أثر حجم العينة وطول الاختبار على دقة تقدير معالم الفقرة والقدرة باستخدام نماذج نظرية الاستجابة للفقرة اللامعلمية

حسين عبدالنبي القيسي *

ملخص

هدفت الدراسة إلى الكشف عن أثر حجم العينة وطول الاختبار على دقة تقدير معالم الفقرة والقدرة باستخدام نماذج نظرية الاستجابة للفقرة اللامعلمية، اعتمادا على مؤشري دقة القياس التحيز (BIAS)، والجذر التربيعي لمتوسط مربعات الخطأ (RMSE)، ولتحقيق أغراض الدراسة، تم توليد قدرات لأفراد عينات حجمها 100، و250، و500، و1000، فرد من توزيع طبيعي (1، 0) . وبالاعتماد على معلمة القدرة تم توليد اختبارات وفقاً للتصنيف الآتي (20، و40، و60) فقرة تحت افتراض التوزيع الطبيعي لصعوبة الفقرات (1، 0)، والتوزيع المنتظم لمعلمة التمييز بواقع قيمة ابتدائية (0.4)، وقيمة نهائية (1.2)، والتوزيع المنتظم لمعلمة التخمين بواقع قيمة ابتدائية (0.2) وقيمة نهائية (0.3) على افتراض أن الاختبارات من نوع الاختيار من متعدد، وله أربعة بدائل، تلائم النموذج ثلاثي المعلمة باستخدام برنامج توليد البيانات WinGen .v3.

وللإجابة عن اسئلة الدراسة، استخدمت برمجية TESTGRAF؛ لتقدير معالم الفقرة والقدرة باستخدام طريقة تنعيم النواة KS اللامعلمية، ومن ثم تمت مقارنة المعالم المقدرة مع المعالم الحقيقية (المولدة) باستخدام مؤشري التحيز (BIAS) والجذر التربيعي لمتوسط مربعات الخطأ (RMSE).

وقد أظهرت نتائج الدراسة وجود فروق دالة إحصائياً ($\alpha = 0.05$) في متوسطات مؤشر دقة التقدير (BIAS) في دقة تقدير معلمة التمييز a تعزى لمتغير حجم العينة، كما اظهرت النتائج وجود فروق دالة إحصائية ($\alpha = 0.05$) في متوسطات مؤشر دقة القياس (BIAS) في دقة تقدير معلمة التخمين c تعزى لمتغيري الدراسة (حجم العينة، وطول الاختبار) والتفاعل بينهما، كما بينت النتائج وجود فروق دالة إحصائياً ($\alpha = 0.05$) في متوسطات مؤشر دقة القياس (BIAS) في دقة تقدير معلمة القدرة θ تعزى لمتغيري (حجم العينة، وطول الاختبار)، في حين لم تظهر فروق دالة إحصائياً ($\alpha = 0.05$) في متوسطات مؤشر دقة القياس (BIAS) في دقة تقدير معلمة الصعوبة b.

الكلمات الدالة: (دقة التقدير، ومعالم الفقرة والقدرة، ونماذج نظرية استجابة الفقرة اللامعلمية).

---

* قسم علم النفس، جامعة مؤتة.

تاريخ تقديم البخث: 2014/4/7م. تاريخ قبول البحث: 2015/8/26م.

The effect of the Sample Size and the Llength of Test on the Accuracy Estimation of the Item Parameters By Using Non- Parametric Item Response Theory

Abstract

The Study aimed at investigating the effect of the sample size and length of test on the accuracy estimation of the item-parameter by using the Non- Parametric item- response theory based on the indicators of measurement accuracy (Bias) and Root Mean Square Error (RMSE),To achieve the purpose of the study, the individuals' abilities were generated, The Sample of those participants consisted of (100, 250, 500, 1000) of natural distribution of (0 , 1) according to the category (20, 40, 60) items. Under the assumption of the natural distribution of the item-difficulty (0 , 1) and the distribution for the uniform distribution for the discrimination parameter in a primary value of (0.4) and final value of (1.2) and the uniform distribution for C parameter in a primary value of (0.2) and final value of (0.3) on the assumption of multiple-choice question test that suit the three parameter model using data generating program Win Gen v.3.

To answer the questions of the study ,Test graf programing was used to estimate item- parameter and the ability by using Non-parametric Kernel Smoothing (KS) and then comparing the estimated parameters with the real (generated) ours using Bias and RMSE.

The findings show statistically significant differences at ($\alpha = 0.05$) in the means indicator of Bias attributed to the variable of the sample size . The Findings also indicator statistically significant differences at ($\alpha = 0.05$) in the means of Bias of the estimation in parameter C attributed to the variables of the study (Sample size and test length) and the interaction between them.

In addition, the Findings showed statistically significant differences at ($\alpha = 0.05$) in the means of Bias in the estimation of the ability parameter θ attributed to the mentioned variables, whereas there were no statistically significant differences at ($\alpha = 0.05$) in the means of Bias in the estimation of the difficulty parameter b.

Keywords: Accurate estimation, Item Parameters, Ability, non- Parametric item – response theory.

المقدمة:

يوجد في الوقت الحاضر اطاران يستندان إلى أسس إحصائية للتعامل مع مشكلات القياس النفسي والتربوي هما: النظرية الكلاسيكية (Classical Test Theory: CTT)، ونظرية الاستجابة للفقرة (Item Response Theory: IRT) ويشار إلى نماذج النظرية الكلاسيكية بأنها نماذج ضعيفة كون افتراضاتها يمكن تحقيقها بسهولة في بيانات الاختبار، بينما يشار إلى نماذج نظرية الاستجابة للفقرة على أنها نماذج قوية حيث تستمد قوتها من الافتراضات التي تستند إليها النظرية.

وفي الحقيقة فإن النظريتين تتناولان الأخطاء بطرق مختلفة، فعلى سبيل المثال يمكن افتراض أن الأخطاء تتوزع طبيعيا في نموذج معين، بينما لا يتم افتراض مثل هذا التوزيع للأخطاء في نموذج آخر، وفي نموذج ما فان حجم أخطاء القياس يمكن اعتباره ثابتا عبر تدريج درجات الاختبار بينما في إطار نظرية أخرى يمكن افتراض أن حجم أخطاء القياس يرتبط بالقدرة الحقيقية للمفحوص.

ومن مميزات نظرية الاستجابة للفقرة أن النماذج الخاصة بها تربط الاستجابة على الفقرة بالقدرة وتكون احصائيات الفقرة على نفس تدريج القدرة، وهذا الأمر لا يتوفر في النظرية الكلاسيكية، وكذلك فإن التحديد الدقيق لموقع الفقرة على تدريج القدرة يعطي افضل قياس، بالإضافة إلى وضوح العلاقة بين الأداء على الفقرة والقدرة (Hambleton, 1993).

أن الهدف الأول لأية نظرية في القياس هو تقديم أساس لعمل تنبؤات حول السمات أو القدرات التي يتم قياسها بواسطة فقرات الاختبار، وقد كانت النظرية الكلاسيكية في القياس مستخدمة للوصول إلى هذا الهدف، وفي إطار النظرية الكلاسيكية فإن مفهوم القدرة يتم التعبير عنه بالعلامة الحقيقية وتعرف على أساس أنها توقع العلامة المشاهدة التي يتم الحصول عليها على أساس تطبيق الاختبار على المفحوص عددا كبيرا من المرات، وهنا تقدم نظرية الاستجابة للفقرة أساسا مختلفا، إذ تعتبر القدرات عوامل تؤثر على أداء المفحوصين على فقرات الاختبار، وتوصف العلاقة بين القدرة والأداء على الفقرة في صيغة اقتران منوالي متزايد لكل فقرة وتسمى هذه العلاقة منحنى خصائص الفقرة (Item Characteristic Carve: ICC) وهذه العلاقة هي في أحد نماذجها اقتران لثلاثة معالم: معلم التمييز(a) ويصف ميل المنحنى عند نقطة الانقلاب (Inflection point) أي عند النقطة المناظرة لمعلم الصعوبة (b) عندما يكون المنحنى التراكمي الاوجايف (ogive) سويا، وقيمة معلم التخمين (C) تساوي صفرا، ومعلم الصعوبة هو قيمة (θ) التي تناظر نقطة الانقلاب لمنحنى

الاوجايف، ومعلم التخمين (C) هو المقطع الصادي للتقارب الأدنى للمنحنى أي هو النقطة التي تكون عندها قدرة المفحوص تملك اقل احتمالية للإجابة على الفقرة بشكل صحيح، وهذه المعالم ترتبط بعلاقة رياضية.

وقد بنيت نظرية الاستجابة للفقرة على افتراضات قوية يجب توفرها في البيانات، حتى تؤدي إلى نتائج موثوق بها، ومن أهم هذه الافتراضات افتراض أحادية البعد (Unidimensionality)، ويعني وجود قدرة واحدة تفسر أداء الفرد على الاختبار، أما الافتراض الثاني فيتمثل بالاستقلال الموضعي (Local Independence) ويقصد به أن تكون استجابات الفرد على فقرات الاختبار مستقلة إحصائياً عند مستوى قدرة معين، أي أن استجابة الفرد عن فقرة ما يجب أن لا تؤثر سلبا أو إيجابا على استجابته لفقرة أخرى(Croker & Algina, 1986) . ويذكر هامبلتون وسومينثان (Hambleton & Swaminathan, 1991) إلى أن افتراض الاستقلال الموضعي يكافئ افتراض أحادية البعد، ويعني ذلك أنه إذا تحقق افتراض أحادية البعد في المقياس، فإن المقياس يحقق افتراض الاستقلال الموضعي.

أن أحد أهم المزايا المرتبطة بنظرية الاستجابة للفقرة استقلالية القياس، ويعني ذلك أن تقدير معالم الفقرات يكون مستقلا عن خصائص الأفراد التي استخدمت في تقدير هذه المعالم (Sample Free)، وأن تقدير قدرة الفرد يكون مستقلا عن عينة الفقرات التي تطبق عليه (Item Free) (Hambleton, 1994). وتعد استقلالية القياس بمثابة النقطة المفصلية بين النظرية الكلاسيكية ونظرية الاستجابة للفقرة، وهو ما عبر عنه لورد (Lord, 1980) بخاصية اللاتغاير (Invariance).

ويحاول الفاحص في أي متغير يتعامل معه أن يحدد أرقاما أو رموزا ذات معنى تدل على سمة أو فئة في ذلك المتغير، أو أن يحدد مواقع الأفراد على تدريج معين بحيث يكون لفروق المواقع معنى يتناسب مع الفروق الحقيقية في مقدار ما يمتلكون من السمة موضوع القياس، سواء بمقارنة الموقع للفرد الواحد مع نقطة مرجعية يتم الاتفاق عليها، ويشار إليها بالصفر الافتراضي (Arbitrary Zero)، أو أنها ذات معنى أو مدلول ثابت، وتعني لأي فرد ما تعنيه لأي فرد آخر، ويشار إليها بالصفر المطلق (Absolute Zero). وهذا يعني أن المقياس (Scale) يعرف من

خلال الغرض، ويتلخص في تحديد مواقع الأفراد حسب نوع السمة أو حسب درجة امتلاكهم لها. كما أشار كيرلنجر (Kerlinger, 1973) إلى أن القياس ليس أكثر من لعبة ذات قواعد محددة، يسعى فيها اللاعب للوصول إلى أفضل تماثل (Isomorphism) بين القياس (Measurement) والحقيقة (Reality). وكما هو معروف لا نستطيع أن نعرف هذه الحقيقة، ولكن نعمل على تقدير لهذه الدرجة أو الترتيب والمهم في هذه العملية (التقدير) هو الوصول إلى التوافق التام في المواقع (Rank Isomorphism)، وليس القيم (Value Isomorphism).

وبذلك لا يوجد شك بإمكانية نجاح عملية القياس على مستوى القياس الاسمي أو الرتبي، وذلك أن معظم السمات النفسية والتربوية لا تقع على مقياس النسبة (لعدم توافر الصفر المطلق فيها)، ولا تقع على مقياس فئوي لعدم تساوي الفترات بين القيم (الفرق بين الدرجتين 95 و90 لا يساوي الفرق بين الدرجتين 55 و50). وبذلك فإن القياس النفسي والتربوي يقع على مقياس أرقى من المقياس الرتبي وأقل من الفئوي؛ أي شبه فئوي (Quasi Interval) (عودة، 2010).

ولكن المهتمين في القياس النفسي والتربوي يحللون نتائج القياس بافتراض أنها واقعة على مستوى القياس الفئوي؛ لأن معظم المعالجات الإحصائية اللازمة لتفسير النتائج وتحليلها مبنية على الأقل على مقياس فئوي، ولأن هذه البيانات تمكن المختص من تحويلها تحويلا خطيا، مع العلم أن انتهاك هذا الافتراض (وقوع البيانات على مقياس فئوي) قد يؤدي إلى تشويه النتائج وصعوبة في التفسير.

وتنقسم نماذج نظرية الاستجابة للفقرة إلى نوعين رئيسيين، الأول: يعرف بنماذج نظرية الاستجابة للفقرة المعلمية (Parametric Item Response Theory Models) (PIRT)، حيث يكون شكل دالة استجابة الفقرة (Item Response Function) (IRF) محددة (لوجستية الشكل)، بينما يعرف النوع الثاني بنماذج نظرية الاستجابة للفقرة اللامعلمية (NIRT) (Non- Parametric Item Response Theory Models)، التي لا تحدد شكل محدد لدالة استجابة الفقرة (فقط أن تكون غير متناقصة)، ولا تفترض أي شكل سابق، وكما يشير فان دير لندن وهامبلتون (Van der Linden & Hambleton,1997) يمكن الافتراض أن هذه الدوال أقرب لدوال الاستجابات الحقيقية من تلك التي تعطيها النماذج المعلمية؛ لأنها تعتمد على افتراضات أقل حول النموذج الحسابي.

يوجد اعتبارات عدة يجب مراعاتها عند اختيار النموذج المناسب للبيانات سواء النماذج المعلمية أو النماذج اللامعلمية. فقد شاع استخدام نماذج نظرية الاستجابة للفقرة المعلمية من قبل الباحثين على الرغم من موانعها في تحليل البيانات ذات المستوى الرتبي، إلا أن مصداقية النتائج قد تكون موضع تساؤل عندما لا يتحقق فرض وقوع البيانات على مستوى القياس الفئوي، الأمر الذي تبرره نماذج استجابة الفقرة اللامعلمية والتي لا تضع قيود حول شكل دالة استجابة الفقرة، مما يثير التساؤل حول مدى مطابقة النوعين للبيانات التحصيلية التي يعتبرها البعض رتبية والبعض الآخر يعتبرها شبه فئوية، ومدى دقة النتائج التي تفرزها مثل هذه الاختبارات (Liang, 2010).

إن كبر حجم عينة الأفراد والفقرات التي يتطلبها تطبيق نظرية الاستجابة للفقرة في القياس (Crocker & Algina ,1986) والدالة الرياضية المعقدة التي تفترضها لكل نموذج من نماذجها، أظهرت الحاجة إلى توفر برامج الحاسوب التي تساعد في تحليل بيانتها. ولقد ظهرت برامج حاسوبية كثيرة تستخدم في تحليل بيانات نماذج نظرية الاستجابة للفقرة، ففي عام (1991) ادعى رامسي (Ramsay, 1991) أن إجراءات تقدير برنامج (TESTGRAF) أسرع 500 مرة من البرامج الأخرى دون خسارة في الكفاءة، وأنه يحتاج إلى حجم عينة وطول اختبار أقل من البرامج الأخرى، حيث إن هذا البرنامج يستخدم لنماذج نظرية الاستجابة للفقرة اللامعلمية، كما انتشرت البرامج التي تساعد الباحثين في مختلف العلوم من إنجاز الأبحاث والتوصل السريع للنتائج. ومن بين هذه البرامج، برنامج توليد البيانات في علم القياس والتقويم حيث تطورت برامج التوليد المتعلقة بنظرية الاستجابة للفقرة مع ظهور هذه النظرية في السبعينات ومن أجل تحديد خصائص وفعالية النماذج الرياضية وغيرها من الصيغ الرياضية التي تضمنتها هذه النظرية.

نماذج نظرية استجابة الفقرة اللامعلمية:

تعد نماذج نظرية الاستجابة للفقرة اللامعلمية نماذج إحصائية يمكن استخدامها في القياس النفسي والتربوي لدراسة جودة المقياس وكفاءته من خلال تحليلات جتمان (Guttman Scalogram Analysis). وتستند على مفهوم التراكمية، التي تفترض أن كل فقرة وكل مفحوص يمتلكان موقعا على متصل القدرة، حيث يجيب المفحوص على الفقرة إجابة صحيحة إذا كانت قدرته أعلى من

صعوبة تلك الفقرة، مما يتيح التنبؤ بنمط استجابة المفحوص من معرفة درجته الكلية على الاختبار، وقد لا يتحقق ذلك تجريبيا مما أدى إلى ظهور الفكرة التي جاءت بها النماذج الحديثة لنظرية الاستجابة للفقرة (المعلمية واللامعلمية) والقائمة على الاحتمالية لا التحديد، التي نصت على أن احتمالية الإجابة الصحيحة مرتفعة (لكن $\neq 1$) كلما ازدادت قدرة المفحوص، كما أنها قليلة (لكن $\neq 0$) عند انخفاض قدرة المفحوص (Sijtsma & Hemker, 2002).

كما يشير سيجتسما ومولينار (Sijtsma & Molenaar, 2000) أن النماذج اللامعلمية تقوم على مجموعة من الافتراضات، التي تعد أقل تشدداً من تلك التي تقوم عليها النماذج المعلمية وهي كالآتي:

1- أحادية البعد (Unidimensionality): الاستجابات على الفقرات تتبع متغير كامن أحادي البعد يرمز له بالرمز (θ).

2- الاستقلال الموضعي (Local Independence): احتمالية الإجابة على أي فقرة غير مرتبطة بالاستجابة على أي فقرة أخرى في الاختبار.

3- الإطرادية (Monotonicity): وتعني أنه بازدياد قيمة القدرة θ، تزداد احتمالية الإجابة الصحيحة على الفقرة، أو تبقى ثابتة ضمن مستويات القدرة المختلفة .

4- الإطرادية المضاعفة (Double Monotonicity): وهو الافتراض الأصعب وغير الضروري لتحقق النموذج، والمتضمن امتلاك دوال استجابة غير متقاطعة لفقرات الاختبار التي تشكل التدريج.

تقسم النماذج اللامعلمية إلى قسمين رئيسين، هما:

1- نموذج التجانس الإطرادي (Monotone Homogeneous Model) (MHM) وهو نموذج نظرية الاستجابة للفقرة اللامعلمية والمعروف بنموذج Mokken، ويستخدم لتحليل التدريج للاستجابات الثنائية، وقد وصف سيجتسما (Sijtsma,1998) تحليل التدريج لموكن على أنه نسخة معدلة احتمالية لتحليل التدريج لجتمان، حيث اقترح موكن نموذج التجانس الاطرادي في عام (1971) واشترط أحادية البعد للمقياس، بالإضافة إلى أن تكون دالة

استجابة الفقرة غير متناقصة (الإطرادية) كدالة للقدرة، إذ يختلف بشكل أساسي عن النموذجين ثنائي وثلاثي المعالم في أن دالة استجابة الفقرة ليس بالضرورة أن تأخذ شكلا لوجستيا مما يجعل نموذج موكن أقل تقييدا للبيانات التجريبية من النماذج اللوجستية، كما يتيح هذا النموذج إمكانية ترتيب الأفراد تبعا لمستوى القدرة باستخدام الدرجة الكلية.

2- نموذج الاطراد المضاعف (Double Monotonicity Model) (DMM) هو نموذج موكن الثاني، الذي يفترض جميع افتراضات نموذج التجانس الاطرادي، بالإضافة إلى افتراض عدم تقاطع دوال الاستجابة لفقرات الاختبار، ولكن يسمح لها بالتماس في المناطق المتطرفة، مما يجعل منه نموذجا صعب التحقيق (Sijtsma & Molenaar, 2002).

دقة تقدير معالم الفقرة والقدرة:

إن الخطوة الأساسية والأهم في تطبيق نظرية الاستجابة للفقرة هي تقدير معالم نموذج نظرية الاستجابة للفقرة، التي تحدد خصائص هذا النموذج، كما أن النجاح في عملية التقدير يتوقف على توفير إجراءات مناسبة لتقدير هذه المعالم، سواء في النماذج المعلمية أو اللامعلمية لنظرية الاستجابة للفقرة. وتعتمد الدقة في تقدير المعالم التي يتم الحصول عليها على مبدأين أساسيين، هما: عدم التحيز في التقدير(unbiased)، ومعدل مربعات الأخطاء المتوقعة (Expected Mean Squared Error: EMSE)، حيث يشير التحيز (Bias) إلى توقع الفرق بين تقدير المعلم، وقيمة المعلم الفعلية، ولا يعني تحقق هذا المبدأ أن التقدير دقيق تماما، وذلك لإمكانية وجود تقديرات عدة غير متحيزة، لهذا فإن مبدأ معدل مربعات الأخطاء للتقدير الذي يقاس بواسطة تباين القيمة المقدرة يعطي أسلوبا لاختيار تقدير جيد للمعالم (Casella & Berger, 1990).

وهناك خصائص أخرى مرغوب فيها في التقديرات الإحصائية، منها الاتساق (Consistency) الذي يعني أنه كلما ازداد حجم العينة، فإن قيمة المقدر (Estimator) تقترب من المعلم باحتمالية كبيرة.

وقد رأى وانق و فيزبول (Wang & Vispole, 1998) أن هناك ما لا يقل عن ثلاث حالات يمكن أن يكون التحيز هو الأكثر أهمية من الخطأ المعياري، وهي في مقارنة أوساط المجموعات،

وفي استرجاع تقديرات القدرة من الاختبارات المختلفة على التدريج نفسه، وفي تحديد الإتقان أو عدمه وبالتالي تعد الطريقة التي تعطي تحيزا أقل هي الطريقة المفضلة.

وتشير الدقة إلى المى الذي يتوافق فيه القرار المسند على درجات الاختبار مع القرار الذي يمكن اتخاذه فيما لو كانت الدرجات لا تتضمن أية أخطاء قياس، وبالتالي فإن الدقة لا بد من تقديرها على اعتبار أن الاختبار الذي لا يتضمن أخطاء غير موجود (Crocker & Algina, 1986).

وفي نماذج نظرية الاستجابة للفقرة (اللامعلمية) يعتمد احتمال الإجابة الصحيحة على معلم القدرة (θ)، وعلى معالم الفقرة ذات العلاقة، وهي: معلمة الصعوبة (b)، ومعلمة التمييز (a)، ومعلمة التخمين (c)، حيث إن جميع هذه المعالم غير معروفة، ولذلك فإن مهمة التقدير هي تحديد قيمة (θ) لكل مفحوص بالإضافة إلى معرفة معالم الفقرة من خلال إجابات المفحوصين على فقرات الاختبار. وهناك طرق عدة لتقدير معالم الفقرة والقدرة سواء في النماذج المعلمية أو اللامعلمية، وهذه الطرق تستخدم أساليب التحليل العددي (Numerical Analysis) من خلال برامج يمكن تطبيقها باستخدام الحاسب الإلكتروني، وقد أشار ثيسن ووينر (Thisen & Wainer, 1982) إلى أن تحديد كمية الخطأ في تقدير المعالم ليست عملية سهلة، وذلك بسبب عدم وجود صيغة رياضية محددة تعطي الخطأ في تقدير المعلم كدالة لحجم العينة وعدد الفقرات، ففي أبسط الصيغ الرياضية نحتاج إلى استخدام طرق التكامل العددي (Numerical Integration) لإيجاد الخطأ المعياري لتقدير المعالم.

وهناك طرق عدة للقيام بتقدير معالم نماذج نظرية الاستجابة للفقرة اللامعلمية، ولكن السؤال الذي يبقى قائما والذي يفترض أن تجيب عليه أي نظرية قياس (أو نموذج قياس) هو: ما هي دقة تقدير معالم الفقرة والقدرة؟

وقد بين لورد (Lord, 1980) وجود وسائل عدة للكشف عن دقة تقدير المعالم وجودة الاختبارات، من أبرزها:

- محك معدل مربعات الأخطاء (EMSE أو الخطأ المعياري في التقدير (S.E.E)

- محك الكفاءة النسبية للاختبار (RE) (Relative Efficiency)، ودالة معلومات الاختبار (Test Information Function)(TIF). هذا وسيتم في هذه الدراسة استخدام مؤشرات لتحديد دقة تقدير المعالم في النماذج اللامعلمية لنظرية استجابة الفقرة.

أساليب تقدير معالم الفقرة والقدرة في النماذج اللامعلمي:

هناك أساليب عدة لتقدير معالم الفقرة والقدرة حسب نماذج نظرية الاستجابة للفقرة اللامعلمية، وتختلف هذه الأساليب عن نظيرتها المعلمية في أنها لا تفترض أي شكل حسابي سابق، وتقدر منحنى خصائص الفقرة (ICC)، ثم تجد المعالم من هذا المنحنى، وذلك أنها تفترض وقوع البيانات على مستوى القياس الرتبي، ومن هذه الأساليب ما يأتي:

الانحدار اللامعلمي (Non-Parametric Regression):

أشار دوجلس (Douglas, 1997) أن تقدير دالة استجابة الفقرة ضمن تحليل الانحدار اللامعلمي يتم دون أي افتراضات فيما يخص شكل هذه الدالة (لوجستية الشكل) كما في النماذج المعلمية، حيث يوجد على الأقل طريقتين لتقدير شكل هذه الدالة، وهما:

1. Kernel Smoothing (KS).
2. Isotonic Regression Estimation.

ويضيف أن الأفضلية التطبيقية تعود إلى طريقة (KS)؛ وذلك لوجود برنامج حاسوبي مجاني (TESTGRAF)، الذي يمكن من خلاله تقدير دالة Pi(θ)، حيث يقوم (KS) بحساب Pi(θq)، الذي يدل على احتمالية الإجابة الصحيحة على الفقرة i عند مستوى (q) معين من القدرة من خلال المعادلة الآتية:

$$p_i(\theta_q) = \sum_{a=1}^{N} w_{aq} y_{ima}.$$

حيث إن

$$w_{aq} = \frac{k\left[\frac{\theta_a - \theta_q}{h}\right]y_i}{\sum_{b=1}^{N} k\left[\frac{\theta_a - \theta_q}{h}\right]}$$

w_{aq} : متجهة القدرة للمفحوص a عند مستوى قدرة q، الذي يتم تقديره تبعاً لرتبة المفحوص a مع رتب باقي المفحوصين (N , …….. , b , c).

- y_{ima} : متجه خيار الفقرة الثنائي بطول يساوي N (عدد المفحوصين الكلي) ، والذي يأخذ القيمة 1 في حال اختار المفحوص a الخيار m.
- K : دالة كيرنل التي يمكن تقديرها بطرق عدة باستخدام برنامج (TESTGRAF).
- h : بارامتر التهذيب (Smoothing Parameter)، وهو يعتمد بشكل أساسي على عدد المفحوصين، ويساوي (1.1N) في برنامج TESTGRAF (Ramsay, 2000).

وتم في هذه الدراسة استخدام هذه الطريقة (Kernel Smoothing) لتقدير معالم الفقرة والقدرة، كما يلي:

أولا: يتم الحصول على العلامة الكلية (Xa) لكل متقدم a حيث a =(1، 2، 3….N) وXa هو عدد الإجابات الصحيحة، ويتم إعطاء المتقدمين ذوي العلامات الكلية المتساوية ترتيبا رتبيا بشكل عشوائي.

ثانيا: يتم إعطاء المتقدمين علامات ربعية (كمية) للتوزيع الطبيعي المعياري اعتمادا على رتبهم (Za)، بحيث تكون المساحة تحت دالة الكثافة الطبيعية المعيارية إلى يسار (Za) مساوية a/(N + 1).

ثالثا: يتم تسجيل قيم المؤشر (Yjka =1) عندما يقوم المتقدم (a) باختيار البديل k، وغير ذلك فهو صفر للمتقدم على الرتبة a للبديل k للفقرة j.

وأخيرا تم تقدير احتمال اختيار البديل k للفقرة (j)،[Pjk(θ)] عن طريق تلطيف العلاقة بين متجه المؤشر [Yjka] من الرتبة N ومتجه الكفاءة (Zas) ل(N) من المتقدمين. ويعتبر التلطيف نوعا من المعادلات التي تكون فيها ((Pjk(θ)) عند أي مستوى ل (θ) هو المتوسط الموزون ل (Yjka) للمتقدمين ذوي مستويات القدرة القريبة من (θ) وتظهر معادلة التلطيف التي أعطاها رامسي (1995) بالشكل الآتي:

$$\hat{P}_{jk}(\theta_q) = \frac{\sum_{a=1}^{N} K\left[(\theta_a - \theta_q)/h\right] y_{ikj}}{\sum_{a=1}^{N} k\left[(\theta_a - \theta_q)/h\right]}$$

ولتسريع عملية الحساب دون التضحية بالدقة، فقد تم الحصول على معدل قيم المؤشر $(\hat{P}\,jkr)$ لقيم θ_a التي تقع بين مركزي الفقرات المتجاورة، وكذلك تم حساب المنطقة التي تقع تحت المنحنى الطبيعي المعياري بين مراكز هذه الفترات (ϕ_r). وبعد ذلك تم تبسيط هذه القيم كما بالمعادلة الآتية:

$$\hat{P}_{jk}(\theta_q) = \frac{\sum_{r=1}^{Q} \phi_r k\left[(\theta_q - \theta_r)/h\right] \hat{P}_{ikj}}{\sum_{r=1}^{Q} \phi_r k\left[(\theta_q - \theta_r)/h\right]}$$

وبعد ذلك تم تقدير تباين الخطأ لقيمة المنحنى المقدر بالمعادلة الآتية

$$\hat{S}(\theta_q) = \sum_{a=1}^{N} w_{jka}^2\, \hat{P}_{jk}(\theta_a)(1 - \hat{P}_{jk}(\theta_a))$$

$$w_{jka} = \frac{K\left[(\theta_a - \theta_q)/h\right]}{\sum_{a=1}^{n} K\left[(\theta_a - \theta_q)/h\right]}$$

وبعد أن يتم تقدير منحنى خصائص الخيار (OCC)، يمكن عندها تقدير قدرة المتقدمين للاختبار اعتماداً على هذه المنحنيات. ويمكن تدوير تقديرات القدرة الجديدة للوراء في العملية

كأساس للترتيب وإعادة تقدير المنحنيات، ويمكن الاستمرار بهذه المحاولة حتى تتشابه تقديرات المنحنى والقدرة. وقد ذكر رامسي أن محاولتين إلى ثلاث محاولات تكون كافية للوصول إلى التشابه (Ramsay,2000):

دقة القياس في نماذج نظرية الاستجابة للفقرة اللامعلمية

يشير رامسي (Ramsay, 2000) إلى أنه يتم تقدير دقة القياس في النماذج اللامعلمية من خلال دالة معلومات الفقرة والاختبار كما هو الحال في النماذج المعلمية، حيث تعطى دالة معلومات الفقرة للفقرات الثنائية كما يأتي:

$$I_i(\theta) = \left[\frac{dp_i(\theta)}{d\theta}\right]^2 / \left[p_i(\theta)(1-P_i(\theta))\right]$$

حيث إن Pi(θ): قيمة دالة خصائص الفقرة.

ويتم تحديد دقة القياس في النماذج اللامعلمية من خلال استخدام برنامج (TESTGRAF) برسم متوسط دالة معلومات الفقرة (Mean Item Information Function). وتستخرج دالة معلومات الاختبار بالاعتماد على دوال معلومات فقرات الاختبار كامل حيث تعطى بالمعادلة الآتية:

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta)$$

ثم يستخرج متوسط دالة معلومات الاختبار (Average East Information Function)، حيث يقوم برنامج (TESTGRAF) بعمل مقارنات بين اختبارات مختلفة قد تحتوي أعداداً مختلفة من الفقرات.

الدراسات السابقة:

يتضمن هذا الجزء الدراسات السابقة ذات العلاقة بموضوع الدراسة، التي تم الاطلاع عليها. ومن ثم وضع الباحث تعقيباً عليها يبرز ما تتميز به الدراسة الحالية عنها. وقام الباحث بتناول العديد من الدراسات التي أشارت إلى موضوع دقة القياس، والبحث عن أفضل الوسائل للوصول

بالقياس النفسي إلى درجة من الدقة بما يمكن من اتخاذ قرارات صائبة بالاعتماد على نتائج عملية القياس.

وتمت الإشارة إلى محكات الدقة في تقدير المعالم، والعوامل التي تؤثر في دقة تقدير معالم الفقرة، ومعلمة القدرة، خصوصاً عدد الفقرات وحجم عينة المفحوصين، وأهمية دقة التقدير في بناء الاختبارات واتخاذ القرارات. وفيما يأتي استعراض لأبرز تلك الدراسات:

قام كل من ريكيس ومارك (Reckase & Mark, 1978) بإجراء دراسة باستخدام أسلوب المحاكاة هدفت إلى مقارنة دقة تقدير معلمة القدرة، ومعالم الفقرة في نموذج راش والنموذج اللوجستي الثلاثي، حيث بينت النتائج أن النموذج اللوجستي الثلاثي قد طابق بيانات الاختبار بشكل أفضل من نموذج راش، وأن دقة تقدير معلمة القدرة لنموذج راش أقل من دقة تقدير معلمة القدرة للنموذج اللوجستي الثلاثي. كما بينت الدراسة أن النموذج اللوجستي الثلاثي يحتاج حجم عينة أكبر لمعايرة الفقرات من نموذج راش، وكانت أبرز نتائج هذه الدراسة أن هناك ارتباطاً عالياً بين تقديرات القدرة وفق النموذجين لمعظم البيانات، وأن نموذج راش يفضل استخدامه في حالة العينات الصغيرة.

وأجرى رامسي (Ramsay, 1991) دراسة في الولايات المتحدة الأمريكية هدفت إلى الكشف عن مدى استخدام نماذج كيرنال اللامعلمية، في تقدير رتب القدرة، ولتحقيق هدف الدراسة تم تطبيق طريقة كيرنال على مجموعة من الأمثلة التي تضم متغيرات ثنائية الاستجابة. وقد بينت الدراسة أن طريقة كيرنال من أهم الطرق اللامعلمية، كما بينت الدراسة عدم وجود فقد للبيانات تأثر على دقة تقدير رتب القدرة في المتغيرات ثنائية الاستجابة.

وفي دراسة أعدها فيتزباترك وآن (Fitzpatric & Ann, 2001) هدفت فحص أثر طول الاختبارات، التي تتكون من فقرتين وأربع فقرات وثماني فقرات واثنتي عشرة فقرة وعشرين فقرة على ثبات الاختبار، حيث خصص لكل فقرة درجتين وأربع درجات وست درجات، وقد تم تشكيل عينات حجمها (200، و500، و1000) مفحوص بهدف فحص أثرها (طول الاختبار، وحجم العينة، وطريقة التصحيح) على دقة تقدير المعالم، حيث استخدم أسلوب المحاكاة لأغراض هذه الدراسة، وذلك بتوليد استجابات المفحوصين على اختبار في الرياضيات، وقد تم استخدام معيار الجذر التربيعي لمعدل مربعات الفروق (RMSD)؛ بهدف تقييم دقة تقديرات المعالم. وتوصلت الدراسة إلى

نتائج عدة، منها: أن دقة تقدير المعالم عندما تم استخدام عينة حجمها (200) مفحوص كانت أقل من دقة تقدير المعالم عندما تم استخدام عينة حجمها (500) مفحوص. وكانت دقة تقدير المعالم عندما تم استخدام عينة حجمها (200) مفحوص وعينة حجمها (500) مفحوص أقل من دقة تقديرات المعالم عندما تم استخدام عينة حجمها (1000) مفحوص؛ أي أن دقة التقدير تزداد بازدياد حجم العينة.

وأجرى كونينغ وسيجستما وهامرز (Koning & Sijtsma & Hamers, 2002) دراسة بهدف المقارنة بين نموذجين معلميين وآخرين لامعلميين من نماذج استجابة الفقرة للتعرف إلى الفائدة المرجوة منها في تحليل بيانات الاختبار. وقد قام الباحثون بتطبيق اختبار للاستنتاج الاستقرائي على عينة مكونة من (478) طالباً وطالبة من طلبة الصف الثالث الأساسي. وقد تم تحليل البيانات باستخدام النماذج الآتية: النموذج المعلمي أحادي المعلمة (راش)، ونموذج فيرهيلست المعلمي (Verhelst model)، ونموذج موكن الاطرادي، ونموذج موكن المضاعف الاطرادي. وأظهرت النتائج أفضلية للجمع بين النوعين من النماذج المعلمية واللامعلمية، إذ قدمت النماذج اللامعلمية تدريجات رتبيه للفقرات والأفراد، كما قدمت النماذج المعلمية معلومات مفيدة حول خصائص الفقرات بالإضافة لفائدتها في بعض الجوانب التطبيقية كمعايرة درجات الاختبار، فالنماذج بنوعيها المعلمية واللامعلمية قدمت معلومات مختلفة باستخدام إحصائيات مختلفة، حيث فضلت الدراسة الجمع بينهما لتحسين نوعية الاختبار وجودته.

وأجرت ليو ودونبار وكولن (Lew, Dunbar and Kolen, 2004) دراسة في أيوا بالولايات المتحدة الأمريكية هدفت إلى مقارنة النموذج المعلمي لاختبارات الاختيار من متعدد وطريقة كيرنال اللامعلمية لتقدير منحنى خصائص الفقرة (ICC)، ولتحقيق هدف الدراسة تم استخدام معيار تطبيقي تمثل في الكشف عن مدى استقرار تقديرات المنحنى في حالة تمثيلها بيانياً وبشكل عشوائي. وتم دراسة أثر زيادة معلم التبسيط (h : بارامتر التهذيب (Smoothing Parameter)، وهو يعتمد بشكل أساسي على عدد المفحوصين، ويساوي (1.1N) في برنامج TESTGRAF) على النموذج اللامعلمي، وأثر صغر حجم العينة على النموذجين، وقد أظهرت النتائج وجود اختلافات جزئية بالنسبة للاستقرار ضمن النموذج الواحد (داخل النموذج)، وكان تزايد الاختلافات بسيطاً الذي يعزى للنموذج، وأدى كل من النموذجين إلى نفس النتائج في دقة تقدير منحنى خصائص الفقرة (ICC).

وأجرى فو (Fu, 2010) دراسة بهدف التعرف إلى دقة تقدير معلمة القدرة ومعلمة صعوبة الفقرة باستخدام (5) نماذج من نظرية الاستجابة للفقرة المعلمية واللامعلمية، وكانت النماذج المستخدمة في إطار الدراسة متباينة من حيث مستوى التخمين، وأحجام العينة المستخدمة، وطول الاختبار. وتم في هذه الدراسة توليد مجموعة من الاستجابات بلغت (50) مجموعة من البيانات المولدة باستخدام ظروف اختبار مختلفة. وأشارت نتائج الدراسة إلى أن هناك تبايناً في دقة تقدير معالم الفقرة والمفحوصين حسب مستوى التخمين الموجود في الاختبار، وحجم العينة المستخدمة، وطول الاختبار. كما أشارت النتائج أيضاً إلى أن دقة تقدير معلمة القدرة، ومعالم الفقرة تعتمد على معيار الدقة المستخدم في كل واحد من نماذج استجابة الفقرة.

كما أجرى دي لا توري ويوان (De La Torre & Yuan, 2010) دراسة هدفت إلى التعرف على أثر حجم العينة على دقة تقدير معلمة القدرة ومعالم الفقرة في اختبارات مطورة حسب نماذج نظرية الاستجابة للفقرة (IRT). واستخدمت الدراسة نموذج (HO-IRT) في توليد مجموعة من البيانات ضمن ظروف اختبار مختلفة. حيث تم توليد البيانات باستخدام طريقة مونتي كارلو من أجل التعرف إلى العلاقات بين معلمة القدرة ومعالم الفقرة في الاختبار، وأثر حجم العينة في دقة تقدير معلمة القدرة ومعالم الفقرة. وأشارت النتائج إلى أن حجم العينة يؤثر على قدرة الاختبار في تقدير معلمة القدرة ومعالم الفقرة، وإلى وجود علاقة ارتباطية بين تقدير معلمة القدرة ومعالم الفقرة.

وأجرى الشريفين (2012) دراسة هدفت إلى الكشف عن أثر طريقة تقدير معالم الفقرة وقدرات الأفراد على قيم معالم الفقرة، والخصائص السيكومترية للاختبار، في ضوء تغير حجم العينة. ولتحقيق هدف الدراسة تم بناء اختبار تحصيلي في الفيزياء من نوع الاختيار من متعدد بأربعة بدائل تكون بصورته النهائية من (33) فقرة. وطبق الاختبار على عينة الدراسة المكونة من (1000) طالب وطالبة من طلبة الصف الثاني الثانوي العلمي، وحللت النتائج وفق النموذج الثلاثي المعلمة باستخدام البرمجية (Bilog-Mg). وبينت النتائج وجود فروق ذات دلالة إحصائية (α= 0.05) في متوسطات الأخطاء المعيارية لتقديرات معالم الفقرات تعزى للتفاعل بين طريقة التقدير وحجم العينة، في حين لم تظهر فروق ذات دلالة إحصائية تعزى لمتغير حجم العينة وطريقة التقدير. كما أشارت النتائج إلى وجود فروق ذات دلالة إحصائية (α= 0.05) بين متوسطات الأخطاء المعيارية لتقديرات القدرة للأفراد تعزى لمتغير حجم العينة، وللتفاعل بين طريقة التقدير وحجم العينة، في حين لم تظهر

فروق ذات دلالة إحصائية تعزى لطريقة التقدير، كما بينت النتائج عدم وجود فروق دالة إحصائياً بين معاملات الثبات المقدرة وفق نظرية الاستجابة للفقرة عند أحجام العينة المختلفة (100، و500، و1000). وأشارت النتائج إلى أن دقة تقديرات معلمة القدرة تزداد في حالة عينة الأفراد ذوي القدرة العالية، وعينة الأفراد ذوي القدرة المتدنية عند استخدام طريقة بييز التوقع (EAP)، في حين تزداد الدقة عند مستويات الأفراد ذوي القدرة المتوسطة باستخدام طريقة الأرجحية العظمى (MLE) بغض النظر عن حجم العينة.

التعقيب على الدراسات السابقة:

ومن خلال الدراسات التي تم عرضها سابقاً يمكن استخلاص بعض الاتجاهات في النتائج التي أمكن التوصل إليها فيما يتعلق بدقة تقدير المعالم في بعض نماذج نظرية الاستجابة للفقرة وذلك على النحو الآتي:

1. دقة تقدير المعالم في نماذج نظرية الاستجابة للفقرة المعلمية واللامعلمية تتأثر بطول الاختبار وحجم عينة المفحوصين.
2. إن دقة تقدير المعالم تتغير بشكل واضح تبعاً للنموذج المستخدم في تقدير المعالم، والطريقة المستخدمة في التقدير.
3. قليل من الدراسات قامت – في حدود علم الباحث– بفحص أثر تشكيلات مختلفة لحجم عينة المفحوصين، وطول الاختبار على دقة تقدير معالم الفقرة والقدرة باستخدام نموذج لامعلمي، وهذا يظهر أهمية إجراء الدراسة الحالية.
4. إن الدراسات السابقة لم تستخدم مؤشرات الدقة، مثل: التحيز (BIAS)، والجذر التربيعي لمتوسط مربعات الخطأ (RMSE) للإجابة عن أسئلة الدراسة.
5. تتفق الدراسة الحالية مع الدراسات السابقة في سعيها لتقدير دقة تقدير معالم الفقرة والقدرة باستخدام نظرية الاستجابة للفقرة، في حين تختلف الدراسة الحالية مع الدراسات السابقة في أنها تختبر دقة التقدير باختلاف حجم العينة، وطول الاختبار في النماذج اللامعلمية.

أهداف الدراسة:

سعت الدراسة إلى تحقيق الأهداف الأتية:

1- الكشف عن أثر حجم عينة المفحوصين على دقة تقدير معالم الفقرة والقدرة باستخدام نموذج نظرية الاستجابة للفقرة اللامعلمية.

2- الكشف عن اثر طول الاختبار على دقة تقدير معالم الفقرة والقدرة باستخدام نموذج نظرية الاستجابة للفقرة اللامعلمية.

3- الكشف عن اثر تفاعل حجم عينة وطول الاختبار المفحوصين على دقة تقدير معالم الفقرة والقدرة باستخدام نموذج نظرية الاستجابة للفقرة اللامعلمية.

وسيتم تكوين تشكيلات مختارة من حجوم عينات المفحوصين وأطوال اختبارات بهدف التعرف على اثر هذه التشكيلات على دقة تقدير معالم الفقرة والقدرة في النموذج اللامعلمي لنظرية الاستجابة للفقرة .

مشكلة الدراسة واهميتها:

بالرغم من الميزات التي تمتاز بها نماذج نظرية الاستجابة للفقرة، إلا أن المشكلة في تطبيق نظرية الاستجابة للفقرة، أنها تحتاج إلى حجوم عينات وأطوال اختبار كبيرة، إلا أن نماذج نظرية الاستجابة للفقرة اللامعلمية لا تحتاج إلى حجوم عينات واطوال اختبار كبيرة مقارنة بنظيرتها المعلمية، كما ادعى رامسي (Ramsay, 1991) أن عددا قليلا من المفحوصين لا يزيد عن 100 وعدد فقرات الاختبار 20 هما المطلوبان لتقدير (ICCs) . فقد شاع استخدام نظرية استجابة الفقرة المعلمية من قبل الباحثين على الرغم من موانعها في تحليل البيانات ذات المستوى الرتبي، إلا أن مصداقية النتائج قد تكون موضع تساؤل عندما لا يتحقق فرض وقوع البيانات على مستوى فترة، الأمر الذي تبرره نماذج استجابة الفقرة اللامعلمية والتي لا تضع قيود حول شكل دالة استجابة الفقرة، مما يثير التساؤل حول مدى مطابقة النوعين للبيانات التحصيلية التي يعتبرها البعض رتبيه والبعض الآخر شبه فئوية، ومدى دقة النتائج التي تفرزها مثل هذه الاختبارات (liang ,2010).

اهتمت معظم الدراسات في الكشف عن أثر حجم عينة المفحوصين، وطول الاختبار على دقة تقدير معالم الفقرة والقدرة للنماذج المعلمية، إلا أنه لم يتم فحص أثر تشكيلات مختلفة لطول الاختبار وحجم عينة المفحوصين على دقة تقدير معالم الفقرة والقدرة، لذا تنبثق أهمية هذه الدراسة من الحاجة لمعرفة دقة التقديرات المختلفة (معالم الفقرة ومعلمة القدرة) لتقديرات نظرية الاستجابة للفقرة اللامعلمية باختلاف حجم عينة المفحوصين، وطول الاختبار، الأمر الذي يسمح بتحليل البيانات باستخدام النماذج اللامعلمية حين الإيفاء بافتراضاتها، وتوافر حجم العينة، وطول الاختبار المناسبين لتطبيق هذه النماذج (إضافة إلى الميزات الأخرى التي تمتاز بها النماذج اللامعلمية، مثل سهولة المطابقة، وشكل دالة الفقرة، وأنها تحتاج إلى حجم عينة وطول اختبار أقل).

ويمكن تحديد مشكلة الدراسة من خلال الأسئلة الرئيسة الآتية:

1. هل توجد فروق دالة إحصائياً عند مستوى الدلالة α=0.05 بين متوسطات مؤشرات دقة تقدير معالم الفقرة (a, b, c) المقدرة تعزى لمتغير حجم العينة ومتغير طول الاختبار والتفاعل بينهما ؟
2. هل توجد فروق دالة إحصائياً عند مستوى الدلالة α=0.05 بين متوسطات مؤشرات دقة تقدير معلمة القدرة θ المقدرة تعزى لمتغير حجم العينة ومتغير طول الاختبار والتفاعل بينهما؟

تعريف المصطلحات:

- نماذج نظرية الاستجابة للفقرة اللامعلمية (NIRT): هي النماذج التي لا تحدد شكل دالة استجابة الفقرة، والتي تتطلب افتراضات أقل تشدداً، وتصلح للبيانات الفئوية والرتيبة، التي تستخدم برنامج (TESTGRAF) الذي يستخدم طريقة (KERNEL SMOOTHING (KS)) في التقدير.
- حجم العينة: تعني عدد المفحوصين الذين سيطبق عليهم الاختبار، وفي هذه الدراسة سيتم استخدام أربع عينات بأحجام مختلفة.
- طول الاختبار: يعني عدد الفقرات التي يتكون منها الاختبار، وفي هذه الدراسة سيتم استخدام ثلاثة اختبارات بأطوال مختلفة.

- معلمة القدرة Ability: اللوغاريتم الطبيعي لدالة الأرجحية القصوى Maximum Likelihood Function لإجابة المفحوص اجابة صحيحة على الفقرات التي تعتبر نقطة صفر التدريج عند صعوبتها.
- معلمة الصعوبة Threshold: مستوى القدرة الذي يقابل احتمال 0.50، للإجابة على الفقرة إجابة صحيحة عندما يكون معامل التخمين مساويا للصفر.
- معلمة التمييز Slope: ميل منحنى خصائص الفقرة، الذي يقابل النقطة التي تكون فيها علامة القدرة تساوي صعوبة الفقرة.
- معلمة التخمين Asymptote: هي خط المقاربة الأدنى Lower Asymptote من منحنى خصائص الفقرة، ويمثل احتمالية إجابة المفحوصين ذوي القدرة المتدنية على الفقرة إجابة صحيحة.
- دقة التقدير (Accuracy of Estimation): تعبير يشير إلى جودة التقدير التي يميزها الاحتمالية الكبيرة في أن التقدير قريب من القيمة الحقيقية، حيث يمكن الوصول إلى ذلك باختيار التقدير غير المتحيز (unbiased estimator)، الذي يتصف تباينه بأنه أقل تباين من أي تقدير آخر غير متحيز، وذلك باستخدام الجذر التربيعي لمعدل مربعات الأخطاء (RMSE) (Baker, 2001).

محددات الدراسة:

1- أجريت هذه الدراسة باستخدام طريقة المحاكاة (Simulation)، بيانات مولدة.

2- فيما يتعلق بتوزيعات معالم الفقرة والقدرة المولدة في الدراسة الحالية، فقد تم تبني التوزيع الطبيعي (0،1) لمعلمة القدرة ومعلمة الصعوبة، والتوزيع المنتظم لمعلمتي التمييز والتخمين.

3- تم إجراء هذه الدراسة بتكرار واحد لعملية تقدير معالم الفقرة والقدرة.

منهجية الدراسة وإجراءاتها:

توليد البيانات:

تم استخدام المنهج التجريبي في هذه الدراسة بتوليد بيانات باستخدام المحاكاة (لأن هذا الأسلوب يمكننا من ضبط المتغيرات الخارجية، ولأن هدف الدراسة فحص منهجية عمل نماذج نظرية

الاستجابة للفقرة اللامعلمية) من الاستجابات الثنائية (1، 0) لعينات تحاكي عينات المجتمع الأصلي بطريقة المونتي كارلو (Monte Carlo) بحجوم عينات وأطوال اختبار مختلفة، وذلك باستخدام برنامج (WinGen v.3)، بحيث يتراوح أحجام العينات من (100 إلى 1000) مفحوص تحت افتراض التوزيع الطبيعي للقدرة (1، 0)، واختبارات يتراوح طولها من (20−60) فقرة تحت افتراض التوزيع الطبيعي لصعوبة الفقرات (1، 0)، والتوزيع المنتظم لمعلمة التمييز بواقع القيمة الابتدائية (0.4) والقيمة النهائية (1.2)، والتوزيع المنتظم لمعلمة التخمين بواقع القيمة الابتدائية (0.2) والقيمة النهائية (0.3) على افتراض أن الاختبار المُولدة بياناته هو اختبار الاختيار من متعدد وله أربعة بدائل، تلائم النموذج ثلاثي المعلم، والجدول(1) يبين المتوسطات الحسابية والانحرافات المعيارية الخاصة بمعالم الفقرة الحقيقية لكلٍّ من معلمة التمييز والصعوبة والتخمين لكل حالة من الحالات المشمولة بالدراسة الناتجة عن تفاعل متغيري الدراسة (حجم العينة، وطول الاختبار).

جدول (1) المتوسطات الحسابية والانحرافات المعيارية لمعالم الفقرة الحقيقية وفقاً لمتغيري الدراسة (طول الاختبار، حجم العينة)

	الإحصائي	عدد الفقرات								
		20 فقرة			40 فقرة			60 فقرة		
		A	b	C	a	b	C	a	b	C
100	المتوسط الحسابي	0.89	-0.02	0.24	0.76	0.25	0.25	0.83	-0.03	0.25
	الانحراف المعياري	0.24	0.92	0.03	0.23	1.01	0.03	0.23	0.88	0.03
250	المتوسط الحسابي	0.89	-0.02	0.24	0.76	0.25	0.25	0.83	-0.03	0.25
	الانحراف المعياري	0.24	0.92	0.03	0.23	1.01	0.03	0.23	0.88	0.03
500	المتوسط الحسابي	0.89	-0.02	0.24	0.76	0.25	0.25	0.83	-0.03	0.25
	الانحراف المعياري	0.24	0.92	0.03	0.23	1.01	0.03	0.23	0.88	0.03
1000	المتوسط الحسابي	0.89	-0.02	0.24	0.76	0.25	0.25	0.83	-0.03	0.25
	الانحراف المعياري	0.24	0.92	0.03	0.23	1.01	0.03	0.23	0.88	0.03

والجدول (2) يبين معلم القدرة الحقيقية لكل حالة من الحالات المشمولة بالدراسة الناتجة عن تفاعل متغيري الدراسة (حجم العينة، وطول الاختبار).

جدول (2) المتوسطات الحسابية والانحرافات المعيارية لمعلم القدرة الحقيقية وفقاً لمتغيري الدراسة (حجم العينة، وطول الاختبار)

حجم العينة (فرد)	الإحصائي	طول الاختبار 20 فقرة	40 فقرة	60 فقرة
100 (فرد)	القيمة الصغرى	-1.75	-1.75	-1.75
	المتوسط الحسابي	-0.02	-0.03	-0.03
	الانحراف المعياري	0.80	0.80	0.80
	القيمة العظمى	1.43	1.43	1.43
250 (فرد)	القيمة الصغرى	-2.11	-2.07	-2.11
	المتوسط الحسابي	-0.01	-0.01	0.00
	الانحراف المعياري	0.92	0.91	0.91
	القيمة العظمى	2.78	2.78	2.78
500 (فرد)	القيمة الصغرى	-2.86	-2.80	-2.73
	المتوسط الحسابي	0.04	0.02	0.02
	الانحراف المعياري	0.95	0.95	0.95
	القيمة العظمى	2.37	2.37	2.37
1000 (فرد)	القيمة الصغرى	-3.25	-3.25	-2.39
	المتوسط الحسابي	0.04	0.05	0.05
	الانحراف المعياري	1.00	0.98	0.97
	القيمة العظمى	3.71	3.71	3.06

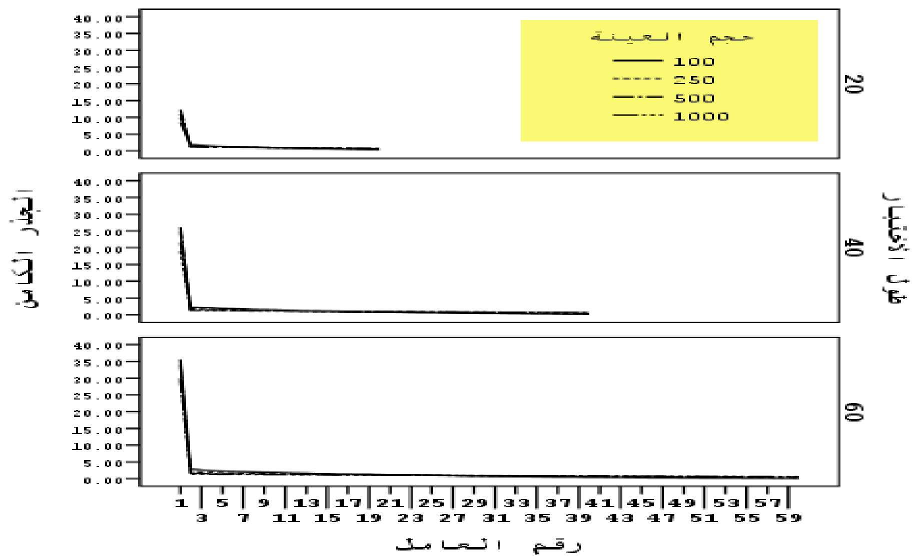
والجدول (3) يبين نتائج التحليل العاملي للبيانات المُوَّلدة عندما يكون طول الاختبار (20، و40، و60) وحجم العينة (100، و250، و500، و1000)، الذي تم استخدامه بهدف التحقق من افتراض أحادية البعد.

جدول(3) نتائج التحليل العاملي للبيانات المُوَّلدة وفقاً لمتغيري الدراسة (حجم العينة، وطول الاختبار)

حجم العينة	طول الاختبار	رقم العامل	الجذر الكامن	التباين المفسر %	التباين المفسر التراكمي	الجذر الكامن الأول / الجذر الكامن الثاني	(الجذر الكامن الأول - الجذر الكامن الثاني) / (الجذر الكامن الثاني - الجذر الكامن الثالث)
100	20	1	8.512	42.6	42.6	4.61	26.30
		2	1.848	9.2	51.8		
		3	1.595	8.0	59.8		
	40	1	24.746	61.9	61.9	11.23	259.75
		2	2.203	5.5	67.4		
		3	2.116	5.3	72.7		
	60	1	35.554	59.3	59.3	12.54	165.83
		2	2.836	4.7	64.0		
		3	2.639	4.4	68.4		
250	20	1	9.843	49.2	49.2	6.93	110.52
		2	1.420	7.1	56.3		
		3	1.344	6.7	63.0		
	40	1	19.015	47.5	47.5	10.78	214.88
		2	1.765	4.4	52.0		
		3	1.684	4.2	56.2		
	60	1	29.942	49.9	49.9	14.68	438.74
		2	2.040	3.4	53.3		
		3	1.976	3.3	56.6		
500	20	1	12.311	61.6	61.6	9.64	233.20
		2	1.277	6.4	67.9		
		3	1.230	6.1	74.1		
	40	1	26.111	65.3	65.3	17.74	621.76
		2	1.472	3.7	69.0		
		3	1.432	3.6	72.5		
	60	1	29.154	48.6	48.6	17.55	415.90
		2	1.661	2.8	51.4		
		3	1.595	2.7	54.0		

حجم العينة	طول الاختبار	رقم العامل	الجذر الكامن	التباين المفسر %	التباين المفسر التراكمي	$\frac{\text{الجذر الكامن الأول}}{\text{الجذر الكامن الثاني}}$	$\frac{(\text{الجذر الكامن الأول} - \text{الجذر الكامن الثاني})}{(\text{الجذر الكامن الثاني} - \text{الجذر الكامن الثالث})}$
1000	20	1	10.985	54.9	54.9	9.49	190.19
		2	1.158	5.8	60.7		
		3	1.106	5.5	66.2		
	40	1	21.545	53.9	53.9	16.58	818.90
		2	1.299	3.2	57.1		
		3	1.274	3.2	60.3		
	60	1	33.951	56.6	56.6	24.73	1313.86
		2	1.373	2.3	58.9		
		3	1.348	2.2	61.1		

يلاحظ من الجدول 3، أن كافة قيم التباين المفسر عندما يكون طول الاختبار (20، 40، 60) وحجم العينة (100، 250، 500، 1000) قد تخطت الـ 20% كمؤشر أوَّل على أحادية البعد، كما وتخطت جميع قيم عملية قسمة الجذر الكامن الأول على الجذر الكامن الثاني قيمة 2 كمؤشر ثانٍ على تحقق أحادية البعد، وكذلك أظهرت نتائج عملية قسمة حاصل طرح الجذر الكامن الثاني من الجذر الكامن الأول على حاصل طرح الجذر الكامن الثالث من الجذر الكامن الثاني قيماً ضخمة مما يشير إلى تحقق افتراض أحادية البعد كمؤشر ثالث، والشكل 1، يبين نتائج التحليل العاملي للبيانات المُوَّلدة وفقاً لمتغيري الدراسة (طول الاختبار، وحجم العينة) باستخدام الجذور الكامنة وعدد العوامل كمؤشر رابع على تحقق أحادية البعد.



شكل 1: رسم بياني يوضح نتائج التحليل العاملي للبيانات المولدة وفقاً لمتغيري الدراسة

(حجم العينة، وطول الاختار)

ومن ثم تم تقدير معالم الفقرات وقدرات الافراد باستخدام النموذج الثلاثي المعلمة باستخدام برنامج (Testgraf version of July 2001)، حيث إن هذا البرنامج يستخدم أسلوب تنعيم النواة (Kernel Smoothing) (اللامعلمي) في تقدير معالم الفقرة والقدرة للإجابة عن أسئلة الدراسة.

المعالجات الإحصائية:

تم استخدام المعالجات الإحصائية الآتية:

للإجابة عن سؤال الدراسة الأول؛ تم حساب المتوسطات الحسابية والانحرافات المعيارية لمؤشر دقة القياس ممثلةً بـ (BIAS) الخاصة بمعالم فقرات الاختبار (a, b, c) المقدرة، حيث تم حساب مؤشر الدقة BIAS باستخدام المعادلة $Bias(a,b,c^*) = \sum_{i=1}^{N}(a,b,c_i^* - a,b,c_i)/N$، متبوعة بإجراء تحليل التباين الثنائي 2-Way Interaction ANOVA للكشف عن مواطن الدقة في كل حالات الدراسة، كما تم حساب المتوسطات الحسابية لمؤشر دقة القياس ممثلةً بـ (RMSE)

الخاصة بمعالم فقرات الاختبار (a, b, c) المقدرة، حيث تم حساب مؤشر الدقة RMSE (Root Mean Square Error) باستخدام المعادلة $RMSE(a,b,c^*) = \sqrt{\sum_{i=1}^{N}(a,b,c_i^* - a,b,c_i)^2/N}$، حيث $a,b\,,c_i^*$ هي معالم الفقرات المقدرة ، (a, b, c) هي معالم الفقرات الحقيقية للفقرة i، N عدد الفقرات.

وللإجابة عن سؤال الدراسة الثاني؛ تم حساب المتوسطات الحسابية والانحرافات المعيارية لمؤشر دقة القياس ممثلةً بـ (BIAS) الخاصة بمعلم القدرة المقدرة، حيث تم حساب مؤشر الدقة BIAS باستخدام المعادلة $Bias(\theta^*) = \sum_{i=1}^{N}(\theta_i^* - \theta_i)/N$، متبوعة بإجراء تحليل التباين الثنائي 2-Way Interaction ANOVA للكشف عن مواطن الدقة في كل حالات الدراسة، كما تم حساب المتوسطات الحسابية لمؤشر دقة القياس ممثلةً بـ (RMSE) الخاصة بمعلم القدرة المقدرة، وتم حساب مؤشر الدقة RMSE (Root Mean Square Error) باستخدام المعادلة $RMSE(\theta^*) = \sqrt{\sum_{i=1}^{N}(\theta_i^* - \theta_i)^2/N}$، حيث θ_i^* القدرة المقدرة للمفحوص i، θ_i القدرة الحقيقية للمفحوص i، N عدد المفحوصين (Weiss, 2009).

نتائج الدراسة ومناقشتها

أولاً: النتائج المتعلقة بالإجابة عن السؤال الأول: "هل توجد فروق دالة إحصائياً عند مستوى الدلالة α=0.05 بين متوسطات مؤشرات دقة تقدير معالم الفقرة (a, b, c) المقدرة تعزى لمتغير حجم العينة ومتغير طول الاختبار والتفاعل بينهما ؟"

وللإجابة عن هذا السؤال فقد تم تقدير قيم معالم الصعوبة والتمييز والتخمين لكل تشكيلات الاختبار باختلاف حجم العينة وطول الاختبار، ثم تم حساب المتوسطات الحسابية والانحرافات المعيارية لتقديرات معالم الفقرة باختلاف متغيري (حجم العينة، وطول الاختبار)، وذلك كما هو مبين في الجدول 4.

جدول (4) المتوسطات الحسابية والانحرافات المعيارية لتقديرات معالم الفقرة باختلاف باختلاف متغيري (حجم العينة، وطول الاختبار)

حجم العينة	الإحصائي	طول الاختبار								
		20 فقرة			40 فقرة			60 فقرة		
		a	b	C	a	b	C	a	b	C
100 فرد	المتوسط الحسابي	0.414	2.705	0.452	0.782	-1199757.295	0.383	0.776	2.905	0.453
	الانحراف المعياري	0.73	5.70	0.23	0.90	7587948.93	0.29	0.95	9.25	0.27
250 فرد	المتوسط الحسابي	0.732	1.852	0.441	0.481	28.497	0.320	0.526	1.317	0.407
	الانحراف المعياري	0.92	2.85	0.25	0.69	153.65	0.18	0.50	2.01	0.24
500 فرد	المتوسط الحسابي	0.220	-758199.424	0.219	0.329	0.910	0.299	0.386	1.473	0.273
	الانحراف المعياري	0.14	3390773.62	0.16	0.27	2.04	0.21	0.45	5.57	0.23
1000 فرد	المتوسط الحسابي	0.507	0.896	0.444	0.427	0.574	0.306	0.485	0.475	0.342
	الانحراف المعياري	0.25	1.05	0.23	0.25	1.35	0.19	0.26	1.58	0.18

يلاحظ من الجدول 4 وجود فروق ظاهرية بين المتوسطات الحسابية الخاصة بتقديرات معالم الفقرات (a, b, c) ناتجة عن اختلاف مستويات المتغيرين، حيث يلاحظ أن أكبر قيمة لمتوسط تقديرات معلمة التمييز a قد كانت 0.782 بانحراف معياري 0.90، عندما كان حجم العينة 100 وطول اختبار 40 ، أما متوسط تقديرات معلمة الصعوبة b فقد كانت أكبر قيمة لها كبيرة جدا وذلك بسبب عدم قدرة هذه الطريقة (KS) تقدير صعوبة الفقرة التي أجاب عنها أو لم يجيب عنها كل

المفحوصين، أما متوسط تقديرات معلمة التخمين c فقد كانت أكبر قيمة لها 0.453 بانحراف معياري 0.27 عندما كان حجم العينة 100 فرد وطول الاختبار 60 فقرة .

وكذلك تم حساب المتوسطات الحسابية والانحرافات المعيارية الخاصة بمؤشر التحيز BIAS كمؤشر من مؤشرات دقة التقدير لمعالم الفقرات باختلاف متغيري (حجم العينة، وطول الاختبار)، وذلك كما هو مبين في الجدول 5.

جدول(5) المتوسطات الحسابية والانحرافات المعيارية لمؤشر التحيز لتقديرات معالم الفقرات باختلاف متغيري (حجم العينة وطول الاختبار)

طول الاختبار											الإحصائي	حجم العينة
60 فقرة				40 فقرة				20 فقرة				
مؤشر التحيز لمعالم الفقرات:												
C	b	a		C	b	a		c	b	a		
0.20	2.94	-0.05		0.13	-1199757.55	0.02		0.21	2.72	-0.47	المتوسط الحسابي	100 فرد
0.27	9.20	1.02		0.29	7587949.12	0.92		0.23	5.55	0.69	الانحراف المعياري	
0.16	1.35	-0.30		0.07	28.24	-0.28		0.20	1.87	-0.15	المتوسط الحسابي	250 فرد
0.24	1.94	0.54		0.18	153.48	0.71		0.24	3.01	1.03	الانحراف المعياري	
0.02	1.51	-0.44		0.05	0.66	-0.43		-0.02	-758199.40	-0.67	المتوسط الحسابي	500 فرد
0.22	5.63	0.51		0.21	1.60	0.35		0.16	3390773.74	0.189	الانحراف المعياري	
0.09	0.51	-0.34		0.05	0.32	-0.33		0.20	0.92	-0.38	المتوسط الحسابي	1000 فرد
0.18	1.33	0.31		0.18	1.14	0.35		0.23	1.14	0.32	الانحراف المعياري	

يلاحظ من الجدول(5) وجود فروق ظاهرية بين القيم المطلقة للمتوسطات الحسابية الخاصة بمؤشر التحيز لتقديرات معالم الفقرات (a, b, c) ناتجة عن اختلاف مستويات المتغيرين، حيث يلاحظ أن أقل قيمة مطلقة لمتوسط مؤشر التحيز لمعلمة التمييز a قد كانت 0.02 بانحراف معياري 0.92، عندما كان حجم العينة 100 وطول الاختبار 40، اما معلمة الصعوبة b فقد كانت أقل قيمة مطلقة لمتوسط مؤشر التحيز 0.32 بانحراف معياري 1.14 عندما كان حجم العينة 1000 وطول الاختبار 40، أما معلمة التخمين c فقد كانت أقل قيمة مطلقة لمتوسط مؤشر التحيز 0.02 بانحراف معياري 0.16 عندما كان حجم العينة 500 وطول الاختبار 20 فقرة.

وفي ضوء ما تقدم، تم إجراء تحليل التباين الثنائي 2-Way Interaction ANOVA لقيم مؤشر التحيز BIAS لتقديرات معلمة الفقرات (a) المقدرة باختلاف متغيري (حجم العينة، طول الاختبار)، وذلك كما هو مبين في الجدول 6.

جدول (6) نتائج تحليل التباين الثنائي لقيم مؤشر التحيز لتقديرات معلمة الفقرات (a) المقدرة باختلاف متغيري (حجم العينة وطول الاختبار)

مصدر التباين	مجموع المربعات	درجة الحرية	متوسط مجموع المربعات	قيمة ف المحسوبة	الدلالة الإحصائية
طول الاختبار	0.771	2	0.386	1.531	0.217
حجم العينة	2.378	3	0.793	3.147	0.025
طول الاختبار× حجم العينة	1.715	6	0.286	1.135	0.341
الخطأ	117.379	466	0.252		
الكلي	122.244	477			

يتضح من الجدول 6 عدم وجود فروق دالة إحصائياً عند مستوى الدلالة α=0.05 بين المتوسطات الحسابية لقيم مؤشر التحيز BIAS لتقديرات معلمة الفقرات (a) المقدرة تعزى لمتغير طول الاختبار، والتفاعل بين متغيري طول الاختبار وحجم العينة؛ مما يعني عدم تأثر دقة تقدير معلمة التمييز بطول الاختبار والتفاعل بين متغيري حجم العينة وطول الاختبار، كما يتبين من الجدول 6 وجود فروق دالة إحصائياً عند مستوى الدلالة α=0.05 بين المتوسطات الحسابية لقيم مؤشر التحيز BIAS لتقديرات معلمة الفقرات (a) المقدرة تعزى لمتغير حجم العينة؛ أي أن هناك تأثيرا لمتغير حجم العينة في دقة تقدير معلمة التمييز، ولكون المتغير متعدد المستويات، فقد تم إجراء اختبار levene للكشف عن انتهاك تجانس التباين من عدمه بين الأوساط الحسابية لقيم مؤشر التحيز لتقديرات معلمة تمييز الفقرة المقدرة وفقا للمتغير، حيث بلغت قيمة F المحسوبة لاختبار levene ما مقداره 1.012 عند درجتي حرية (11 للبسط، 466 للمقام) وهي غير دالة إحصائياً عند مستوى الدلالة (α = 0.05) مما يعني إمكانية استخدام اختبار Schefee للمقارنات البعدية، لتحديد لصالح أي من حجوم العينات قد كانت الفروق بينها دالة إحصائيا، حيث تبين وجود فروق ذات دلالة إحصائية بين حجم العينة 100 من جهة وحجوم العينات (250،500،1000) على التوالي، ولصالح حجم العينة 100، كما تبين وجود فروق ذات دلالة إحصائية بين حجم العينة 250 من جهة وحجوم العينات (500، 1000) على التوالي، ولصالح حجم العينة 250، وقد يعزى ذلك إلى النموذج الرياضي المستخدم في طريقة التقدير (KS)، حيث أن هذه الطريقة تستخدم الرتب في تقدير معالم الفقرة، وأنها تعمل ترتيب عشوائي للرتب المتشابهة (لا يوجد أفراد أو فقرات لها نفس الرتبة)، وهذا يؤدي خسارة في المعلومات، لذلك عندما يكون حجم العينة قليل نسبيا فان دقة التقدير تكون أكبر لأن احتمالية وجود رتب متساوية يكون أقل، ولم يتبين وجود فروق ذات دلالة إحصائية بين حجم العينة 500 وحجم العينة 1000، وهذه النتيجة تتوافق مع ما ادعاه رامسي (Ramsay,1991) أنه يمكن تقدير معالم الفقرات باستخدام عدد قليل من الفقرات والمفحوصين.

كما تم إجراء تحليل التباين الثنائي 2-Way Interaction ANOVA لقيم مؤشر التحيز BIAS لتقديرات معلمة الفقرات (b) المقدرة باختلاف متغيري (حجم العينة، وطول الاختبار)، وذلك كما هو مبين في الجدول 7.

جدول (6) نتائج تحليل التباين الثنائي لقيم مؤشر التحيز لتقديرات معلمة الفقرات (b) المقدرة باختلاف متغيري (طول الاختبار، وحجم العينة)

مصدر التباين	مجموع المربعات	درجة الحرية	متوسط مجموع المربعات	قيمة ف المحسوبة	الدلالة الإحصائية
طول الاختبار	4461624783675.400	2	2230812391837.700	0.844	0.431
حجم العينة	5742335895015.290	3	1914111965005.100	0.724	0.538
طول الاختبار×حجم العينة	19429912163911.200	6	3238318693985.200	1.225	0.292
الخطأ	1231975654788660.000	466	2643724581091.550		
الكلي	1261609527631270.000	477			

يتضح من الجدول(7) عدم وجود فروق دالة إحصائية عند مستوى الدلالة α=0.05 بين المتوسطات الحسابية لقيم مؤشر التحيز BIAS لتقديرات معلمة الفقرات (b) المقدرة تعزى لمتغير طول الاختبار، أو متغير حجم العينة، أو التفاعل بينهما (طول الاختبار، حجم العينة)؛ مما يعني عدم تأثر دقة تقدير معلمة الصعوبة بمتغيري حجم العينة وطول الاختبار والتفاعل بينهما، وأن الفروق في المتوسطات الحسابية ناتجة عن الصدفة، وقد يعزى كبر قيم المتوسطات الحسابية إلى أن طريقة التقدير المستخدمة في النموذج الثلاثي اللامعلمي (KS) لا تستطيع تقدير معلمة صعوبة الفقرة التي أجاب عليها أو لم يجيب عليها جميع الأفراد، لذلك قدرت صعوبة بعض الفقرات كبيرة جدا (الفقرات التي لم يجيب عليها كل الأفراد) ومرة أخرى قدرت صعوبة بعض الفقرات قليلة جدا (الفقرات التي أجاب عنها كل الأفراد)، بعكس طريقة التقدير المستخدمة في النماذج المعلمية (طريقة الأرجحية العظمى الهامشية) والتي تستطيع تقدير معالم الفقرة التي أجاب أو لم يجيب عليها جميع الأفراد.

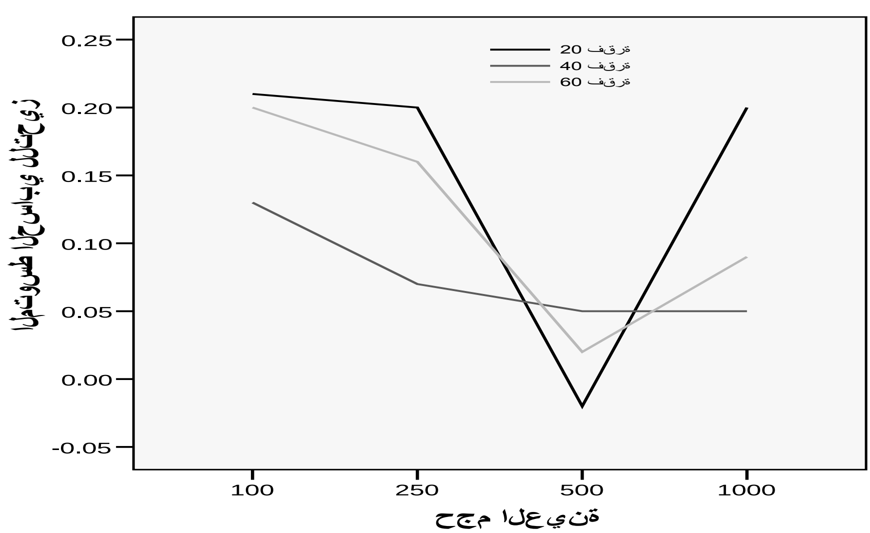
كما تم إجراء تحليل التباين الثنائي 2-Way Interaction ANOVA لقيم مؤشر التحيز BIAS لتقديرات معلمة الفقرات (c) المقدرة باختلاف متغيري (حجم العينة، وطول الاختبار)، وذلك كما هو مبين في الجدول 8.

جدول (8) نتائج تحليل التباين الثنائي لقيم مؤشر التحيز لتقديرات معلمة الفقرات (c) المقدرة باختلاف متغيري (حجم العينة، وطول الاختبار)

مصدر التباين	مجموع المربعات	درجة الحرية	متوسط مجموع المربعات	قيمة ف المحسوبة	الدلالة الإحصائية
طول الاختبار	0.311	2	0.155	5.055	0.007
حجم العينة	0.543	3	0.181	5.889	0.001
طول الاختبار× حجم العينة	0.754	6	0.126	4.090	0.001
الخطأ	14.315	466	0.031		
الكلي	15.922	477			

يتضح من الجدول (8) وجود فروق دالة إحصائياً عند مستوى الدلالة α=0.05 بين المتوسطات الحسابية لقيم مؤشر التحيز BIAS لتقديرات معلمة الفقرات (c) المقدرة تعزى لمتغير طول الاختبار ومتغير حجم العينة والتفاعل بينهما؛ مما يعني تأثر دقة تقدير معلمة التخمين بمتغيري حجم العينة وطول الاختبار والتفاعل بينهما، ولكون المتغير متعدد المستويات، فقد تم إجراء اختبار levene للكشف عن انتهاك تجانس التباين من عدمه بين الأوساط الحسابية لقيم مؤشر التحيز لتقديرات معلمة تمييز الفقرة المقدرة وفقا للمتغير، حيث بلغت قيمة F المحسوبة لاختبار levene ما مقداره 2.012 عند درجتي حرية (11 للبسط، 466 للمقام) وهي غير دالة إحصائياً عند مستوى الدلالة (α = 0.05) مما يعني امكانية استخدام اختبار Schefee للمقارنات البعدية، لتحديد لصالح أي من حجوم العينات قد كانت الفروق بينها دالة إحصائيا، حيث تبين عدم وجود فرق ذي دلالة إحصائية بين حجم العينة 100 وحجم العينة 250، كما تبين وجود فرق ذي دلالة إحصائية بين حجم العينة 100 من جهة وحجوم العينات (1000،500) من جهة أخرى ولصالح حجم العينة 500، و1000، كما تبين وجود فرق ذي دلالة إحصائية بين حجم العينة 250 من جهة وحجوم العينات (500، 1000) من جهة أخرى ولصالح حجم العينة 500 ، 1000على التوالي، كما تبين وجود فرق ذي دلالة إحصائية بين حجم العينة 500 وحجم العينة 1000 ولصالح حجم العينة 500 ،اما متغير

طول الاختبار، (بعد التأكد من تجانس التباين لإمكانية استخدام اختبار Scheffe للمقارنات البعدية) فقد تبين وجود فرق ذي دلالة إحصائية بين طول الاختبار 20 فقرة، و40 فقرة ولصالح طول الاختبار 40 فقرة، وتبين وجود فرق ذي دلالة إحصائية بين طول الاختبار 40 فقرة، و60 فقرة ولصالح طول الاختبار 40 فقرة، ولم يتبين وجود فرق ذي دلالة إحصائية بين طول الاختبار 20 فقرة، و60 فقرة، مما يعني أن هذا النموذج مناسب ويكون أكثر دقة عندما يكون التخمين أقل ما يمكن أو في النماذج التي تفترض التخمين صفر، ولتوضيح أثر التفاعل بين متغيري حجم العينة وطول الاختبار في دقة تقدير معلمة التخمين (c)؛ أي أن هناك تأثيرا مشتركا لكل من العاملين، فقد تم تمثيل هذا التفاعل بيانيا كما في الشكل رقم(2) وذلك بالاستعانة بالأوساط الحسابية في الجدول رقم (5).



شكل (2) التفاعل لمغيري الدراسة (حجم العينة، وطول الاختبار) لقيم مؤشر التحيز BIAS لتقديرات معلمة الفقرات (a) المقدرة

من الشكل رقم (2) أن متوسط مؤشر الدقة BIAS قد كانت اقل ما يمكن عند حجم العينة 500، وطول الاختبار 20 ، يليها متوسط مؤشر الدقة BIAS عند حجم العينة 500، وطول الاختبار40، وكان متوسط مؤشر الدقة PIAS اكبر شيء عند حجم العينة 500، وطول الاختبار

60 فقرة، مما يعني أن أفضل دقة تقدير لمعلمة التخمين c كانت عند حجم العينة 500 فرد وطول الاختبار 20 فقرة.

كذلك في ضوء ما تقدم، فقد تم حساب قيم مؤشر RMSE كمؤشر من مؤشرات دقة التقدير لمعالم الفقرات باختلاف متغيري (حجم العينة، وطول الاختبار)، وذلك كما هو مبين في الجدول(9).

جدول (9) قيم مؤشر RMSE لتقديرات معالم الفقرات باختلاف متغيري (حجم العينة، وطول الاختبار)

طول الاختبار	حجم العينة	RMSE لمعلمة التمييز			RMSE لمعلمة الصعوبة			RMSE لمعلمة التخمين	
20 فقرة	100 فرد		0.825			6.057			0.308
	250 فرد		1.016			3.484			0.309
	500 فرد		0.692			3390773.597			0.161
	1000 فرد		0.489			1.441			0.303
40 فقرة	100 فرد		0.912			7587948.714			0.313
	250 فرد		0.752			154.160			0.186
	500 فرد		0.549			1.714			0.208
	1000 فرد		0.478			1.172			0.188
60 فقرة	100 فرد		1.008			9.587			0.337
	250 فرد		0.619			2.347			0.285
	500 فرد		0.670			5.783			0.223
	1000 فرد		0.461			1.412			0.204

يلاحظ من الجدول 9، أن كافة قيم مؤشر دقة التقدير RMSE لمعالم فقرات الاختبار ذي البيانات الموَّلدة (a, b, c) عندما كان حجم العينة 100فرد وطول الاختبار 20 فقرة، قد كانت أصغر ظاهرياً من قيم مؤشر دقة التقدير RMSE لمعالم فقرات الاختبار ذي البيانات الموَّلدة (a, b, c) عندما كان حجم العينة(250، 500، 1000) فقرة وطول الاختبار (40، 60) فقرة في معظم حالات الدراسة تقريبا. مما يعني أن دقة تقدير معالم الفقرة (a, b, c) عند حجم عينة 100فرد وطول اختبار 20 فقرة قد كان أكثر دقة ظاهرياً بما كان عليه الأمر عندما كنت حجوم العينات (250،

500، 1000) وأطول الاختبارات (40، 60)، وهذه النتيجة تتوافق مع نتائج تحليل التباين لمؤشر دقة التقدير BIAS، وقد يعزى ذلك إلى طريقة التقدير (KS) المستخدمة في التقدير، والتي تعتمد في تقديرها على الرتب، حيث إنها تحول الدرجات الخام إلى رتب، ثم تقدر منحنى خصائص الفقرة، ومن ثم تقدر قيم المعالم، فكلما زاد حجم العينة وطول الاختبار يؤدي إلى زيادة عدد الرتب وبالتالي زيادة عدد الرتب المتشابهة وهذا يؤدي إلى خسارة في المعلومات، وبالتالي إلى انخفاض دقة التقدير. كما أن هذه النتيجة تتوافق مع الأدب النظري في أن كفاءة النموذج اللامعلمي تكون أفضل عند حجوم العينات وأطوال الاختبارات الصغيرة.

ثانياً. النتائج المتعلقة بسؤال الدراسة الثاني: "هل توجد فروق دالة إحصائياً عند مستوى الدلالة α=0.05 بين متوسطات مؤشرات دقة تقدير معلمة القدرة θ المقدرة تعزى لمتغير حجم العينة ومتغير طول الاختبار والتفاعل بينهما؟"

للإجابة عن هذا السؤال؛ تم حساب المتوسطات الحسابية والانحرافات المعيارية لقيم القدرة θ المقدرة باختلاف متغيري (حجم العينة، وطول الاختبار)، والجدول 10 يبين هذه النتائج.

جدول (10) المتوسطات الحسابية والانحرافات المعيارية لقيم القدرة θ المقدرة باختلاف متغيري (حجم العينة، وطول الاختبار)

طول الاختبار	الإحصائي		حجم العينة			
			100 فرد	250 فرد	500 فرد	1000 فرد
20 فقرة	المتوسط الحسابي		0.584	0.311	0.214	-0.034
	الانحراف المعياري		2.28	1.88	1.81	1.79
40 فقرة	المتوسط الحسابي		-0.096	0.022	0.040	-0.023
	الانحراف المعياري		1.70	1.48	1.49	1.47
60 فقرة	المتوسط الحسابي		0.124	0.076	0.033	0.009
	الانحراف المعياري		1.51	1.35	1.34	1.30

يلاحظ من الجدول (10) وجود فروق ظاهرية بين قيم المتوسطات الحسابية الخاصة بمعلمة القدرة (θ) المقدرة ناتجة عن اختلاف مستويات المتغيرين، حيث يلاحظ أن أكبر قيمة لمتوسط معلمة القدرة (θ) المقدرة قد كانت 0.584 بانحراف معياري 2.28، عندما كان حجم العينة 100 فرد

وطول اختبار 20 فقرة، كما يلاحظ من الجدول 10 أن أقل قيمة لمتوسط معلمة القدرة (θ) المقدرة قد كانت 0.009 بانحراف معياري 1.30 عندما كان حجم العينة 60 فقرة وطول الاختبار 1000 فرد .

كما تم حساب المتوسطات الحسابية والانحرافات المعيارية لمؤشر التحيز BIAS كمؤشر من مؤشرات دقة التقدير لتقديرات القدرة θ المقدرة باختلاف متغيري (حجم العينة، وطول الاختبار)، وذلك كما هو مبين في الجدول (11).

جدول (11) المتوسطات الحسابية والانحرافات المعيارية لمؤشر التحيز لتقديرات القدرة θ المقدرة باختلاف متغيري (حجم العينة وطول الاختبار)

طول الاختبار	الإحصائي		حجم العينة			
			100 فرد	250 فرد	500 فرد	1000 فرد
20 فقرة	المتوسط الحسابي للتحيز		0.615	0.329	0.224	-0.057
	الانحراف المعياري للتحيز		1.79	1.42	1.33	1.25
40 فقرة	المتوسط الحسابي للتحيز		-0.065	0.040	0.050	-0.046
	لانحراف المعياري للتحيز		1.21	1.02	0.97	0.92
60 فقرة	المتوسط الحسابي للتحيز		0.155	0.094	0.042	-0.013
	الانحراف المعياري للتحيز		0.98	0.73	0.70	0.63

يلاحظ من الجدول (11) وجود فوق ظاهرية بين القيم المطلقة لقيم المتوسطات الحسابية لمؤشر التحيز لتقديرات القدرة ناتجة عن اختلاف مستويات المتغيرين، حيث يلاحظ أن أقل قيمة مطلقة لمتوسط مؤشر التحيز لتقديرات معلمة القدرة (θ) قد كانت 0.013 بانحراف معياري 0.63، عندما

كان حجم العينة 1000 فرد وطول اختبار 60 فقرة، كما يلاحظ من الجدول 11 أن أكبر قيمة مطلقة لمتوسط مؤشر التحيز لتقديرات معلمة القدرة (θ) قد كانت 0.615 بانحراف معياري 1.79 عندما كان حجم العينة 100 وطول الاختبار 20 فقرة .

وفي ضوء ما تقدم، تم إجراء تحليل التباين الثنائي 2-Way Interaction ANOVA لقيم مؤشر التحيز BIAS لتقديرات القدرة θ المقدرة باختلاف متغيري (حجم العينة، وطول الاختبار)، وذلك كما هو مبين في الجدول 12.

جدول (12) نتائج تحليل التباين الثنائي لقيم مؤشر التحيز BIAS لتقديرات القدرة θ المقدرة باختلاف متغيري (حجم العينة، وطول الاختبار)

مصدر التباين	مجموع المربعات	درجة الحرية	متوسط مجموع المربعات	قيمة ف المحسوبة	الدلالة الإحصائية
طول الاختبار	19.822	2	9.911	7.132	0.001
حجم العينة	41.831	3	13.944	10.034	0.000
طول الاختبار× حجم العينة	15.471	6	2.579	1.855	0.085
الخطأ	7696.219	5538	1.390		
الكلي	7773.343	5549			

يتضح من الجدول (12) وجود فروق دالة إحصائياً عند مستوى الدلالة α=0.05 بين المتوسطات الحسابية لقيم مؤشر التحيز BIAS لتقديرات القدرة (θ) المقدرة تعزى لمتغير طول الاختبار ومتغير حجم العينة؛ مما يعني تأثر دقة تقدير معلمة القدرة بمتغيري حجم العينة وطول الاختبار، وللكشف عن مواقع الفروق بين متوسطات مؤشر الدقة (BIAS)، وبعد التأكد من تجانس التباين للمتغيرين حجم العينة، وطول الاختبار كون المتغيران متعددي المستويات باستخدام اختبار levene، تم استخدام اختبار شفيه Schefee للمقارنات البعدية، لمتغيري حجم العينة وطول الاختبار على التوالي، حيث تبين وجود فرق ذي دلالة إحصائية بين حجم العينة 100 فرد من جهة و(250، 500، 1000) فرد من جهة أخرى، ولصالح 250 فرد و500 فرد و1000 فرد على التوالي، كما تبين وجود فرق ذي دلالة إحصائية بين حجم العينة 250 من جهة و(500، 1000)

فرد من جهة أخرى، ولصالح 500 فرد و1000 فرد على التوالي، مما يعني أن دقة تقدير معلمة القدرة (θ) تزداد بزيادة حجم العينة، أما متغير طول الاختبار فقد تبين وجود فرق ذي دلالة إحصائية بين طول الاختبار 20 فقرة و40 فقرة ولصالح طول اختبار 40 فقرة، كما تبين وجود فرق ذي دلالة إحصائية بين طول اختبار 20 فقرة و60 فقرة ولصالح 60 فقرة، مما يعني أن دقة تقدير معلمة القدرة كانت أفضل في حجوم العينات واطوال الاختبار الكبيرة، وقد يعزى ذلك إلى باراميتر التهذيب ((Smoothing Parameter (h)، المستخدم في برنامج (TESTGRAF)، حيث أنه يتأثر في حجم العينة ويزداد بشكل ملاحظ مع زيادة حجم العينة، ويساوي N 1,1؛ وبالتالي يزداد h بزيادة حجم العينة؛ مما يعني زيادة دقة التقدير، وقد يعزى ذلك إلى العلاقة الارتباطية بين دقة تقدير معلمة القدرة ومعالم الفقرة، حيث أن دقة تقدير معلمة التخمين تزداد بزيادة حجم العينة، وهذا يؤثر على دقة تقدير معلمة القدرة، لأن دقة تقدير معلمة القدرة تكون اكبر عندما يكون التخمين اقل ما يمكن، وهذه النتيجة تتوافق مع نتيجة دراسة دي لاتوري وايوان (De la Torre & Yuan,2010) التي أشارت إلى أن حجم العينة يؤثر على قدرة الاختبار في تقدير معلمة القدرة ومعالم الفقرة، وإلى وجود علاقة ارتباطية بين تقدير معلمة القدرة ومعالم الفقرة، كما تتفق هذه النتيجة مع نتائج دراسة فو (Fu, 2010) والتي أشارت إلى أن هناك تباينا في دقة تقدير معالم الفقرة والقدرة حسب مستوى التخمين الموجود في الاختبار، وحجم العينة المستخدمة، وطول الاختبار، كما أشارت النتائج أيضاً إلى أن دقة تقدير معلمة القدرة، ومعالم الفقرة تعتمد على معيار الدقة المستخدم في كل واحد من نماذج نظرية الاستجابة للفقرة.

وكذلك تم حساب قيم مؤشر RMSE كمؤشر من مؤشرات دقة التقدير لتقديرات القدرة θ المقدرة باختلاف متغيري (حجم العينة، وطول الاختبار)، وذلك كما هو مبين في الجدول 13.

جدول (13) قيم مؤشر RMSE لتقديرات القدرة θ المقدرة باختلاف متغيري (حجم العينة، وطول الاختبار)

طول الاختبار	حجم العينة			
	100 فرد	250 فرد	500 فرد	1000 فرد
20 فقرة	1.795	1.232	1.227	1.117
40 فقرة	1.282	0.945	0.944	0.887
60 فقرة	1.089	0.838	0.730	0.606

يلاحظ من الجدول (13) ، أن كافة قيم مؤشر دقة التقدير RMSE لتقديرات القدرة (θ) قد كانت أصغر ظاهريا عندما كان طول الاختبار 60 فقرة وحجم العينة 1000 مفحوص. مما يعني أن دقة تقدير معلمة القدرة (θ) قد كان أكثر دقة ظاهرياً بما كان عليه الأمر عندما كانت حجوم العينات(100، 250، 500) واطول الاختبارات (40، 20) فقرة، وهذه النتيجة تتفق مع دراسة هامبلتون وكوك (Hambleton & Cook, 1980) التي توصلت إلى أنه بزيادة حجم العينة، وعدد الفقرات، تزداد دقة التقدير لقدرة الفرد باستخدام النموذج الثلاثي المعلمة.

الخلاصة والتوصيات:

تبين للباحث من خلال نتائج الدراسة بأنه لا يوجد فرق دال إحصائياً في دقة تقدير معلمة الصعوبة (b)، تعزى لمتغيري حجم العينة وطول الاختبار والتفاعل بينهما، إلا أن هناك تأثيراً لمتغيري الدراسة (حجم العينة، وطول الاختبار) والتفاعل بينهما على دقة تقدير معلمة التخمين (c)، حيث أظهرت النتائج وجود فرق دال إحصائياً في دقة تقدير معلمة التخمين (c) يعزى لمتغير حجم العينة، وبعد إجراء المقارنات البعدية باستخدام اختبار شافيه Schefee تبين ان دقة تقدير معلمة التخمين (c) لا تختلف عندما كان حجم العينة (100، 250) وهي اقل دقة مما كان عليه الحال عندما كان حجم العينة (500، 1000)، كما أظهرت النتائج وجود فرق دال إحصائياً في دقة تقدير معلمة التخمين (c) يعزى لمتغير طول الاختبار، ولصالح طول الاختبار 40 فقرة، وبينت النتائج وجود فرق دال إحصائياً في دقة تقدير معلمة التخمين (c) يعزى للتفاعل بين متغيري حجم العينة وطول الاختبار، واظهرت النتائج وجود فرق دال إحصائياً في دقة تقدير معلمة التمييز (a) يعزي لمتغير حجم العينة، وبعد إجراء اختبار شافية Schefee للمقارنات البعدية تبين أن هناك فرق دال

إحصائياً في دقة تقدير معلمة التمييز (a) ولصالح العينات الصغيرة (100، 250)، واعتمادا على ذلك فان الباحث يوصي مطوري الاختبارات بإمكانية استخدام طريقة (KS) اللامعلمية، لتقدير معالم الفقرة في الفقرات التي يكون فيها التخمين صفر او اقل ما يمكن (مثل الفقرات المقالية)، ويوصي باستخدامها عندما لا تفي البيانات بافتراضات النماذج المعلمية.

كما كشفت نتائج التحليل وجود فرق دال إحصائياً في دقة تقدير القدرة (θ) يعزى لمتغير طول الاختبار، كما تبين وجود فرق دال إحصائياً في دقة تقدير معلمة القدرة (θ) يعزى لمتغير حجم العينة، وبعد إجراء المقارنات البعدية باستخدام اختبار شافيه Scheffe تبين أن دقة تقدير القدرة (θ) كانت أكثر عند حجم العينات الكبيرة (500، 1000) مما كان عليه الأمر عند حجوم العينات الصغيرة (100، 250)، مما يعني أن دقة تقدير القدرة (θ) تحتاج إلى حجوم عينات كبيرة نسبيا، مقارنتا بما كان عليه الأمر عند تقدير معالم الفقرة. واعتمادا على ذلك فان الباحث يوصي مطوري الاختبارات بإمكانية استخدام طريقة (KS) اللامعلمية في تقدير معالم الفقرات، وتحديدا عندما تكون حجم عينة المفحوصين قليلة، حيث تكون دقة التقدير لمعالم الفقرة عالية، مما ينعكس إيجاباً على جودة التقديرات والتخلص من مشكلة حجم العينة وطول الاختبار الكبيرين التي تتطلبها النماذج المعلمية. كما يوصي الباحث باستخدام هذه الطريقة في الغرف الصفية لتقدير معالم فقرات الاختبارات عندما تكون حجوم العينات قليلة، وفي الاختبارات التي يكون فيها التخمين اقل ما يمكن، ويوصي الباحث باستخدام طريقة KS لتقدير قدرات الأفراد في منتصف متصل القدرة؛ بمعنى أن دقة تقدير معلمة القدرة تكون اقل في طرفي متصل القدرة، وذلك ان هذه الطريقة لا تستطيع تقدير قدرة الأفراد الذين أجابوا أو لم يجيبوا إجابة صحيحة على كل الفقرات.

وأخيرا يوصي الباحث بإجراء دراسات أخرى مماثلة باستخدام بيانات حقيقية، ودراسة أثر حجم العينة وطول الاختبار على دقة تقدير معالم الفقرة والقدرة تحت ظروف خاصة من توزيعات مستويات القدرة ومستويات معالم الفقرات، (منتظم، ملتوي لليمين، ملتوي لليسار)، ويوصي الباحث استخدام النموذج ثنائي المعلمة والنموذج أحادي المعلمة لتقصي دقة تقدير معالم الفقرة والقدرة باختلاف متغيري حجم العينة وطول الاختبار، وذلك للسيطرة على معلمة التخمين، كما يوصي بإجراء دراسات مماثلة لكن في ظل تغيير بارامتير التهذيب h .

المراجع

الشريفين، نضال كمال.(2012) .أثر طريقة تقدير معالم الفقرة وقدرات الأفراد على قيم معالم الفقرة، والخصائص السيكومترية للاختبار، في ضوء تغير حجم العينة. المجلة التربوية، 26(104)،177−238 .

عودة، احمد سليمان. (2010). القياس والتقويم في العملية التدريسية. اربد: دار الأمل للنشر والتوزيع.

Baker.(2001). The Basics of Item Response Theory (2nd Ed.). ERIC Clearinghouse on Assessment & Evaluation: USA.

Casella, G., & Berger, R.L.,(1990).Statistical Inference, Belmont, CA, Duxbury. Coombs, C. H. (1964). A theory of data, New York, Wiley.

Crocker, L., &Algina, J. (1986). Introduction to classical and modern test theory. NY: Holt, Rinhart &Winston.

De la Torre, Jimmy; Yuan Hong (2010). Parameter Estimation With Small Sample Size A Higher-Order IRT Model Approach. Applied Psychological Measurement.. 34 (4), p267-285.

Douglas, J. (1997). Joint consistency of nonparametric item characteristic curv and ability estimates. Psychometrika, 47 , 7-28.

Douglas, J. (1997). Joint consistency of nonparametric item characteristic curves and ability estimation. Psychometrika, 62, 7-82.

Fitzpatrick,& Ann . R. (2001) . The effects of test length and sample Size on the reliability and equating of tests Composed of constructed – response item. Applied Measurement in Education, 14(1): 412-425.

Fu, Qiong (2010) Comparing Accuracy of Parameter Estimation Using IRT Models in the Presence of Guessing. Unpublished PhD dissertation, Illinois University, USA.

Hambelton, K.(1993). Principles and selected applications of item response theory. in R. L.linn(Ed) Educational Measurement. (3 ed.) (pp.147-200)phoenix: theory press.

Hambleton, R. K., & Cook, L. (1980). The robustness of latent trait models and effects of test length and sample size on the precision of ability estimates. In D. J. Weiss (Ed.), Proceeding of the 1979 Computerized Adaptive Testing Conference, 36-52. Minneapolis, Minnesota: University of Minnesota, Department of Psychology. Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Hambleton, R. Swaminathan, H. and Rogers, H. (1991) .Fundamental of item response theory .New York: Sage.

Hambleton, R.,. & Swaminthan, H. (1985). Item response Theory Principle and Application. Boston: Kluwer: Nijhoff Publishing.

http://www.psych.mcgill.ca/faculty/ramsy.htm.

Kerlinger, F. N. and Pedhazur, E.J.(1973) Multiple regression in Behavioral research. New York: Holt, Rinchart and Winston, Inc,

Kingma, J .,& Tenvergert, E. (1985). A Nonparametric Scale Analysis of development of Conservation. Applied Psychological Measurement, 9(2),375-387.

Koning, E.; Sijtsma, K.&Hamers, J.(2002). Comparison of Four IRT Models When Analysing Two Test for Inductive Reasoning .Applied Psychological Measurement, 26(3), 302-320

Lew, P., Dunbar, S., and Kolen, M. (2004). A Compression of parametric Approaches to Item Analysis for multiple – choice tests, Education and psychological measument, G4 (4), 565-587

Liang, T .(2010).An Assessment of the Nonparametric Approach of Evaluating the fit of Item Response Model. Dissertation Abstract International, (UMI No. 3397726).

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

Psychometrica, 47 , 397- 412

Ramsay, J. (1991). Kernel Smoothing Approaches to Non-para Metric Item characteristic Curve Estimation. Psychometrika, 56 (4), 611 – 630.

Ramsey, J.(2000). Test Graf: A program For the graphical analysis of multiple-choice tests and questionnaire data [Computer Software and manual] .Retrieved from

Reckase, &Mark, D. (1978) A Comparison of the one and three Parameter logistic model for item calibration. Paper Presented at the annual meeting of the American Educational Research Association, Toronto,Canada. {on - Line}Available : http :// eric .ed .gov.

Reeve, B. (2004). Applications of item response theory (IRT) modeling for building an evaluating questionnaires measuring patient- reported outcomes. Web site: httb://outcomes.cancer.gov/conference/irt/ree-ve.pdf.

Sijtsma, K .& Hamker, B . T .(2000). A taxonomy of IRT models for ordering of persons and item using simple sum scores .Journal of Educational and Behavioral Statistics, 25(2), 391-415 .

Sijtsma, K .(1998). Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. Applied Psychological Measurement, 22,3-31.

Sijtsma, K. & Molenaar, I. (2002). Introduction to Nonparametric Item Response Theory. Sage Publication, International Educational and Professional Publisher. Thousand Oaks : London . New Delhi.

Thisen , D., & Wainer, H, (1982). Some standard errors in item response theory.

Van der Linden, W.J, &Hambleton , R . K. (1997). Part IV: Non-Parametric models introduction in W. J. van der Linden & R. K. Hambleton, (1994), Handbook of modern item response theory (pp.347- 349).New York : Springer- Verlag.

Wang, T., & Vispoel, W. P. (1998). Properties of ability Estimation Methods in Computerized Adaptive Testing. Journal of Educational Measurement, 35(2), 109 – 135 .

Weiss, D. J. (2009). Termination Criteria in Computerized Adaptive Tests: Variable – Length CATs Are Not Biased. Paper Presented in the 2007 GMAC Conference on Computerized Adaptive Testing .Minneapolis .