلمة الفرد استُخدم اختبار شيفيه للكشف عن تلك المواقع لكل عامل على حدة، ويظهر الجدول 8 مواقع تلك الفروق.

الجدول 8

نتائج المقارنات الثنائية بين الأوساط الحسابية لقيم RMSE لمعلمة قدرة الفرد حسب طول الاختبار و حجم العينة

العامل	طول الاختبار	الوسط الحسابي	طول الاختبار					
			10	25	50	75	100	300
طول الاختبار	**10**	0.518		0.142*	0.217*	0.166*	0.233*	0.221*
	25	0.376			0.075*	0.024	0.091*	0.079*
	50	0.301				0.050*	0.016	0.004
	75	0.352					0.067*	0.055*
	100	0.285						0.012
	300	0.297						

حجم العينة	الوسط الحسابي	حجم العينة					
		100	250	500	1000	2000	4000
100	0.442		0.093*	0.102*	0.109*	0.111*	0.108*
250	0.349			0.009	0.016	0.017	0.014
500	0.340				0.007	0.009	0.006
1000	0.333					0.002	0.001
2000	0.332						0.003
4000							

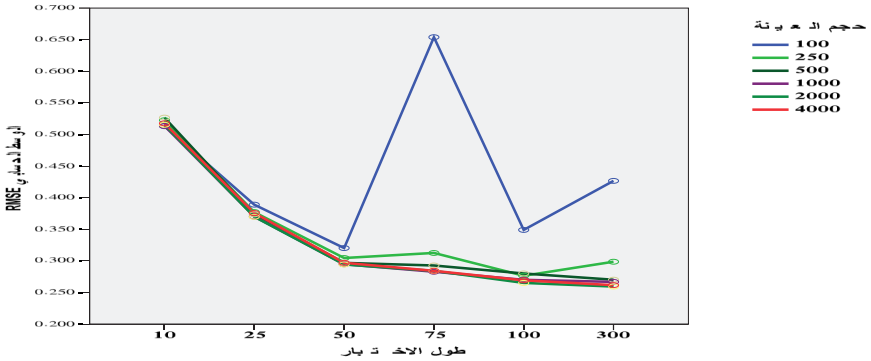
حجم العينة

* دال إحصائيا عند مستوى الدلالة α = 0.05

كشفت نتائج المقارنات الثنائية بين المتوسطات الحسابية لدقة تقدير معلمة الفرد الواردة في الجدول (8) والمتعلقة بعامل طول الاختبار بأن هناك فروقا دالة إحصائيا بين الوسط الحسابي لدقة تقدير معلمة قدرة الفرد عند مستوى طول الاختبار 10 فقرات، وبين الأوساط الحسابية لبقية مستويات طول الاختبار (25، 50، 75، 100، 300 فقرة)؛ إذ جاءت هذه الفروق لصالح الاختبار المكون من 100 فقرة الذي بلغت قيمة الوسط الحسابي له (0.285) ويعكس من ثَمَّ فعالية أكثر في تقدير معلمة قدرة الفرد. وقد كشفت الفروق أيضا أن هناك فرقا دالا إحصائيا بين الوسط الحسابي لدقة تقدير معلمة قدرة الفرد عند مستوى طول 25 فقرة، وبين الوسطين الحسابيين للمعامل نفسه عند مستوى طول 100 و 300 فقرة، وجاء لصالح الاختبار المكون من 100 فقرة، بينما لم يكن هناك فرق دال إحصائيا بين المستوى نفسه (25 فقرة)، وبين الاختبار المؤلف من 75 فقرة. وكشفت النتائج كذلك وجود فروق دالة إحصائيا بين الوسط الحسابي لدقة تقدير معلمة قدرة الفرد عند مستوى طول 50 فقرة والوسط الحسابي للمعامل نفسه عند مستوى طول 75 فقرة ولصالح الاختبار المكون من 50 فقرة، حيث كان أعلى دقة في تقدير قدرة الفرد، وجدير بالذكر بأنه لم تكن هناك فروق دالة إحصائيا بين الأوساط الحسابية لدقة تقدير معلمة قدرة الفرد عند مستويات طول الاختبار (50، 100، 300 فقرة).

أما بالنسبة لعامل طول العينة، فقد بينت نتائج المقارنات الواردة في الجدول 8 أن هناك فروقا دالة إحصائيا بين المتوسطات الحسابية عند مستوى حجم 100 فرد وبقية المستويات لحجم العينة (250، 500، 1000، 2000، 4000)، حيث كان هذا المستوى من حجم العينة الأقل دقة في تقدير معلمة قدرة الفرد، مما يعني أن هذا المستوى من حجم العينة أيضا لا يصلح عند استخدام النموذج اللوجستي ثلاثي المعلمة لتقدير قدرة الفرد ومعالم الفقرات كما ورد سابقا. وأشارت نتائج المقارنة أيضا بأن الفروق بين المتوسطات الحسابية لم تكن دالة إحصائيا بين مستويات حجم العينة (250، 500، 1000، 2000، 4000).

وبالرجوع إلى نتائج تحليل التباين الواردة في الجدول 7، فقد أشارت النتائج بأن هناك أثرا مشتركا لكل من العاملين طول الاختبار وحجم العينة، وقد مُثِّلَ هذا الأثر المشترك بيانيا كما هو موضح في الشكل 2.



الشكل 2

التمثيل البياني للتفاعل بين طول الاختبار وحجم العينة في دقة تقدير قدرة الفرد

يلاحظ من الشكل (2) بأن دقة تقدير قدرة الفرد كانت تزداد عندما كان مستوى طول الاختبار يزيد عن 25 فقرة وحجم العينة أعلى من 250 فردا. ويتضح من الشكل أيضا أن أعلى دقة لتقدير قدرة الفرد كانت عند مستويات طول للاختبار 100 و 300 فقرة عند حجم العينة 4000 فرد، مما يعني أنه عند استخدام النموذج اللوجستي ثلاثي المعلمة مع مستويات طول للاختبارات كبيرة فإننا بحاجة إلى عينات كبيرة لا تقل عن 2000 مفحوص للوصول إلى أفضل التقديرات لمعالم الفرات وقدرة الأفراد.

وعلى الرغم من تباين دقة تقديرات معالم الفقرات وقدرة الأفراد عبر مستويات الطول المختلفة للاختبار ومستويات حجم العينة المختلفة ودلالة هذا التباين، فقد وُجِد وسيط معاملات الارتباط بين المعالم المقدرة والحقيقية لكل مجموعة من مجموعات البيانات التي وُلِّدَت حسب تصميم الدراسة لكل عامل من عوامل الدراسة باعتباره مؤشرا إحصائيا آخر على دقة التقديرات، ويوضح الجدول (9) هذه القيم.

الجدول 9

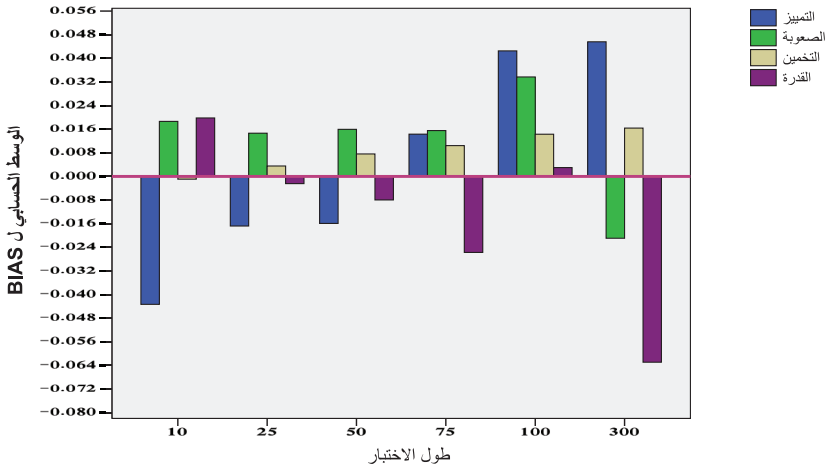
وسيط قيم معاملات الارتباط بين المعالم المقدرة والمعالم الحقيقية

حجم العينة	المعلمة	طول الاختبار					
		10	25	50	75	100	300
100	التمييز	0.565	0.575	0.533	0.582	0.595	0.637
	الصعوبة	0.964	0.965	0.973	0.958	0.974	0.563
	التخمين	0.230	0.466	0.560	0.228	0.485	0.356
	القدرة	0.848	0.927	0.956	0.958	0.970	0.969
250	التمييز	0.737	0.687	0.677	0.740	0.748	0.780
	الصعوبة	0.982	0.983	0.985	0.985	0.986	0.963
	التخمين	0.200	0.580	0.649	0.441	0.546	0.334
	القدرة	0.856	0.933	0.955	0.957	0.964	0.969
500	التمييز	0.830	0.786	0.785	0.819	0.848	0.862
	الصعوبة	0.988	0.989	0.992	0.990	0.992	0.989
	التخمين	0.383	0.638	0.733	0.542	0.647	0.578
	القدرة	0.848	0.927	0.956	0.959	0.965	0.969
1000	التمييز	0.892	0.858	0.874	0.904	0.900	0.915
	الصعوبة	0.991	0.993	0.995	0.994	0.995	0.993
	التخمين	0.401	0.714	0.829	0.641	0.730	0.655
	القدرة	0.858	0.927	0.957	0.962	0.965	0.969
2000	التمييز	0.926	0.914	0.925	0.943	0.948	0.952
	الصعوبة	0.995	0.996	0.997	0.997	0.997	0.996
	التخمين	0.564	0.817	0.867	0.732	0.816	0.765
	القدرة	0.849	0.928	0.958	0.961	0.966	0.765
4000	التمييز	0.960	0.950	0.959	0.963	0.970	0.969
	الصعوبة	0.996	0.998	0.998	0.998	0.998	0.998
	التخمين	0.629	0.874	0.924	0.824	0.854	0.844
	القدرة	0.854	0.928	0.957	0.960	0.966	0.970

يتضح من النتائج الواردة في الجدول (9) بأن قيم معاملات الارتباط بين معلمة القدرة المقدرة والقدرة بشكل عام تحسنت بين المعالم المقدرة والحقيقية بزيادة طول الاختبار

وحجم العينة لتصل إلى شبه الارتباط التام، خصوصا فيما يتعلق بمعلمة التمييز، والصعوبة وقدرة الأفراد. وتظهر النتائج كذلك أن أقل قيم لمعاملات الارتباط بين المعالم المقدرة والمعالم الحقيقة عندما كان حجم العينة 100 مفحوص، في حين كانت أعلى القيم لمعاملات الارتباط عندما كان حجم العينة 4000 مفحوص وطول الاختبار 300 فقرة، ويتضح كذلك بان قيم معاملات الارتباط بين معلمة القدرة المقدرة والحقيقية قد تقاربت عندما زاد طول الاختبار على 10 فقرات لتصبح قيمها أعلى من 0.90 على الرغم من أن قيمه عالية عندما كان طول الاختبار مكون من 10 فقرات، حيث كانت القيم أعلى من 0.80، وهذا يؤكد حقيقة فاعلية النموذج اللوجستي في تدريج قدرة الأفراد مع كافة الأطوال المختلفة للاختبارات سواء كانت قصيرة أو متوسطة أو كبيرة الطول، ولكن للوصول إلى أفضل التقديرات خصوصا فيما يتعلق بمعالم الفقرات، فإن ذلك يتطلب حجم عينة كبير حيث يتضح ذلك من نتائج الجدول نفسه.

وللكشف عن نوع التحيز الموجود في التقديرات لكل من معالم الفقرات وقدرة الأفراد باختلاف طول الاختبار وحجم العينة، وُجِدَ الوسط الحسابي لقيم التحيز(BIAS) لكل معلمة من معالم الفقرات ولقدرة الفرد حسب طول الاختبار وحجم العينة، وتمثيل ذلك بيانيا عن طريق الأعمدة، لكل عامل على حدة، ويوضح الشكل 3 التمثيل البياني للتحيز في التقديرات حسب طول الاختبار.



الشكل 3

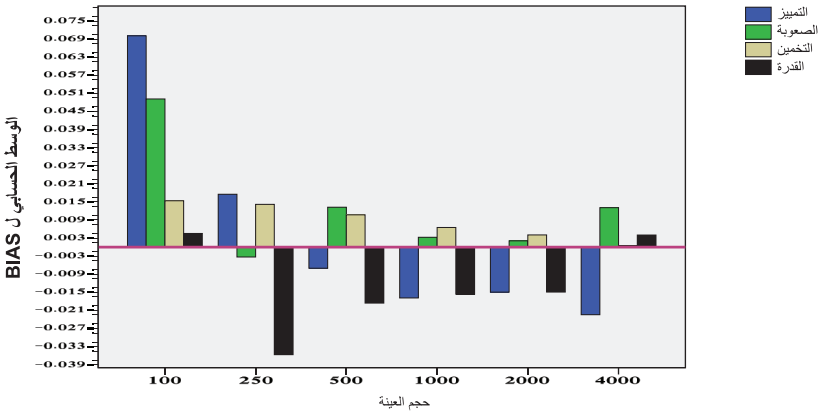
التمثيل البياني للوسط الحسابي لقيم التحيز لكل من معالم الفقرات وقدرة الأفراد حسب طول الاختبار

يظهر من الشكل (3) بأن التقديرات لمعلمة التمييز كانت متحيزة في الاتجاه السالب

عندما كانت أطوال الاختبار مكونة من 10، 25 و 50 فقرة الذي يشير إلى تخفيض (underestimation) التقديرات مما كانت عليه، ومع ذلك فهي كانت قريبة من الصفر عند مستوى طول للاختبار 25 و 50 فقرة، في حين كان التحيز في التقديرات للمعلمة نفسها موجبا عند مستوى طول للاختبار 75، 100 و 300 فقرة، وكانت أعلى قيمة له عندما كان طول الاختبار 300 فقرة. أما بالنسبة لمعامل الصعوبة فقد كان هناك تحيز موجب للتقديرات عند الأطوال (10، 25، 50، 75، 100 فقرة)، وعلى الرغم من المغالاة في التقديرات عند هذه الأطوال، إلا أنها تعد قريبة من الصفر، وقد اظهر أيضا تحيزا سالبا لتقدير الصعوبة عند مستوى طول 300 فقرة وفي الوقت نفسه قريبا من الصفر.

ويظهر من الشكل 3 كذلك بأن التقديرات في معلمة التخمين كانت متحيزة في الاتجاه الموجب لجميع أطوال الاختبار باستثناء الاختبار الذي طوله 10 فقرات، حيث كان التحيز في الاتجاه السالب ويساوي تقريبا صفر، ولكن بشكل عام كان التحيز في التقديرات لهذا المعلم قريبا من الصفر ولجميع أطوال الاختبار. أما بالنسبة للتحيز في تقديرات القدرة للأفراد فقد أظهر الشكل 3 أن هناك تحيزا سالبا عند مستوى طول للاختبار (25، 50، 75، 100) وموجبا عند مستوى طول (10، 300 فقرة) وقد كانت قيم التحيز لتقديرات معلمة القدرة قريبة من الصفر عندما كان طول الاختبار (25، 50 ، 100 فقرة)، وتخفيضا في التقديرات عندما أصبح طول الاختبار 300 فقرة.

أما فيما يتعلق بعامل حجم العينة، فقد تم أيضا تمثيل الأوساط الحسابية للتحيز في تقديرات معالم الفقرات وقدرة الأفراد بيانيا كما يتضح في الشكل 4.



الشكل 4

التمثيل البياني للوسط الحسابي لقيم التحيز لكل من معالم الفقرات وقدرة الأفراد حسب حجم العينة

يتضح من الشكل (4) بأن هناك تحيزًا موجبا في تقدير معلمة التمييز عندما كان مستوى حجم العينة (100، 250) حيث كانت أعلى قيمة له عند مستوى حجم عينة 100، في حين أظهرت التقديرات لمعلمة التمييز تحيزا سالبا عند بقية مستويات حجم العينة (500، 1000، 2000، 4000)، وكان قريبا من الصفر عند 250، 2000 مفحوص. وقد أظهر الشكل تحيزا موجبا في تقدير معلمة الصعوبة عند جميع مستويات حجم العينة باستثناء حجم العينة 250، وقد كانت أقل قيمة له عندما كان مستوى حجم العينة 2000 و 1000 مفحوص. وعلى الرغم من ذلك فقد كانت تلك القيم قريبة من الصفر. أما بالنسبة لتقديرات معلمة التخمين فقد أظهرت التقديرات تحيزا موجبا عند جميع مستويات حجم العينة، وقد أخذ بالتناقص مع زيادة حجم العينة وقيمه كانت تقترب من الصفر عندما زاد حجم العينة على 1000 مفحوص. وفيما يتعلق بتقديرات معلمة القدرة، فإنه يتضح من الشكل نفسه بأنه كان هناك تحيز سالب عند مستويات حجم العينة (250، 500، 1000، 2000)، في حين كان التحيز في التقديرات موجبا عند المستويين من حجم العينة (100، 4000) وكانت قيمه الأقل عند هذين المستويين.

مناقشة النتائج:

أظهرت النتائج المتعلقة بالسؤال الأول بأن طول الاختبار وحجم العينة يؤثران في دقة تقدير معالم الفقرات (التمييز، والصعوبة، والتخمين). وأهم ما توصلت إليه الدراسة الحالية وجود الأثر لتفاعل كل من طول الاختبار وحجم العينة في دقة تقدير معالم الفقرات. وإن المتصفح للنتائج الواردة في الجدول (2) يجد أن هناك نزعة عامة في دقة تقدير معالم الفقرات، حيث يلاحظ أن الدقة في تقدير معالم الفقرات تزداد بزيادة طول الاختبار وحجم العينة، ومثل هذه النتيجة جاءت متفقة إلى حد كبير مع نتائج الدراسات (الدرابيع ، 2010 ؛ Hambleton & Cook, 1980; Hulin, Lisak & Drasgow, 1982). إلا أن ما توصلت إليه الدراسة الحالية في أن الدقة في معالم الفقرات تكون أعلى ما يمكن عندما يكون طول الاختبار 100 فقرة وحجم العينة 4000 مفحوص لا تتفق مع ما أشارت إليه نتائج الدراسات التي تناولت أثر طول الاختبار وحجم العينة على دقة تقدير معالم الفقرات باستخدام النماذج ثنائية التدريج، حيث إن هذه الدراسات تناولت مستويات طول تراوحت ما بين (10 إلى 80 فقرة) ومستويات لحجم العينة تراوحت ما بين 50 إلى 2000 مفحوص.

ومن النتائج المهمة التي توصلت إليها هذه الدراسة ولم تتطرق لها نتائج الدراسات السابقة أهمية حجم العينة باعتباره عاملا مهما في دقة تقدير معالم الفقرة؛ إذ أسهم في تباين دقة تقدير معالم الفقرات أكثر من عامل طول الاختبار، وهذا ما أظهره مؤشر الدلالة العملية مربع ايتا حيث كانت قيمه لمعلمة التمييز، والصعوبة والتخمين (0.74 ، 0.28، 0.77) على التوالي. وهذا يؤكد حقيقة أهمية تمثيل العينة لخصائص المجتمع المستهدف فيما يتعلق بالاختبارات والمقاييس التي تلعب دورًا مهما في دقة القياس المرتبطة مع أخطاء القياس، حيث يقل الخطاء العيني كأحد مصادر أخطاء القياس بزيادة حجم العينة،

مما يعني ضرورة التخطيط الجيد لاختيار عينة المعايرة عندما يأخذ بعين الاعتبار دقة تقدير معالم الفقرات. ومن الجدير بالذكر أيضا بأن المتصفح للنتائج الواردة في الجدول 2 يجد بأن زيادة حجم العينة كان تأثيره على دقة تقديرات معلمة الصعوبة والتخمين أقل من تأثيره في دقة تقدير معلمة التمييز، ويعزى السبب في ذلك بأن معامل التمييز هو في الأصل معامل ارتباط، وحجم العينة يؤدي دورًا في قيمة معامل الارتباط. وفيما يتعلق بعامل طول الاختبار يلاحظ بأن أثره في دقة تقديرات معالم الفقرات كان قليلا مقارنة بحجم العينة (انظر الجدول 2)، حيث يوجد تغير طفيف في دقة تقديرات معالم الفقرات عندما تغير طول الاختبار من 25 فقرة إلى 300 فقرة، وهذا يعزز الأساس النظري لقابلية النموذج اللوجستي ثلاثي المعلمة من حيث إنه يعد الأكثر فعالية في تقدير معالم الفقرات عند استخدام مستويات متنوعة من أطوال الاختبارات (قصيرة، متوسطة، طويلة) مع مستويات كبيرة لحجم العينة (1000، 2000، 4000).

إضافة لذلك، فقد أشارت نتائج الدراسة إلى نتيجة مهمة فيما يتعلق بوجود فروق ذات دلالة إحصائية لتفاعل كل من طول الاختبار وحجم العينة على دقة تقديرات معالم الفقرات، فكما يتضح من الشكل (1) بأن هناك تأثيرا مشتركا لكلا العاملين في دقة تقدير معالم الفقرات، حيث يتضح بأن أعلى دقة لمعالم الفقرات كانت عند مستوى حجم العينة 4000 مفحوص لتصبح الأقل عند مستوى حجم العينة 100، وكانت الأفضل عندما كان مستوى طول الاختبار 25 ، 50، 75، 100 فقرة. وهذه النتيجة تؤكد أهمية الانتباه إلى حجم العينة عندما يتطلب الأمر استخدام النموذج اللوجستي ثلاثي المعلمة في معايرة فقرات الاختبار بأن لا يقل حجم العينة عن 1000 مفحوص للوصول إلى أفضل التقديرات، ونقطة أخرى جديرة بالذكر أنه عند استخدام النموذج نفسه لمعايرة فقرات الاختبارات الطويلة على غرار الاختبارات المكيفة وبنوك الأسئلة فإنه يجب أن يكون حجم العينة أكبر من 2000 مفحوص وليس 1000 مفحوص كما أشارت نتائج الدراسات السابقة من مثل دراسة (Hambleton & Cook, 1980)، مما يعني أن استخدام عينات صغيرة الحجم في معايرة فقرات الاختبار يؤثر سلبا على دقة التقديرات ودقة القياس.

إن النتائج التي وردت في الجدول 9 جاءت متفقة مع نتائج الدراسات (Hulin,Lisak & Drasgow, 1982; Luked & Baur, 2009) التي تناولت معاملات الارتباط بين تقديرات معالم الفقرات (التمييز، والصعوبة، والتخمين) المقدرة والمعالم الحقيقة باختلاف حجم العينة وطول الاختبار باعتباره مؤشرا إحصائيا في دقة تقدير معالم الفقرات، حيث كانت قيم معاملات الارتباط بين معالم الفقرات المقدرة والحقيقة تزداد بزيادة حجم العينة وطول الاختبار لتصل إلى شبه الارتباط التام عندما كان حجم العينة 4000 وطول الاختبار 300 فقرة، وجدير بالذكر أن هذه النتيجة جاءت متفقة مع نتائج فحص دلالة الفروق بين المتوسطات الحسابية لدقة تقدير معالم الفقرات؛ حيث كانت لصالح مستوى حجم العينة 4000 مفحوص، وأن هذا التوافق بين دلالة الفروق بين الأوساط الحسابية لدقة تقدير معالم الفقرات ومعاملات الارتباط بين القيم المقدرة والحقيقية، يؤكد أهمية حجم العينة في دقة تقدير معالم الفقرات عند استخدام النموذج اللوجستي ثلاثي المعلمة في تدريج فقرات الاختبار، وفاعلية النموذج في تقدير معالم الفقرات.

أما فيما يتعلق بنتائج السؤال الثاني المتمركز حول دقة تقديرات معلمة القدرة للفرد، فقد أشارت النتائج الواردة في الجدول 7 وجود أثر لكل من طول الاختبار، وحجم العينة، والتفاعل بينهما في دقة تقدير معلمة الفرد. وإن نتائج هذا التفاعل تشير بشكل عام أنه عند استخدام حجم عينة أكبر من 250 فرد فإن دقة التقدير في معلمة قدرة الفرد تميل إلى التقارب، وهذا التقارب يزداد بشكل واضح عندما يكون طول الاختبار 50 فقرة فما فوق وحجم العينة 1000 فرد فما فوق. واتفقت هذه النتيجة إلى حد ما مع نتائج الدراسات السابقة (الدرابيع ، 2010 ؛ Hambleton & Cook, 1980; Hulin, Lisak & Drasgow, 1982; Glass, 2005) التي أشارت إلى أن حجم عينة 1000 فرد مع طول اختبار 50 فقرة مناسب للحصول على أفضل التقديرات لمعلمة قدرة الفرد، واختلفت مع هذه الدراسات في أنها أظهرت بأن أفضل دقة لتقدير قدرة الفرد عندما كان حجم العينة 2000 و 4000 فرد عند مستوى طول 100 و300 فقرة، وهي متفقة مع النتائج المتعلقة بمعالم الفقرات، مما يؤكد حقيقة فاعلية النموذج اللوجستي ثلاثي المعلمة في دقة تقديرات معالم الفقرات وقدرة الأفراد عند استخدام حجم عينة 2000 فما فوق مع أطوال للاختبارات تزيد على 50 فقرة للوصول إلى أفضل التقديرات.

وبالنظر إلى كل عامل على حدة فقد بينت النتائج بأن عامل الطول في الاختبار كان الأكثر إسهاما في تباين معلمة القدرة من خلال النظر إلى مؤشر الدلالة العملية مربع ايتا (انظر الجدول 7)، فقد أسهم بنسبة 35% في تباين قيم RMSE لمعلمة قدرة الفرد. وهذا ما أكده مؤشر معامل الارتباط بين معلمة القدرة المقدرة ومعلمة القدرة الحقيقة حيث وصلت إلى شبه الارتباط التام عندما زاد طول الاختبار على 10 فقرات كما هو موضح في الجدول 9، واتفقت أيضا مع نتائج دراسة (Hambleton & Traub,1973) بأن النموذج ثلاثي المعلمة يعد أفضل النماذج الثنائية في تقدير معلمة القدرة للفرد. وذلك يعزى بأن النموذج ثلاثي المعلمة هو أقل النماذج الثنائية تشددا ويتضمن معلمة التخمين، حيث يلجا الأفراد ذوو القدرة المتدنية إلى عملية التخمين، فكما يظهر من الشكل 2 بأن أعلى دقة لتقدير معلمة القدرة للفرد كانت عند مستويات طول 100 و 300 فقرة عند حجم العينة 4000 فرد، في حين كانت الأقل دقة عند مستوى حجم العينة 100 فرد وطول اختبار 10 فقرات.

وعلى الرغم من دلالة حجم العينة في دقة تقديرات معلمة الفرد، إلا أن إسهامه كان قليلا في تباين قيم RMSE لمعلمة القدرة، وهذا ما أكدته دلالة الفروق بين الأوساط الحسابية لدقة تقدير معلمة القدرة بأنها لم تكن دالة إحصائيا عند مستويات لحجم العينة (250 ، 500، 1000، 2000، 4000). وبشكل عام أظهر هذا العامل بأن مستوى حجم العينة 100 لا يصلح إذا أراد مطور الاختبار استخدام النموذج اللوجستي ثلاثي المعلمة، حيث إن التقديرات لمعالم الفقرات وقدرة الأفراد عند استخدام هذا المستوى من حجم العينة تكون غير دقيقة ، حيث لوحظ ارتفاع في قيم RMSE لمعالم الفقرات وقدرة الأفراد.

وفيما يتعلق بنوع التحيز الموجود في تقديرات معالم الفقرات وقدرة الأفراد، فقد أظهر الشكلان 3 و 4، بأن التحيز في التقديرات كان متباينا من حيث المغالاة في التقديرات

(Overestimation)، أو تخفيض التقديرات (Underestimation) سواء لمعالم الفقرات أو قدرة الأفراد، ولدى تتبع سلوك التحيز في التقديرات لمعالم الفقرات وقدرة الأفراد عبر مستويات حجم العينة، ومستويات طول الاختبار من خلال الأوساط الحسابية لقيم التحيز(BIAS) لوحظ بشكل عام أن الفرق المطلق بين التقديرات المقدرة لمعالم الفقرات من النموذج اللوجستي ثلاثي المعلمة والمعالم الحقيقية تقترب من الصفر عندما كان طول الاختبار 50 فقرة فأعلى وحجم العينة 2000 و4000 فرد. مما يشير إلى أن النموذج اللوجستي ثلاثي المعلمة يعطي تقديرات دقيقة لمعالم الفقرات وغير متحيزة عندما يزيد طول الاختبار على 50 فقرة ويكون مستوى حجم العينة كبيرا أعلى من 2000 فرد. وهذا ينسجم مع ما أشار إليه لورد(Lord, 1980) بأن النموذج اللوجستي ثلاثي المعلمة يعد نموذجا فعالا عند استخدام اختبارات ذات أطوال متوسطة أو كبيرة مع حجم عينات كبيرة.

كما دلت النتائج المتعلقة بمعلمة قدرة الأفراد من خلال الشكلين 3 و4، بأن هناك تخفيضا في التقديرات (تحيز في الاتجاه السالب) لمعلمة القدرة، ولكن اقتربت الفروق بين الأوساط الحسابية لقيم التحيز في تقديرات معلمة القدرة للفرد من الصفر عندما تغير طول الاختبار من 25 فقرة إلى 100 فقرة عبر مستوى حجم العينة 4000 فرد. وإن مثل هذه النتيجة تؤكد حقيقة تفاعل طول الاختبار مع حجم العينة في إنتاج تقديرات دقيقة سواء المتعلقة بتقديرات معالم الفقرات أو معلمة قدرة الفرد، وهو ما افتقرت إليه الدراسات السابقة في تناول هذا الأثر المشترك لكلا العاملين، وهذا ينسجم نظريا مع الأساس النظري للنموذج اللوجستي ثلاثي المعلمة بأن قدرة الأفراد المقدرة من النموذج لها علاقة بدالة خصائص الاختبار الذي يعتمد على معالم الفقرات التي تتأثر كما أشارت نتائج الدراسة بطول الاختبار وحجم العينة، لما لهما من أثر معاكس في تقليل الانحراف المعياري لأخطاء القياس، ومن ثَمَّ زيادة دقة القياس والذي يترتب عليه إنتاج تقديرات دقيقة لمعالم الفقرات وقدرة الأفراد.

الاستنتاجات والتوصيات:

إن النتائج التي توصلت إليها الدراسة الحالية تشير عموما إلى كفاءة النموذج اللوجستي ثلاثي المعلمة في إنتاج تقديرات دقيقة لمعالم فقرات الاختبار وقدرة الأفراد عندما يستخدم مع اختبارات ذات أطوال تزيد على 50 فقرة مع حجم عينة كبير يزيد على 2000 مفحوص. ذلك لأن درجة تعقيد النموذج تعتمد على عدد المعالم المشتقة منه، والنموذج ثلاثي المعلمة يعد نموذجًا معقدا مقارنة مع النماذج ثنائية البعد الأخرى (أحادي المعلمة، وثنائي المعلمة) مما يترتب عليه الانتباه إلى حجم العينة وطول الاختبار عند استخدامه في المجال العملي للوصول إلى أفضل التقديرات لمعالم الفقرات وقدرة الأفراد، خصوصا في المواقف التي يتطلب منها عمل تأويلات مختلفة بالاستناد إلى درجات الاختبار.

وعلى الرغم من النتائج المهمة التي توصلت إليها الدراسة الحالية، التي تعد بمثابة دليل للباحثين عند استخدام هذا النموذج بمدى أهمية طول الاختبار وحجم العينة باعتبارهما عاملين مهمين في التأثير في دقة التقديرات لمعالم الفقرات وقدرة الأفراد، إلا أن بعض القيود على هذه الدراسة تحد من تعميم النتائج بشكل عام، وذلك لوجود عوامل أخرى قد تؤدي دورًا في تأثيرها في دقة التقديرات، إضافة إلى طول الاختبار وحجم العينة، من مثل طريقة التقدير المستخدمة في تقدير الفقرات وقدرة الأفراد والبرمجية المستخدمة لإنتاج التقديرات، بالإضافة إلى عدد البيانات المولدة في كل خلية من خلايا تصميم الدراسة (Replication). ففي هذه الدراسة ثُبِّتَت البرمجية وطريقة التقدير وعدد مرات التوليد. وفي ضوء ذلك وللتحقق من كفاءة النموذج اللوجستي ثلاثي المعلمة في تقديره لمعالم الفقرات وقدرة الأفراد تحت شروط مختلفة تصبح الحاجة ملحة لإجراء دراسات مماثلة تتناول عوامل مختلفة منها: طرق تقدير الفقرة والأفراد، والبرمجية المستخدمة، وعدد مرات التوليد لتوفير مؤشرات أكثر قوة عن حجم العينة وطول الاختبار المناسبين عند استخدام النموذج اللوجستي ثلاثي المعلمة، ويمكن أيضا إجراء دراسات تتناول مقارنة كفاءة النماذج الثنائية في تقدير معالم الفقرات وقدرة الأفراد تحت شروط حجم العينة وطول الاختبار ومتغيرات أخرى.

المراجع

أولا : المراجع العربية:

الثوابيـة، أحمـد محمـود (2010). أثـر حجـم العينـة علـى تقديـر صعوبـة الفقـرة والخطـأ المعيـاري فـي تقديرهـا باسـتخدام نظريـة الاسـتجابة للفقرة.مجلـة جامعـة دمشـق،-225،26 556.

الدرابيـع، ماهـر (2001). فعاليـة النمـوذج اللوغاريتمـي ذي المعلمـة الواحـدة « نمـوذج راش» فـي دقـة تقديـر قـدرة الفـرد ومعامـل صعوبـة الفقـرة باختـلاف حجـم العينـة وطـول الاختبار.مجلـة دراسـات-العلوم الإنسـانية 208-197،1.

عـلام ، صـلاح الديـن (2005). نمـاذج الاسـتجابة للمفـردات الاختياريـة أحاديـة البعـد ومتعـددة الأبعـاد وتطبيقاتهـا فـي القيـاس النفسـي والتربـوي (ط1)، القاهـرة: دار الفكـر العربـي.

ثانياً: المراجع الأجنبية:

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

Agresti, A., & Finnlay, B. (2009). Statistical methods for the social science (4th Ed). Upper Saddle River, NJ: Person- Prentice Hall.

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

Barnes, L. B., & Wise, S. L.(1991). The utility of a modified one- parameter IRT model with small sample sizes. Applied Measurements in Education,4,143-157.

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.

De Gruijter, Daton. M. & Van der Kamp, L. J. Th.(2005). Statistical Test Theory for Education and Psychology.

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

Farish, Stephen.J. (1984). Investigating item stability: An empirical investigation into the variability of item statistics under conditions of varying sample design and sample size. Occasional paper No 18. Australian Council for Educational Research, Hawthorn.[on-line]. http://eric.ed.gov.

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.

Glass, G .(2005). The impact of item parameter estimation of computerized adaptive testing with item cloning. Law School Admission Council Computerized Testing Report 02-06 November.

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

Hambleton, R .K, and Cook, L .(1980). The Robustness of Latent Trait Models and effect of test length and sample size on the precision of ability estimate. In D. J. Weiss(Ed), Proceding of the 1979 Computerized Adaptive Testing Conferenc, 36-52. Minneapolis, Minnesota: University of Minnesota Department of Psychology. Psychometric Methods Program , Computerized Adaptive Testing Laboratory.

Hambleton , R .K & Jones, R. W.(1994). Item parameter estimation errors and their influence on test information function. Applied Measurement in Education,3, 171-186.

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

Hambleton, R. K., & Traub, R .(1971). Information curves and efficiency of Three Logistic Test Model. British Journal of Mathematical and Statistical Psychology, 24, 273-281.

Hambleton, R. K., & Traub, R .(1971).Analysis of empirical data yusing tow logistic latent trait models. British Journal of

Mathematical and Statistical Psychology, 26, 195-211

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.

Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT Models. Educational Measurement: Issues and Practice, 8, 35-41.

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

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.

Hulin, C. L.,Lissak, R. L.,& Drasgow, F. (1982). Recovery of two-and three-parameter logistic item characteristic curve: A Monte Carlo study. Applied Psychological Measurements, 6, 249-260.

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.

Kirisci, L. , Hsu, T. ,& Yu, L.(2001).Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. Applied Psychological Measurements, 25, 146-162.

Lord, F. M.(1980). Application of item response theory to practical testing problems. New Jersey: Lawrence Erlbaum Associates, Publishers.

Lord, F. M.(1986). Maximum likelihood and Bayesian parameter estimation in item response theory. Journal of Educational Measurement, 23, 157-162.

Lord, F .M, & Novick, M .R.(1968). Statistical Theories of mental

test scores. London, Addison, Wesley: Publishing Company.

McDonnald, R. P.(1999).Test theory: A unified treatment. Mahwah, Nj: Lawrence Erlbaum Assoiates, Inc.

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

Pelton, D. R.(2002). The accuracy of unidimensional measurement models in the presence of deviation for underlying assumptions. Unpublished PhD thesis ,Brigham Young University , Department of Instructional Psychology and Technology.

Rondall, P .(2007). Estimating the standard error of the maximum likelihood abilty estimator in adaptive using the posterior weighted test information. Educational and Psychological measurement, 67, 958- 975.

Stocking, M.L.(1990). Specifying optimum examines for item response theory. Psychometrika,3,461-475.

Stone, M . and Yumoto, F .(2004). The effect of sample size for estimating Rasch / IRT parameters with dichotomous items, Journal of Applied Measurement, 1, 48 - 61.

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

Thissen, D., & Wainer, H. (1983). Some standard errors in item response theory. Pschometrika, 47, 397-412.

Toland, M .D.(2008).Determining the accuracy of item parameter standard error of estimation in Bilog-MG3. Unpublished doctoral dissertation , The University of Nebraska-Lincoln AAT 3317288.

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

Wainer, H., & Mislevy, R. J.(1990). Item response theory, Item calibration, and proficiency estimation, In H. Winer(Ed),

Computerized Adaptive Testing: A Primer. Hillsdale, NJ: Lawrence Erlbaum Associate.

Wilcox, R. R.(1988). Simulation as a research technique . In J. P. Keeves(Ed).Educational Research, Methodology and Measurement :An International Handbook (pp 134-138). New York: Pergamen Press.

Yen, W.(1987). A comparison of the efficiny and acuraacy 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. Lincolnwood, IL: Scientific Software International, Inc

Investigating the Accuracy of Estimation of a Three-Parameter Logistic Model of Item Parameter and Individuals' Ability in Light of Test Length and Sample Size: Simulation Study

Zaid S. Bani Ata

Faculty of Education - Yarmouk University

Irbid - Jordan

Abstract

This study aimed at detecting the accuracy of estimation of a three-parameter logistic model of item parameter and idividuals' ability in light of test length and sample size. To achieve this goal fifty replicated binary responses were generated using WINGEN program on six test lengths (10, 25 ,50, 75, 100, 300) and six sample sizes (100, 250, 500, 1000, 2000,4000).The generated data were analyzed for each cell resulting from the intersection of sample size and test length by Bilog– Mg3 programs to estimate the item parameter, the abilities of persons, and the computed values of RMSE and BIASE.. The results revealed significant differences due to sample size and test length and their interaction in the accuracy estimates of item parameter and ability of persons. The results also revealed that the mean values of RMSE of item parameter and ability of idividuals decreased as the test length exceeded 50 items and the sample size exceeded 2000 subjects. Moreover, the correlation coefficients between estimated and true parameters supported this result, the correlation was almost complete. The value of BIAS was very close to zero

Key words: Simulation**,** Three- Parameter Logistic Model, sample size, test length , estimation accuracy