أثر صعوبة الفقرة وحجم العينة في دقة معادلة درجات الاختبارات باستخدام نظرية استجابة الفقرة (IRT)

يوسف عبد العاطي المحروق*

وزارة التربية والتعليم، مملكة البحرين

استُلم بتاريخ: ٢٠١٤/١٠/٣٠ عُدل بتاريخ: ٢٠١٥/٣/٣ قُبل بتاريخ: ٢٠١٥/٤/٣١

المستخلص: هدفت هذه الدراسة إلى معرفة أثر صعوبة الفقرة وحجم العينة في دقة معادلة درجات الاختبارات باستخدام نظرية استجابة الفقرة (IRT)، وذلك من خلال دراسة المتغيرات الآتية: حجم العينة: وقد استخدمت ثلاثة مستويات: ٢٠٠، ٦٠٠، ١٠٠٠ وهذه الأحجام للعينات تعتبر مناسبة لطرق المعادلة باستخدام تصميم المجموعات العشوائية، ومتغير مستويات الصعوبة، وله مستويان: التشابه في متوسط معدل الصعوبة للاختبار، والاختلاف في متوسط معدل الصعوبة، تم التوصل إلى نتائج دقة المعادلة في ظل استخدام المتغيرات السابقة منفردة ومجتمعة، ولمقارنة دقة المعادلة، تم توليد بيانات تجريبية باستخدام برمجية (Wingen2)، تم معادلة درجات الاختبارات باستخدام الدرجات الملاحظة في نظرية استجابة الفقرة كمعيار رئيسي للمعادلة. أظهرت النتائج أن حجم العينات الكبير يقلل من الخطأ المعياري للمعادلة ويقلل من البواقي المعيارية. كما أظهرت النتائج أن النماذج المختلفة في صعوبتها تميل قيم الخطأ المعياري وقيم RMSE إلى الانخفاض عندما تختلف مستويات الصعوبة فيها، والنماذج المتشابهة في صعوبتها تميل قيم الخطأ المعياري وقيم RMSE إلى الارتفاع عندما تتشابه مستويات الصعوبة فيها.

كلمات مفتاحية: نظرية استجابة الفقرة، معادلة الاختبارات.

Effect of Item Difficulty and Sample Size on the Accuracy of Equating by Using Item Response Theory

Yousef A. Al Mahrouq*
Ministry of Education, Kingdom of Bahrain

Abstract: This study explored the effect of item difficulty and sample size on the accuracy of equating by using item response theory. This study used simulation data. The equating method was evaluated using an equating criterion (SEE, RMSE). Standard error of equating between the criterion scores and equated scores, and root mean square error of equating (RMSE) were used as measures to compare the method to the criterion equating. The results indicated that the large sample size reduces the standard error of the equating and reduces residuals. The results also showed that different difficulty models tend to produce smaller standard errors and the values of RMSE. The similar difficulty models tend to produce decreasing standard errors and the values of RMSE.

Keywords: Item Response Theory (IRT), equating test.

* yousif.almahrooq@moe.gov.bh

سيطرت نظرية القياس النفسي التربوي الكلاسيكية على الفكر التربوي، وظل العاملون بالقياس النفسي يستخدمون مبادئ وأسس هذه النظرية في بناء الاختبارات والمقاييس بأشكالها المختلفة، وتفسير الدرجات المتحققة عليها لفترة طويلة من الوقت، وعلى الرغم من الاستخدام الواسع للنظرية الكلاسيكية في عملية القياس النفسي والتربوي إلا أنها تعاني من جوانب قصور أشار إليها هامبلتون ولندن Linden, (Hambelton & Van der 1982) منها: اختلاف خصائص فقرات الاختبار باختلاف عينات المفحوصين المستخدمين في معايرة الفقرات، إذ أن خصائص فقرات الاختبار تعتمد على مستوى وتوزيع قدرات أفراد هذه العينات، فصعوبة الفقرات تكون أعلى مما هي في الواقع في حالة اختيار عينة تكون في مستوى قدرتها أعلى من مستوى قدرة الأفراد في المجتمع، كما تكون أدنى مما يجب في حالة اختيار عينة تكون في مستوى قدرتها اقل من مستوى قدرة أفراد المجتمع. كذلك من جوانب القصور التي تعاني منها النظرية الكلاسيكية ان الاختبارات في النظرية الكلاسيكية لا تزودنا بتقدير دقيق للمفحوصين ذوي القدرة العالية وللمفحوصين ذوي القدرة المتدنية؛ وذلك لأن الاختبارات المبنية على أساس النظرية الكلاسيكية في القياس تأخذ في اعتبارها عند اختيار الفقرات أن تتلاءم مع الأفراد متوسطي القدرة.

ولتلافي العيوب التي ظهرت في النظرية الكلاسيكية ومن اجل الوصول إلى قياس أكثر موضوعية، حاول علماء القياس المعاصرون الاستفادة من التقدم التكنولوجي في التوصل إلى طرق سيكو متريه جديدة تساعد في التغلب على بعض نواحي القصور في النظرية الكلاسيكية، لذلك انبثقت عن هذه الطرق الجديدة نماذج تسمى نماذج السمات الكامنة أو نظرية الاستجابة للفقرة (Item Response Theory)، أو ما يسمى بالنظرية الحديثة في القياس التي تفترض أن هناك دالة احتمالية بين معلمتين (Two Parameters) احداهما تتعلق بقدرة الفرد والأخرى تتعلق بالفقرة التي يختبر بها، وبالتالي فان هذه النظرية تهدف إلى التوصل إلى قيم تقديرية لكل واحدة من هاتين المعلمتين، ومن ثم استخدام هذه القيم في تقدير احتمالية الإجابة الصحيحة لكل مفحوص بدرجه عالية من الدقة والثبات.

لقد بين هامبلتون وسوامنثيان (Hambelton & Swaminathan, 1985) مزايا رئيسيه لنظرية استجابة الفقرة منها: أن تقدير قدرة الفرد يكون مستقلا عن عينة الفقرات التي تطبق عليه، أي أن تقديرات القدرة للأفراد متحررة من خصائص الفقرات المستخدمة في تقدير القدرة (Item free). كذلك ومع افتراض وجود عدد كبير من الأفراد، يكون تقدير معالم الفقرات مستقلا عن عينة الأفراد التي استخدمت في تقدير هذه المعالم (Sample free). إن نظرية الاستجابة للفقرة تقوم على افتراضات أساسية ذكرها هامبلتون وساميثيان (Hambelto & Swaminathan,1991) وهي: افتراض أحادية البعد (Unidimensionality)، وافتراض الاستقلال الموضعي (Local Independence) والمطابقة لمنحنى خصائص الفقرة (Item Characteristics Curve (ICC)، والتحرر من سرعة الأداء (Speediness)، كذلك تعتمد نظرية استجابة الفقرة على وجود عدد من النماذج الرياضية المستخدمة لمطابقة النموذج للبيانات التي من أهمها: النماذج اللوجستيه الأحادية والثنائية الثلاثية والرباعية. إن النجاح في استخدام نماذج نظرية الاستجابة للفقرة يقوم على مجموعة من المتطلبات منها: أحادية البعد للاختبار، مطابقة فقرات الاختبار للنموذج المستخدم، حجم العينة المستخدمة، توفر البرمجيات الحاسوبية المناسبة، والطريقة المستخدمة في تقدير

معالم الفقرات ومعلمة القدرة (Hambelton & Swaminathan, 1991).

لقد ظهرت عدة تطبيقات عملية لنظرية استجابة الفقرة ومن هذه التطبيقات: الكشف عن التحيز في فقرات الاختبارات، بنوك الأسئلة، الاختبارات التكيفية، بناء الاختبارات وتحليلها وتقنينها، وكذلك معادلة الدرجات على صور الاختبار المختلفة. ويصنف المختصون في القياس والتقويم التربوي والنفسي طرق معادلة الاختبارات إلى نوعين: الطرق التي تعتمد على النظرية التقليدية في الاختبارات Classical Test Theory.(CTT) أما النوع الثاني فيُصنّفُ ضمن الطرق التي تعتمد على النظرية الحديثة في القياس التربوي (نظرية الاستجابة للفقرة) Item Response Theory (IRT).

إن معادلة الاختبارات المعتمدة على الدرجات الخام raw scores ضمن النظرية التقليدية للاختبارات قد لا تكون مرغوبة؛ وذلك بسبب إخفاقها في تحقيق جميع شروط معادلة الاختبارات، وهي العدالة والمساواة، والتماثل، واللاتباين (Hambleton & Swaminathan, 1985). لذلك، فإن معادلة الاختبار عن طريق نظرية استجابة الفقرة (النظرية الحديثة في القياس التربوي والنفسي-Item Response Theory) تحل الكثير من المشكلات التي عجزت عن حلها النظرية التقليدية في الاختبارات؛ بشرط أن يكون النموذج IRT Model المستخدم في النظرية الحديثة هذه مطابقاً للبيانات المُعدة لعملية المعادلة. ويشير باكر والكارني (Baker & Alkarni, 1991, p 147) إلى أنه من الإسهامات الكبيرة للنظرية الحديثة في القياس التربوي والنفسي -والتي اصطلح على تسميتها بنظرية الاستجابة للفقرة (- Item Response Theory IRT) في الممارسة التربوية، وذلك لقدرتها على وضع عدة اختبارات ومجموعات من المفحوصين على تدريج مشترك common scale في عملية القياس؛ وإمكانية استخدامها في المعادلة الأفقية والرأسية للاختبار.

تقوم نظرية الاستجابة للفقرة على مجموعة من الافتراضات، تؤدي إلى التفسير الصحيح لنتائج الاختبار ومعادلته، بشرط أن يتم تطبيقها بشكلٍ صحيح ودقيق. حيث تفترض هذه النظرية أن أداء المفحوصين في الاختبار يمكن تفسيره عن طريق السمة أو السمات الكامنة latent Traits المراد قياسها، والتي لا يمكن قياسها بصورة مباشرة. إذ يتم استخدام الدرجات التي تم تقديرها للمفحوص في تلك السمة في التنبؤ بأدائه في اختبار ما أو في فقرة من الاختبار؛ لأن العلاقة الحقيقية بين الدرجات المشاهدة (الخام) للمفحوص والسمة المراد قياسها لا يمكن الحصول عليها بطريق مباشر. ومنة هنا، تقوم نظرية الاستجابة لمفردة الاختبار بوصف هذه العلاقة بواسطة دالة تعتمد على مجموعة من الافتراضات، هي: أحادية البُعْدْ (أحادية السمة) Unidimensional حيث يقيس الاختبار سمة واحدة فقط؛ والاستقلال الموضعي local independence وهو استقلال أداء المفحوص على فقرة اختبار عن أدائه على فقرة أخرى. وكذلك جعل السمات الأخرى التي تؤثر على أداء المفحوص ثابتة ومتسقة Consistent. أما الافتراض الأخير فهو افتراض اللاتباين Invariance والذي يعني أن معالم (معلمات) Parameters الفقرة (الصعوبة، والتمييز، والتخمين) لا تعتمد على التوزيع الإحصائي للسمة المراد قياسها؛ وأن المعالم التي تصف أداء المفحوصين لا تعتمد على فقرات الاختبار (Hambelton, Swaminathan, & Rogers, 1991).

خطوات معادلة الاختبارات باستخدام نظرية استجابة الفقرة:

وضح كل من هامبلتون وسواميناثان (Hambleton & Swaminthan, 1985)، وكذلك كولن وبرينان (Kolen & Brennan, 1995)، الخطوات الضرورية لمعادلة الاختبارات بواسطة نظرية استجابة الفقرة؛ وهي كالآتي:

- اختيار التصميم المناسب لمعادلة الاختبار مع الأخذ بعين الاعتبار

خصائص مجموعة المفحوصين وطبيعة الاختبارات المراد معادلتها.

- إختيار النموذج المناسب الذي يطابق التصميم المناسب والاختبار المناسب (نموذج راش أو غيره من نماذج هذه النظرية).
- بناء تدريج مشترك يربط العلاقة بين السمة المراد قياسها ومَعْلَم الفقرة Item Parameter.
- إختيار التدريج المناسب لوضع درجات الإختبار. أي هل تُكْتَبُ الدرجات كدرجات خام (درجات مُشَاهَدَة)، أو على صورة درجات قُدرة ability scores، أو على صورة درجات حقيقية مقدرة estimated true score، ويمكن أن تقوم بهذه المهمة المعقدة رياضياً بعض البرامج الحاسوبية، مثل برنامج BILOG، وبرنامج MULTILOG وبرنامج LOGIST، وغيرها. ومن المتعذر عملياً إلى حد كبير القيام بهذه العملية يدوياً، وخاصة في هذه النظرية.

ويوضح كوكر وايجنور (Cooker & Eignor, 1991) الآليات الأساسية لعملية معادلة الإختبارات في هذه النظرية، والتي يمكن إيجازها في الآتي:

- إختيار التصميم المناسب: هناك ثلاثة تصميمات أساسية تستخدمها هذه النظرية لمعادلة الاختبارات؛ وهي تصميم المجموعة المفردة، وتصميم المجموعات العشوائية، والتصميم ذو الاختبار المشترك. ويعتمد حجم العينة الملائم لإجراء معادلة الاختبار بشكل صحيح على العدد الملائم للمفحوصين للحصول على تقديرات مستقرة للمعلمات المستخدمة في النموذج المختار لوضع الدرجات ومعلمات القدرة (السمة) على تدريج واحد. وفي عملية المعايرة هذه نحتاج إلى عينة تصل إلى ٣٠٠٠ مفحوص. وإذا كانت هناك فقرات مشتركة بين صورتي الإختبار على هيئة اختبار مشترك Anchor Test، فيجب أن تعكس هذه الفقرات المحتوى والخصائص الإحصائية لصورتي الاختبار؛ وألا تقل نسبة فقرات الاختبار المشترك عن ٢٠% من الطول الاحتمالي للاختبار. كما أن اختيار أي من التصميمات المشار إليها أعلاه يرتبط بنوعية البرنامج الحاسوبي المستخدم لمعادلة الاختبار.
- وضع تقديرات المعالم (المعلمات) على تدريج مشترك: لو افترضنا أن مجموعتين من المفحوصين طُبقت عليهما نفس المجموعة من فقرات الاختبار، وتم تقدير معلمات الفقرة (الصعوبة والتمييز) لكل مجموعة على حده، وأن النموذج الرياضي المستخدم هو النموذج ثنائي المعلم؛ حيث أن منحنيات خاصية الفقرة (ICC: Item Characteristic Curves) مستقلة عن المجموعتين المستخدمتين لرسم هذه المنحنيات. ويمكن أن نستنتج من ذلك أن تقديرات معلمات الفقرات متطابقة في كل من المجموعتين على حده. ويجب هنا الأخذ بعين الاعتبار أثر خطأ المعاينة Sampling Error. ولكن الواقع ليس كذلك! إذ يجب إجراء تحويل رياضي معين يوحد التدريج في عملية المعايرة. ويعتمد ذلك طبعاً على نوع البرنامج الحاسوبي المستخدم في إجراء عملية التحويل الرياضي للحصول على نقطة أصل ووحدة قياس للسمة ولمستوى الصعوبة. ويكون متوسط درجات القدرة (السمة) هو الصفر وانحرافها المعياري هو الواحد الصحيح (برنامج LOGIST يمكن أن يحقق ذلك). ويجب ملاحظة أن عملية التحويل الرياضي التي تضع الدرجات وتقديرات معلمات السمة على نفس التدريج تستهدف الحصول على معلمَي المَيْلْ

Slop والقاطع Interception، بحيث يكون الوسط الحسابي والانحراف المعياري لتوزيع مستويات صعوبة الفقرات، التي تم تقديرها في عملية المعايرة الثانية للدرجات، مساويان لنظيريهما اللذين تم تقديرهما في المعايرة الثانية للدرجات، مساويان لنظيريهما اللذين تم تقديرهما في المعايرة الأولى للدرجات.

- معادلة درجات الاختبار: تُعتبر عملية معادلة الاختبار منتهية إذا تم تقدير المعلمات لفقرات كل من صورتي الاختبار المستهدف، وتم وضعها على تدريج مشترك، وتم كذلك الحصول على تقدير للسمة المراد قياسها لدى المفحوص؛ بحيث تكون هي نفسها في أي من صورتي الاختبار. ويؤخذ خطأ القياس بعين الاعتبار هنا. وبناء على ذلك، يكون التعبير عن الدرجات الخام بما يكافئها من درجات السمة. أما إذا أخفق البرنامج الحاسوبي المستخدم في استخرجا نتيجة السمة، يلجأ المختصون إلى ترجمة أو تحويل أي درجة من درجات القدرة إلى الدرجة الحقيقية المقدرة المناظرة لها في صورتي الاختبار؛ واعتبارها الدرجة التي تمت معادلتها في الاختبار. أما الصورة الرياضية للدوال التي تربط بين درجات القدرة وتقديرات الدرجات الحقيقية فهي كالتالي:

$$\hat{T}_x = \sum_{i=1}^{n} \hat{P}_i(\theta)$$

$$\hat{T}_y = \sum_{j=1}^{n1} \hat{P}_j(\theta)$$

$\hat{T}_x$ = الدرجة الحقيقية المقدرة للصورة الأولى من الاختبار. $\hat{T}_y$ = الدرجة الحقيقية المقدرة للصورة الثانية من الاختبار. $\hat{P}_j(\theta)$ = الدالة المقدّرة لاستجابة الفقرة في الفقرات للصورة الأولى للاختبار.

$\hat{P}_j(\theta)$ = الدالة المقدّرة لاستجابة الفقرة للصورة الثانية في الاختبار. علماً بأن التحويل الرياضي للدرجات في كل من صورتي الاختبار يكون مستقلاً عن مجموعة المفحوصين التي تم الحصول على بيانات معادلة الاختبار منها لإجراء هذه التحويل. ويجب أن نلاحظ هنا أنه إذا كانت الصورة القديمة للاختبار المراد معادلته أكثر صعوبة من الصورة الجديدة في بعض المستويات، فإنها تُعطي تقديراً منخفضاً للدرجة الحقيقية المطلوب الوصول إليها عن طريق تقدير درجة السمة. (الدوسري، ٢٠٠٤)

الفوائد العملية والتطبيقية لمعادلة الاختبارات باستخدام نظرية استجابة الفقرة

تشير نتائج الأبحاث التي أجريت على معادلة الاختبار بواسطة هذه النظرية، إلى أن معادلة الاختبار بهذه النظرية لها فوائد جمة من الناحيتين العملية والتطبيقية (Kolen & Brennan, 1995) يمكن إيجازها في الآتي:

- معادلة الاختبار بهذه النظرية هي الأفضل عندما تكون الاختبارات المختلفة فقراتها في مستويات الصعوبة مطبقة على مفحوصين من مجموعات غير عشوائية، تختلف في مستويات السمة المراد قياسها.
- معادلة الاختبار بهذه النظرية، وبسبب خصائصها المشروحة سلفاً، تؤدي إلى تحويلات رياضية للدرجات تكون مستقلة عن مجموعة أو مجموعات المفحوصين التي طُبقت عليها الاختبارات.
- المعادلة بنظرية IRT هي أفضل من نظريتها في النظرية التقليدية للقياس التربوي، وخاصة في النهايات العليا لتدريج الدرجات، حيث غالباً ما تتم عملية اتخاذ القرارات الهامة، وحيث يمكن معادلة الدرجات الخام مع كل قيم درجات لسمة.
- المعادلة بهذه النظرية تسمح باستخدام الصورة القديمة للاختبار لمعادلة درجاتها بالصورة الجديدة للاختبار متى ما تم

$$P(\theta) = \frac{1}{1+e^{D(\theta-b)}}$$

$$P(\theta) = \frac{1}{1+e^{Da(\theta-b)}}$$

$$P(\theta) = \frac{1}{1+e^{Da(\theta-b)}}$$

- وضع الدرجات في كلا الاختبارين، وكذلك تقديرات معلمات القدرة، على تدريج واحد.
- معادلة الاختبار بهذه النظرية تسمح بمعادلة الاختبار على مستوى الفقرة الواحدة، وذلك قبل الشروع في تطبيق الاختبار في صوره المتعددة، وقبل تطبيق صورتيه أيضاً. ويتم ذلك إذا توافرت البيانات القبلية على مستوى الفقرة الواحدة، مع إمكانية معايرتها. ثم يتم وضع تقديرات المعلمات على تدريج مشترك، وهذه الخاصية لا يمكن الحصول عليها من خلال معادلة الاختبار بواسطة النظرية التقليدية في الاختبارات (Kolen, 1988).

يمكن التعبير عن العلاقة الرياضية بين درجة المفحوص (الدرجة الخام) على الفقرة ودرجته على سُلّمْ السِّمة المراد قياسها بعدة نماذج رياضية في نظرية استجابة الفقرة؛ وهي نموذج راش Rash أحادي المَعْلَمَةْ (نسبةً إلى العالم الرياضي الدنماركي جورج راش)، والنموذج ثنائي المَعْلَمة Two-Parameter Model، والنموذج ثلاثي المَعْلَمَةْ Three-Parameter Model؛ ويعبر عن هذه النماذج رياضيا كما يأتي:

نموذج راش أحادي المَعْلَمْ

النموذج ثنائي المَعْلَمْ

النموذج ثلاثي المَعْلَمْ

حيث: P (θ) = احتمال السمة المراد قياسها.a = معامل تمييز الفقرة (قيمته ثابتة في نموذج راش).

^{b}= معامل صعوبة الفقرة.C = مَعْلَمْ التخمين للفقرة.D = مَعْلَمْ تدريج Scale Parameter يجعل الدالة اللوجستية Logistic Function قريبة إلى الحد الأقصى من دالة القوسية الطبيعية Normal Ogive Function وقيمة هذا لمَعْلَمْ تساوي ١.٧، e= الثابت الرياضي الذي يحول الدالة التي تربط بين الدرجة الخام والسمة المراد قياسها، من دالة مالا نهاية إلى دالة احتمالية تحصر العلاقة بين أداء الطالب على الفقرة وصعوبة الفقرة بين الصفر والواحد الصحيح. وقيمة هذا الثابت الرياضي تساوي ٢,٨١٧.

طرق معادلة الاختبارات باستخدام نظرية استجابة الفقرة

هناك ثلاث طرق رئيسية لمعادلة الاختبار بهذه النظرية وهي(Kolen & Brennan, 1995).

(أ) معادلة الاختبارات باستخدام درجات القدرة (السّمة) Ability Score Equating.

في هذا النوع من معادلة الاختبارات نفترض أن كلا الاختبارين (X, Y) يقيسان نفس القُدرة (السمة) (θ)، لذلك فعند تقدير معلمة القدرة ومعالم الفقرات في نفس الوقت لكل من الاختبارين فإننا نضعهما على نفس التدريج؛ حيث θx = θy. وللقيام بهذه الخطوة فإننا نحتاج إلى معادلة درجات القدرة أثناء معايرة calibration الاختبارات. وتعني المعايرة هنا تقدير معلمات الفقرات، كالصعوبة والتمييز. وينتج عن ذلك علاقة خطية بين تدريج القدرة وتقدير معلمات الفقرات. بشرط أن يكون التقديران منفصلين. ولتحديد القيم (ΘX) و(Θy) يمكن استخدام عدة طرق نذكر منها:

- طرق الانحدار (Regression methods).
- طريقة المتوسط الحسابي والانحراف المعياري (Mean and sigma).
- طريقة (Robust Mean and Sigma).
- طريقة خصائص المنحنى (Characteristic curve methods).

(ب) معادلة الاختبارات باستخدام الدرجة الحقيقية True-Score Equating

في النظرية التقليدية للاختبارات، يُطْلَقُ على الدرجة المتوقعة للمفحوص (g) على الفقرة مصطلح "الدرجة الحقيقية". بينما في نظرية استجابة الفقرة، تصاغ رياضياً على الصورة:

$$\sum_{1=1}^{n} P_1(\theta_g)$$

فلو افترضنا أن الاختبار بصورتيه يقيس نفس السمة، وأن معالم الفقرات لكلا الصورتين أو الاختبارين قد تم وضعهما على نفس التدريج؛ فإن الدرجة الحقيقية الصحيحة للطالب هي ζ في الاختبار X، وتكون η في الاختبار Y؛ وتكونان مرتبطتين بدرجات القدرة (θ) بواسطة دالة خاصية الاختبار Test Characteristic Curve، بالصورة الرياضية الآتية (Lord, 1980).

$$\eta = \sum_{1=1}^{n} P_1(\theta) \qquad \zeta = \sum_{1=1}^{n} P_1(\theta)$$

ويجب أن تكون تدرجات القياس للدرجات ζ η، مستقلة عن مجموعة المفحوصين n في الاختبار، وأن يكون تدريج قياس درجات القدرة مستقلاً عن عدد فقرات الاختبارين، كما يُشترط أن تكون مستويات صعوبة الفقرات متطابقة في الاختبارين.

ومن عيوب معادلة الاختبار بهذه الطريقة، ان الدرجات الحقيقية المُقدّرة لا تناظر الدرجات الخام مناظرة واحد لواحد. والعيب الآخر هو أن الدرجات الحقيقية المُقدّرة تُوضع على نفس التدريج مع الدرجات الخام Peterson et. al., (1989).

(ج) معادلة الاختبارات باستخدام طريقة الدرجات المشاهَدَة (الدرجات الخام) Observed-score Equating

من المشكلات التي يواجهها المختصون في معادلة الاختبار بطريقة الدرجات الحقيقية، أن هذه الطريقة لا تنتج عنها درجات معادلة للمفحوص الذي درجته الخام أدنى من مستوى الصدفة(التخمين) (C)؛ وذلك لان العلاقة بين الدرجات الخام ليست مثل العلاقة بين الدرجات الحقيقية. ففي الدرجات الخام تكون أدنى درجة هي الصفر، وفي الدرجات الحقيقية تكون أدنى درجة هي، أي درجة التخمين.

$$\sum_{1=1}^{n} (C)_i$$

تقوم معادلة الاختبار بطريقة الدرجات الخام على فكرة التنبؤ بالتوزيع النظري للدرجات الخام للاختبار عن طريق بناء التوزيع التكراري الذي تمثله الدالة $f(X \mid \theta)$ للدرجات الخام للاختبار لمفحوص قدرته (θ). فإذا وجدنا أن دوال الاستجابة لكل فقرة من فقرات الاختبار متطابقة، بحيث يكون P1(θ) = P(θ)، فإن التكرار النسبي للدرجات الصحيحة (X) للمفحوص (g) يمكن حسابه رياضياً بالمعادلة الآتية ضمن توزيع ذي الحدين:

كما يمكن معادلة الاختبار بهذه الطريقة للوصول إلى دالة التكرار النسبي ذي الحدين بإتباع الخطوات الآتية:

$$\int(X|\theta_g) = \binom{n}{x} P_x Q^{n-x}$$

١. وضع معلمات القدرة ومعلمات الفقرات على تدريج مشترك لكل المجموعات والاختبارات.

٢. الحصول على التوزيع التكراري الهامشي marginal frequency distribution للدرجات في الاختبار الأول باستخدام تقديرات المعلمات على الاختبار، وتقديرات معلمات القدرة، باستخدام الدالة الرياضية الآتية:

$$\int(X) = \sum_{1-1}^{n} \int(X|\theta_g)$$

٣. تكرار الخطوة رقم (٢) للاختبار الثاني.

٤. إجراء معادلة للاختبار بطريقة الرتب المئينية المتساوية بين الدرجات الخام في الاختبار الأول ونظيراتها في الاختبار الثاني، وذلك باستخدام التوزيع التكراري الهامشي الذي تم إنشاؤه. ومن المهم جداً

في معادلة الاختبار بهذه الطريقة تغطية مدى الدرجات الخام بالكامل.

أما فيما يتعلق بالدراسات السابقة المتعلقة بمعادلة الاختبارات باستخدام نظرية استجابة الفقرة (IRT) فقد أجرى (Huigin, 2004) دراسة هدفت هذه الدراسة إلى التحقق من طرق المعادلة في نظرية استجابة الفقرة مع وجود القيم الشاذة، تم توليد بيانات تجريبية باستخدام تصميم المجموعات غير المتكافئة ذات الجذع المشترك وتحت ظرف اختلاف قدرة المفحوصين (المجموعات المتكافئة ن(٠،١) مقابل ن (١،١) ، كما تم تطبيق عشر طرق من طرق المعادلة في نظرية استجابة الفقرة لمعادلة البيانات التي تم توليدها تجريبيا. ولتقييم دقة هذه الطرق في المعادلة تم استخدام الأخطاء المنتظمة. أشارت النتائج انه عندما تكون القيم الشاذة متضمنة في البيانات، فان أداء الطرق العشرة المستخدمة في المعادلة تؤدي إلى ظهور اختلافات في قدرة المفحوصين وفي توزيع نقاط الدرجة للقيم الشاذة.

وأجرى (الصمادي، ٢٠٠٦) دراسة هدفت إلى الكشف عن فاعلية طرق تصحيح الصواب والخطأ المتعدد وتأثيرها على دقة معادلة الاختبارات باستخدام نماذج النظرية الحديثة للقياس. ولتحقيق هذا الهدف قام الباحث ببناء اختبار تحصيلي مؤلف من (٣٥) فقرة لكل منها أربعة بدائل من نوع فقرات الصواب والخطأ المتعدد في مبحث الرياضيات. تكونت عينة الدراسة من (٨٧٣) طالبا وطالبة من الصف الأول الثانوي موزعين على عشرة مدارس تشمل تسعا وعشرين شعبة في تخصصات الأدبي والإدارة المعلوماتية والعلمية للعام الدراسي (٢٠٠٤ / ٢٠٠٥). أظهرت النتائج أن طريقة التصحيح الرابعة -وهي إعطاء الطالب علامة واحدة لكل بديل تم الإجابة عليه بشكل صحيح – والتي تعتمد على مراعاة المعرفة الجزئية، كانت الأكثر دقة في قياس قدرات الأفراد، وتعطي معلومات أكثر عن الاختبار، كما كانت الأكثر فاعلية في معادلة الاختبار.

أما (Mao, 2006) فقد أجرى دراسة هدفت إلى فحص دقة تقديرات الأخطاء المعيارية لمعادلة كيرنيل تحت ظروف مختلفة من: حجم العينة ودرجة التمهيد واختيار(bandwidth) وخصائص توزيعات الدرجة وذلك باستخدام تصميم المجموعات العشوائية لتقدير الخطأ المعياري للمعادلة (SEE) حيث تم اعتباره معيارا للمعادلة. وقد أشارت النتائج أن دقة تقدير الخطأ المعياري في المعادلة (SEE) كان أفضل في العينات ذات الحجم الكبير وذات إل (bandwidth) الكبير، كذلك أشارت النتائج أن الدرجات العالية من التمهيد تميل إلى إنتاج حجم أكبر من الأخطاء المعيارية للمعادلة مقارنة مع الدرجات المنخفضة من التمهيد.

وأجرى (Robert, 2007) دراسة هدفت إلى مقارنة الدرجات الحقيقية في نظرية استجابة الفقرة ونظرية استجابة الفقرة بالاعتماد على المعادلة الموضعية، قام الباحث باستخدام اختبارات مختلفة وبنماذج مختلفة لفحص الأداء؛ وذلك لتقييم نتائج معادلة الدرجات الحقيقية مع معادلة الدرجات المشاهدة في نظرية استجابة الفقرة تحت ظروف مختلفة: طول الجذع المشترك، فقدان بعض البيانات، طريقة التدريج، وتوزيع قدرات المفحوصين. تم تقييم نتائج المعادلة بالاعتماد على البواقي المعيارية في المعادلة والتحيز. أشارت النتائج أن تقديرات الدرجات الحقيقية أظهرت تقديرات اقل في معياري التحيز والبواقي المعيارية، كذلك لم توفق الدراسة للإشارة بوضوح لاختيار طريقة التدريج لكلا طريقتي المعادلة.

وأجرى (Youngwoo, 2007) دراسة هدفت إلى مقارنة الخطأ المعياري للمعادلة باستخدام طرق نظرية استجابة الفقرة والمئينات مع فقرات متعددة الاستجابة باستخدام تصميم المجموعات غير المتكافئة ذات الاختبار المشترك. استخدمت هذه الدراسة بيانات حقيقية من اختبار مخصص لتقييم الكتابة، حيث عدلت البيانات الأصلية لعمل خمس صور من الاختبارات وثلاث صور أخرى للاختبار، وتم تطبيق معايير الخطأ المعياري والبواقي المعيارية على طرق المعادلة

الستة المستخدمة في هذه الدراسة، وبحجم عينات (ن=٢٥٠-٥٠٠-١٠٠٠-١٦٨٠) للاختبارات الخمسة وتم حسابه باستخدام Bootstrap، وبأحجام عينات (٥٠٠-١٠٠٠-١٦٨٠) للاختبارات الثلاثة. أشارت النتائج بصورة عامة أن طريقة المعايرة المشتركة أظهرت أخطاء معيارية وبواقي معيارية أقل من طريقة المعايرة المنفصلة، كذلك أظهرت النتائج أن معادلة الدرجات المشاهدة أنتجت أخطاء معادلة وبواقي معيارية اقل من معادلة الدرجات الحقيقية.

كما أجرى (Yuki, 2008) دراسة هدفت إلى مقارنة الطرق المعلمية واللامعلمية في نظرية استجابة الفقرة في معادلة الدرجات باستخدام تصميم المجموعات غير المتكافئة ذات الجذع المشترك. تم استخدام أربع طرق معلميه لمعادلة الدرجات في نظرية استجابة الفقرة: الدرجات الحقيقية والمشاهدة المعتمدة على تقدير معالم الفقرات المشتركة والمنفصلة، بالإضافة إلى استخدام أربع طرق لامعلمية وهي: معادلة الدرجات الحقيقية والمشاهدة بالاعتماد على توزيع كيرنيل. وحتى تتم المقارنة في الأداء بين الطرق المستخدمة تم الاعتماد على نوعين من البيانات: تم توليد بيانات بالاعتماد على خمسة عوامل: عدد الفقرات، نسبة الفقرات التي تنتهك افتراضات النموذج المعلمي، عدد المفحوصين، توزيعات القدرة لمجموعات المفحوصين، وأخيرا فيما إذا كانت العلامات لها معلم ذات خط تقاربي منخفض أم لا، كذلك تم تطبيق بيانات حقيقية لفحص سلوك طرق المعادلة في المواقف الحقيقية. أشارت نتائج البيانات المولدة والبيانات الحقيقية انه عندما تتحقق افتراضات النموذج، تكون الطرق المعلمية أكثر دقة من الطرق اللامعلمية، كذلك الطرق اللامعلمية تصبح أكثر دقة من الطرق المعلمية عندما تنتهك افتراضات النموذج، ومن النتائج الأخرى التي توصلت إليها الدراسة أن الخط ألتقاربي الأسفل يؤثر في نتائج المعادلة، بالإضافة إلى أن الطرق اللامعلمية هي أكثر دقة مع الدرجات المشاهدة من الدرجات الحقيقية في معادلة الدرجات.

وأجرت (Amanda, 2008) دراسة هدفت إلى مقارنة طرق المعادلة في النظرية الكلاسيكية ونظرية استجابة الفقرة. حيث تم تطبيق سبع طرق مختلفة لمعادلة هذين النوعين من الدرجات وهي: ثلاث طرق كلاسيكية (طريقة توكر الخطية، الطريقة غير الممهدة وطريقة تشين المئينية)، وأربع طرق من طرق معادلة الدرجات في نظرية استجابة الفقرة (النموذج الأحادي، النموذج الثنائي، النموذج الثلاثي، ونماذج الاختيار من متعدد). تم مقارنة الطرق السبعة المستخدمة في هذه الدراسة باستخدام بيانات حقيقية تم جمعها من تطبيق اختبار SAT وبيانات مولدة. أشارت النتائج أن الطريقة التي أنتجت أقل قيمة للتحيز في البيانات الحقيقية والمولدة هي طريقة توكر الخطية.

من مجمل الدراسات السابقة يمكن القول إن هذه الدراسات تناولت معادلة الدرجات من جوانب متعددة، فمنها من تناول معادلة الدرجات باستخدام الاختبارات المشتركة الداخلية أو الخارجية مع التصاميم المختلفة لجمع المعلومات، وبعضها حاول التحقق من دقة معادلة طريقة من بين طرق المعادلة المختلفة، كما تطرقت بعض الدراسات إلى الحديث عن العوامل التي تؤثر في دقة معادلة الدرجات مثل: صعوبة الفقرات، حجم العينات، أو طبيعة البيانات المستخدمة، كما عالجت بعض الدراسات اثر زيادة كل من عدد فقرات الاختبار وحجم العينة على دقة معادلة الدرجات، وقد لاحظ الباحث أن بعض هذه الدراسات عانت من بعض جوانب القصور على الرغم من أن بعض هذه الجوانب تم الإشارة إليها كمحددات في هذه الدراسة مثل اقتصار عينتها على فئة معينة، إضافة إلى اقتصار بعض الدراسات على طريقة واحدة أو اثنتين من طرق معادلة درجات الاختبار المختلفة وخصوصا معادلة الدرجات باستخدام المئينات أو المعادلة الخطية.

ولعل هذه الدراسة تتميز عن الدراسات السابقة في كونها تناولت معادلة درجات الاختبارات متعددة الاستجابة باستخدام IRT والتي لم تبحث

سابقا -في حدود علم الباحث-بصورة كافية، كذلك تتميز هذه الدراسة بتناولها لمتغيرات عديدة مثل حجم العينة، صعوبة الفقرات؛ للوقوف على مدى فاعلية ودقة نظرية استجابة الفقرة في معادلة الدرجات. كما تتميز هذه الدراسة باعتمادها على نظرية استجابة الفقرة كمعيار للمقايسة والمفاضلة للحكم على دقة المعادلة، وذلك أن معظم الدراسات السابقة كانت تستخدم الطريقة المئينية للمفاضلة بين طرق المعادلة المختلفة.

مشكلة الدراسة وأسئلتها

عندما يتم استخدام وتطبيق الاختبار بشكل متكرر، فان فقرات الاختبار يمكن أن تصبح فقرات سهلة ومعروفة للمفحوصين المستقبليين؛ لذلك يتم إعداد نماذج متعددة للاختبارات في العديد من برامج الاختبارات؛ لمنع كشف الاختبار، وعلى الرغم من وجود النماذج المتعددة للاختبارات والتي تم بناؤها بالاعتماد على نفس الخصائص مثل نفس المحتوى، ونفس مستوى الصعوبة، إلا أن نماذج الاختبار لا تكون متكافئة بالضبط؛ ولهذا السبب فان بعض المفحوصين الذين يأخذون الاختبار الأسهل سيكون لهم أفضلية على أولئك الذين يأخذون الاختبار الأصعب. لذلك إذا تمت معادلة درجات الاختبارات المختلفة بطريقة علمية صحيحة، فان المفحوصين سوف يحصلون على نفس الدرجات بغض النظر عن أي اختبار يتم التقدم له.

من هنا، هدفت هذه الدراسة إلى معرفة أثر صعوبة الفقرة وحجم العينة في دقة معادلة درجات الاختبارات باستخدام نظرية استجابة الفقرة (IRT)، واختيرت هذه الطريقة كموضوع لهذه الدراسة كونها من الطرق الأقل شيوعا واستخداما في معادلة الاختبارات. وبعبارة أخرى فان الغرض الأساسي من الدراسة هو التحقق من دقة معادلة درجات الاختبارات ذات الجذع المشترك للمجموعات المتكافئة باستخدام طرق نظرية استجابة الفقرة في معادلة درجات الاختبارات تحت ظروف اختلاف صعوبة الفقرة وحجم العينة. ويمكن صياغة مشكلة الدراسة بالتساؤل الآتي: ما أثر طرق نظرية استجابة الفقرة في دقة معادلة درجات الاختبارات المتعددة الحدود؟

وبالتحديد فان هذه الدراسة تحاول الإجابة عن السؤالين الآتيين:

١. ما أثر صعوبة الاختبار وحجم العينة في تقدير قيم الخطأ المعياري للمعادلة (Standard Error of Equating-SEE) عند النقاط المختلفة على سلم الدرجات؟

٢. ما أثر صعوبة الاختبار وحجم العينة في تقدير قيم البواقي المعيارية للمعادلة (Root Mean Standard Error of Equating-RMSE) عند النقاط المختلفة على سلم الدرجات؟

أهمية الدراسة

تكتسب الدراسة أهميتها من خلال استخدامها تصميم فقرات الجذع المشترك في مقارنة دقة معادلة الاختبارات باستخدام نظرية استجابة الفقرة تحت ظروف: البيانات متعددة الحدود، الفقرات المشتركة، حجم العينة، وصعوبة الفقرة. ويمكن تلخيص أهمية الدراسة في الجوانب الآتية:

الإسهام في إلقاء الضوء على مدى دقة طرق نظرية استجابة الفقرة في معادلة درجات الاختبارات. تبرز أهمية الدراسة في أن معظم الدراسات السابقة كانت تتم باختيار طريقة المئينات أو طريقة المعادلة الخطية، بينما في هذه الدراسة تم استخدام طريقة نظرية استجابة الفقرة للمعادلة للوقوف على مقدار دقتها في معادلة درجات الاختبارات.

أهداف الدراسة

توفير معلومات وإرشادات تساعد المهتمين في بناء الاختبارات لانتقاء أفضل الطرق في معادلة الدرجات وتطبيقها بكل يسر وسهولة.

محددات الدراسة

- اقتصار هذه الدراسة على طريقة واحدة من طرق معادلة درجات الاختبار المختلفة وهي طريقة المعادلة باستخدام نظرية استجابة الفقرة.
- استخدام الخطأ المعياري ومتوسط مربع الأخطاء المعيارية فقط؛ للحكم على دقة المعادلة.
- اقتصار الدراسة على بيانات تجريبية مولدة باستخدام برمجية (Wingen2).

مصطلحات الدراسة

معادلة درجات الاختبارات: هو إجراء إحصائي يتم فيه تحويل سلم الدرجات على أحد الاختبارات إلى سلم الدرجات على الاختبار الآخر، بحيث يمكن معرفة درجة الفرد على أحد الاختبارات إذا علمنا درجته على الاختبار الآخر.

البيانات متعددة الحدود: وهي فقرات اختباريه تكون الاستجابة عليها متعددة (١-٢-٣-٤-٥) حيث يكلف المفحوص بتحديد الاستجابة التي يراها مناسبة.

الطريقة والإجراءات

هدفت هذه الدراسة إلى معرفة دقة طرق نظرية استجابة الفقرة (IRT) في معادلة درجات الاختبارات متعددة الاستجابة، وفيما يلي وصفا للمنهجية المتبعة في هذه الدراسة.

أولا: تصميم الدراسة ومتغيراتها

تكونت متغيرات الدراسة المستقلة من: حجم العينة وتكونت من ثلاثة مستويات: (٢٠٠، ٦٠٠، ١٠٠٠)، ومستويات الصعوبة، وله مستويان: (التشابه في متوسط معدل الصعوبة للاختبار والاختلاف في متوسط معدل الصعوبة). وفي ضوء متغيرات الدراسة وبمستوياتها، كان تصميم الدراسة (٣ X ٢)، أما المتغيرات التابعة فقد تمثلت في معيار الحكم على دقة معادلة درجات الصور المختلفة للاختبار وهو الخطأ المعياري للمعادلة، والجذر التربيعي لمتوسط مجموع مربعات الانحرافات الخطأ ويحسب الخطأ المعياري للمعادلة بالمعادلة الآتية:

$$SEE = \sqrt{\left[\frac{1}{n-1}\sum_{1}^{n}(\hat{e}_x(y_k) - \bar{\hat{e}}_x(y_k)\right]^2}$$

حيث:

n = عدد مرات التكرار. yk = تمثل الدرجات على النموذج y. ex(yk) = هي الدرجة المعادلة على النموذج x من خلال النموذج yk . $\bar{\hat{e}}_x(y_k)$ = هي متوسط درجات المعادلة للدرجة yk عبر مرات التكرارات.

الجذر التربيعي لمتوسط مجموع مربعات الانحرافات الخطأ (Root Mean Squared Error RMSE) ويمكن حساب جذر متوسط مربع الأخطاء المعيارية (RMSE) من خلال المعادلة الآتية:

$$RMSE = \left[\frac{1}{n_s}\sum\frac{1}{n_e}\sum_{j}(y_j - E_j)^2\right]^{\frac{1}{2}}$$

حيث أن: K = عدد المفحوصين. j= رمز المفحوص. ns = عدد العينات والمتمثلة في عدد مرات توليد البيانات. ne = عدد المفحوصين في العينة الواحدة. yj = الدرجة على الصورة الأولى للمفحوص j في ضوء درجته على الصورة الثانية. Ej = الدرجة الحقيقية المتوقعة للمفحوص على الصورة الأولى.

وحسب مربع انحراف الدرجة المعدلة للمفحوص على الصورة الأولى عن درجته الحقيقية المتوقعة، ومن ثم تم إيجاد مجموع مربعات انحرافات كل أفراد العينة في كل مرة، وحسب متوسط مجموع مربعات الانحرافات في كل عينة، ومن ثم متوسط عدد مرات التوليد وهي (٢٠٠)، وبعد ذلك حسب الجذر التربيعي.

إجراءات جمع البيانات

استخدمت هذه الدراسة نظرية استجابة الفقرة (IRT) لتوليد البيانات باستخدام برنامج (WINGEN-2)، فمن خلال إعطاء معالم الفقرات وخصائص الأشخاص، فان البرمجية تستطيع توليد مجموعات مختلفة من العينات تأخذ

اختبارات مختلفة أو مجموعات مختلفة تأخذ نفس الاختبار.

توليد بيانات الاستجابة على الفقرات ومعادلة الدرجات

تم توليد البيانات باستخدام برنامج (Wingen 2) لتوليد عينات عشوائية تمثل مجتمع التوزيعات ولإيجاد معيار ومعادلة الدرجات وتقييم دقة المعادلة وفق الخطوات الآتية:

تحديد المتغيرات التي تتضمنها الدراسة، مثل حجم العينة وصعوبة الاختبار والتوزيع الذي تقع تحته معالم الفقرات، وغيرها من المتغيرات المراد دراستها.

توليد البيانات وفقاً لنموذج نظرية استجابة الفقرة المطلوب.

تقدير المعالم مثل: معلمة الصعوبة، التمييز، التخمين، ومعلمة القدرة وذلك بالاعتماد على استجابات الفقرة التي تم توليدها.

تكرار الخطوة السابقة (N) من المرات لكل خلية في التصميم البحثي.

مقارنة النتائج التي تقيس تأثير المتغيرات المراد دراستها.

تحليل النتائج الخاصة بكل خلية باستخدام الطرق الوصفية والاستدلالية (Harwell, 1997)

رابعا: خطوات توليد البيانات المستخدمة في هذه الدراسة

اعتمد ت هده الدراسة على توليد البيانات باستخدام المحاكاة حيث تم استخدام نفس إجراءات المعادلة في المستويات الثلاث لحجم العينة لكل متغير (طول الاختبار، حجم العينة)، وتتلخص إجراءات عملية توليد البيانات كما يأتي:

١. للنموذج X، يتم حساب معاملات الصعوبة للمجتمع باستخدام توزيع درجات المجتمع.
٢. في كل عملية توليد للبيانات يتم توليد نموذج يتكون من (١.٥) بشكل عشوائي لـ X.
٣. بعد ذلك يتم إعادة التوليد للبيانات حتى يتم الحصول على البيانات المولدة للاختبار X.
٤. إعادة توليد البيانات للنموذج Y.
٥. بعد الحصول على البيانات لكلا النموذجين يتم تطبيق إجراءات معادلة الدرجات المستخدمة في هذه الدراسة.
٦. بعد ذلك يتم إعادة الخطوات السابقة ٥٠٠ مرة لحساب الخطأ المعياري في المعادلة (SEE) وجذر متوسط مربع الأخطاء في المعادلة (RMSE) لكل درجة خلال درجات الاختبار.

البرمجيات المستخدمة

برنامج (Wingen2)

استخدم هذا البرنامج في توليد الاستجابات الثنائية والمتعددة الأبعاد لعينات من الأفراد يولدها البرنامج عشوائياً كي تحاكي عينات المجتمع الأصلي من حيث التوزيعات والخصائص الإحصائية، ويتم من خلال هذا البرنامج إنتاج العديد من عينات الاستجابة تصل إلى (١٠٠,٠٠٠,٠٠٠) مفحوص و(١٠٠,٠٠٠,٠٠٠) فقرة ذات الأحجام المختلفة التي لها خصائص العينات المناظرة لها بالمجتمع الأصلي نفسها. وبالتالي يمكن من خلال هذه البرامج الحصول على عينات عديدة ذات أحجام كبيرة يصعب الحصول عليها من المجتمع الأصلي مما يوفر الوقت والجهد والمال على الباحثين (Hambleton & Han, 2007).

مقاييس المعادلة (إجراءات المعادلة)

لكل مستوى من طول الاختبار وتشابه مستويات الصعوبة في النماذج كان مقياس المعادلة الذي تم اعتماده هو باستخدام طرق المعادلة باستخدام الدرجات الملاحظة.

تحليل البيانات المولدة: التحليل الإحصائي

بالإضافة إلى الفقرات التي تم توليدها لنماذج الاختبار X وy فان هناك معلومات إضافية تم حسابها من البيانات الخام لكل عينة؛ وذلك من اجل استخدامها في تقدير وتقييم خطوات المعادلة. وقد تم تطبيق طريقة المعادلة باستخدام برمجية (Equating Recipes).

النتائج

هدفت هذه الدراسة إلى معرفة أثر صعوبة الفقرة وحجم العينة في دقة معادلة درجات الاختبارات باستخدام نظرية استجابة الفقرة (IRT)، وذلك في ظل ظروف تجريبية مختلفة، وفيما يأتي عرضا لنتائج الدراسة:

أولا: النتائج المتعلقة بالسؤال الأول: ما أثر صعوبة الاختبار وحجم العينة في الخطأ المعياري للمعادلة عند النقاط المختلفة على سلم الدرجات؟

ثانيا: التشابه في مستويات صعوبة النماذج:

تم توليد نموذجين للاختبار بحيث اشتمل كل اختبار على (٦٠) فقرة متشابهة في صعوبتها، ثم ولدت (٢٠٠) عينة عشوائية وبمستويات (٢٠٠-٦٠٠-١٠٠٠) لكل عينة من مجتمع التوزيع، طبقت طريقة (IRT) المستخدمة في هذه الدراسة لتقدير خطأ المعادلة وفيما يأتي توضيح لهذه النتائج:

يظهر الجدول (١) أن قيم الخطأ المعياري للمعادلة لطريقة (IRT) في معادلة درجات الاختبار قد تراوحت بين (٣.٤٢ -٣.٦٣)، حيث بلغت اقل قيمة للخطأ المعياري (٣.٤٢)، في حين بلغت أعلى قيمة (٣.٦٣)، كذلك يظهر الجدول

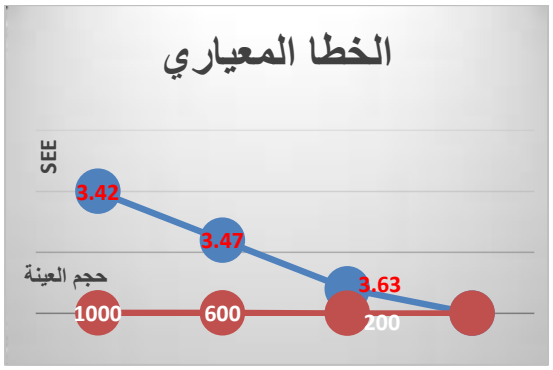
ان قيمة الخطأ المعياري للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تكونت العينة من (١٠٠٠) مفحوص، بلغت قيمة الخطأ المعياري (٣.٤٢)، وعندما بلغ حجم العينة (٦٠٠) مفحوص، بلغت قيمة الخطأ المعياري (٣.٤٧)، وعندما بلغ حجم العينة (٢٠٠) مفحوص، بلغت قيمة الخطأ المعياري (٣.٦٣). ويوضح الشكل رقم (١) قيم

جدول ١

الخطأ المعياري للمعادلة لطريقة (IRT) في معادلة درجات الاختبار المتشابه في مستوى صعوبة الفقرات

حجم العينة	٢٠٠	٦٠٠	١٠٠٠
الخطأ المعياري للمعادلة	٣,٣٦	٣,٤٧	٣،٤٢

الخطأ المعياري للمعادلة لطريقة (IRT) في معادلة درجات الاختبار الذي يتألف من (٦٠) فقره والمتشابه في مستوى صعوبة الفقرات على سلم الدرجات وعبر اختلاف حجم العينات.



شكل ١. الخطأ المعياري للمعادلة لطريقة (IRT) في معادلة درجات الاختبارات المتشابهة في مستوى صعوبة الفقرات

يلاحظ من الشكل السابق ان قيمة الخطأ المعياري للمعادلة تنخفض مع ازدياد حجم العينة.

ب-الاختلاف في مستويات الصعوبة:

تم بناء نموذجين للاختبار مكونين من (٦٠) فقرة ومختلفين في مستوى الصعوبة، ثم تم توليد (٢٠٠) عينة عشوائية ولكل مستوى من مستويات حجم العينة (٢٠٠-٦٠٠-١٠٠٠). طبقت طريقة (IRT) في معادلة درجات الاختبار المستخدمة في هذه الدراسة على هذه العينات العشوائية؛ لتقدير الخطأ المعياري للمعادلة، حيث تم التوصل الى النتائج الاتية:

جدول ٢.

الخطأ المعياري للمعادلة لطريقة (IRT) في معادلة درجات الاختبار المختلف في مستوى صعوبة الفقرة

حجم العينة	٢٠٠	٦٠٠	١٠٠٠
الخطأ المعياري للمعادلة	٣.٣٨	٠.٩٥	٠.٩٠

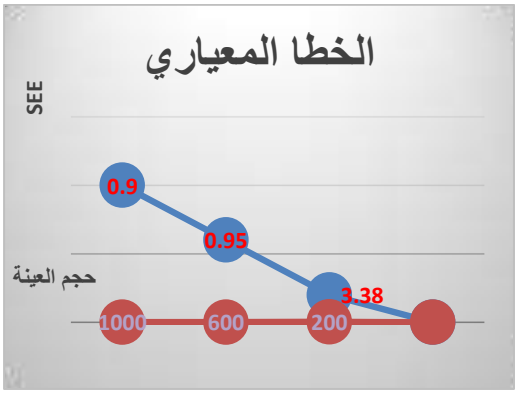
يظهرالجدول (٢) أن قيم الخطأ المعياري للمعادلة للطرق لطريقة (IRT) في معادلة درجات الاختبار الذي يتألف من (٦٠) فقره والمختلف في مستوى صعوبة الفقرات قد تراوحت بين (٠.٩٠ -٣.٣٨)، حيث بلغت أقل قيمة للخطأ المعياري (٠.٩٠)، في حين بلغت أعلى قيمة (٣.٣٨)، كذلك يظهر الجدول ان قيمة

جدول ٣.

قيم (RMSE) لطريقة (IRT) في معادلة درجات الاختبار الذي يتألف من (٦٠) فقره والمتشابه في مستوى صعوبة الفقرات

حجم العينة	٢٠٠	٦٠٠	١٠٠٠
قيم (RMSE)	٤.٩٩	٤.١٢	٤.٠١

الخطأ المعياري للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تكونت العينة من (١٠٠٠) مفحوص، بلغت قيمة الخطأ المعياري (٠.٩٠)، وعندما بلغ حجم العينة (٦٠٠) مفحوص، بلغت قيمة الخطأ المعياري (٠.٩٥)، وعندما بلغ حجم العينة (٢٠٠) مفحوص، بلغت قيمة الخطأ المعياري (٠.٩٠). ويوضح الشكل (٢) قيم الخطأ المعياري للمعادلة لطريقة (IRT) في معادلة درجات الاختبار الذي يتألف من (٦٠) فقره والمختلف في مستوى صعوبة الفقرات:



شكل ٢. الخطأ المعياري للمعادلة لطريقة (IRT) في معادلة درجات الاختبار الذي يتألف من (٦٠) فقره والمختلف في مستوى صعوبة الفقرات

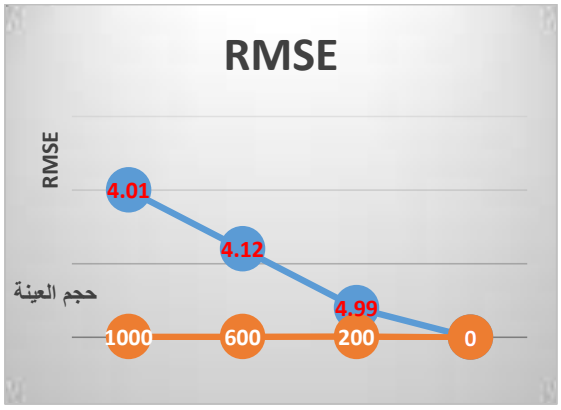
يوضح الشكل السابق أن قيمة الخطأ المعياري للمعادلة تنخفض مع ازدياد حجم العينة، وان المنحنيات تميل إلى النزول وتقترب من الصفر عندما يرتفع حجم العينة، وهذا يشير الى ان ارتفاع حجم العينة يقلل من قيمة الخطأ المعياري للمعادلة (SEE).

النتائج المتعلقة بالسؤال الثاني: ما أثر صعوبة الاختبار وحجم العينة في البواقي المعيارية للمعادلة عند النقاط المختلفة على سلم الدرجات؟

أولا: التشابه في مستويات صعوبة النماذج:

تم توليد نموذجين للاختبار بحيث اشتمل كل اختبار على (٦٠) فقرة متشابهة في صعوبتها، ثم ولدت (٢٠٠) عينة عشوائية وبمستويات (٢٠٠-٦٠٠-١٠٠٠) لكل عينة من مجتمع التوزيع، طبقت طريقة (IRT) المستخدمة في هذه الدراسة لتقدير جذر متوسطات مربعات الفروق وفيما يأتي توضيح لهذه النتائج:

يظهر الجدول (٣) أن قيم (RMSE) لطريقة (IRT) في معادلة درجات الاختبار الذي يتألف من (٦٠) فقره والمتشابه في مستوى صعوبة الفقرات قد تراوحت بين (٤.٠١ -٤.٩٩)، حيث بلغت اقل قيمة ل (RMSE)(٤.٠١)، في حين بلغت أعلى قيمة (٤.٩٩)، كذلك يظهر الجدول ان قيم (RMSE) للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تألفت العينة من (١٠٠٠) مفحوص، بلغت قيمة (RMSE)(٤.٠١)، وعندما بلغ حجم العينة (٦٠٠) مفحوص، بلغت قيمة (RMSE)(٤.١٢)، وعندما بلغ حجم العينة (٢٠٠) مفحوص، بلغت قيمة (RMSE) (٤.٩٩). ويوضح الشكل رقم (٣) قيم (RMSE) لطريقة (IRT) في معادلة درجات الاختبار الذي يتألف من (٦٠) فقره والمتشابه في مستوى صعوبة الفقرات



شكل ٣. قيم RMSE لطريقة (IRT) في معادلة درجات الاختبار الذي يتألف من (٦٠) فقره والمتشابه في مستوى صعوبة الفقرات

يظهر الشكل السابق ان قيم (RMSE) تتأثر بحجم العينة، فكلما كبر حجم العينة قلت قيمة (RMSE) وإذا نقص حجم العينة ارتفعت قيم (RMSE).

ب-الاختلاف في مستويات الصعوبة:

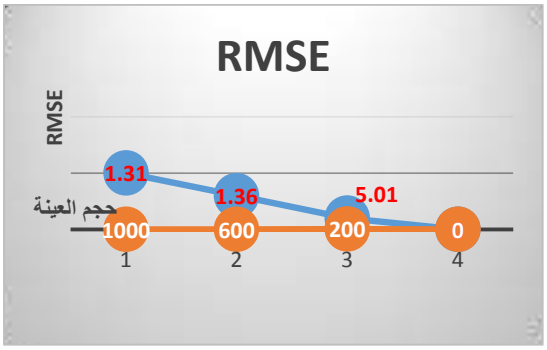
تم بناء نموذجين للاختبار مكونين من (٦٠) فقرة ومختلفين في مستوى الصعوبة، ثم تم توليد (٢٠٠) عينة عشوائية ولكل مستوى من مستويات حجم العينة(٢٠٠-٦٠٠-١٠٠٠). طبقت طريقة (IRT) في معادلة درجات الاختبار المستخدمة في هذه الدراسة على هذه العينات العشوائية؛ لتقدير جذر متوسط مربعات الفروق في المعادلة، حيث تم التوصل الى النتائج الاتية:

يوضح الجدول رقم (٤) قيم (RMSE) لطريقة (IRT) في معادلة درجات الاختبار الذي يتألف من (٦٠) فقره والمتشابه في مستوى صعوبة الفقرات:

جدول ٤.
قيم (RMSE) لطريقة (IRT) في معادلة درجات الاختبار الذي يتألف من (٦٠) فقره والمختلف في مستوى صعوبة الفقرات

حجم العينة	٢٠٠	٦٠٠	١٠٠٠
قيم (RMSE)	٥.٠١	١.٣٦	١.٣١

يظهر الجدول (٤) أن قيم (RMSE) لطريقة (IRT) في معادلة درجات الاختبار الذي يتألف من (٦٠) فقره والمختلف في مستوى صعوبة الفقرات قد تراوحت بين (١.٣١ -٥.٠١)، حيث بلغت اقل قيمة ل (RMSE)(١.٣١)، في حين بلغت أعلى قيمة (٥.٠١)، كذلك يظهر الجدول ان قيم (RMSE) للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تألفت العينة من (١٠٠٠) مفحوص، بلغت قيمة (RMSE)(١.٣١)، وعندما بلغ حجم العينة (٦٠٠) مفحوص، بلغت قيمة (RMSE) (١.٣٦)، وعندما بلغ حجم العينة (٢٠٠) مفحوص، بلغت قيمة (RMSE)(٥.٠١). ويوضح الشكل (٤) قيم (RMSE)لطريقة (IRT) في معادلة درجات الاختبار الذي يتألف من (٦٠) فقره والمختلف في مستوى صعوبة الفقرات:



شكل ٤. قيم (RMSE) لطريقة (IRT) في معادلة درجات الاختبار الذي يتألف من (٦٠) فقره والمختلف في مستوى صعوبة الفقرات

يوضح الشكل السابق ان قيم RMSE تنخفض مع ازدياد حجم العينة، كما ان المنحنيات تميل الى النزول وتقترب من الصفر عندما ترتفع حجم العينة، وهذا يشير الى ان ارتفاع حجم العينة يقلل من قيم RMSE.

مناقشة النتائج

مناقشة النتائج المتعلقة بالسؤال الأول: ما اثر صعوبة الاختبار وحجم العينة في الخطأ المعياري للمعادلة عند النقاط المختلفة على سلم الدرجات؟

أظهرت النتائج أن قيم الخطأ المعياري للمعادلة لطريقة (IRT) في معادلة درجات الاختبار المتشابه في مستويات الصعوبة قد تراوحت بين (٣.٦٣ -٣.٤٢)، حيث بلغت أقل قيمة للخطأ المعياري (٣.٤٢)، في حين بلغت أعلى قيمة (٣.٦٣)، كذلك يظهر الجدول أن قيمة الخطأ المعياري للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تكونت العينة من (١٠٠٠) مفحوص، بلغت قيمة الخطأ المعياري (٣.٤٢)، وعندما بلغ حجم العينة (٦٠٠) مفحوص، بلغت قيمة الخطأ المعياري (٣.٤٧)، وعندما بلغ حجم العينة (٢٠٠) مفحوص، بلغت قيمة الخطأ المعياري (٣.٦٣).

كما أظهرت النتائج أن قيم الخطأ المعياري للمعادلة لطريقة (IRT) في معادلة درجات الاختبار الذي يتألف من (٦٠) فقره والمختلف في مستوى صعوبة الفقرات قد تراوحت بين (٣.٣٨ -٠.٩٠)، حيث بلغت أقل قيمة للخطأ المعياري (٠.٩٠)، في حين بلغت أعلى قيمة

(٣.٣٨)، كذلك يظهر الجدول أن قيمة الخطأ المعياري للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تكونت العينة من (١٠٠٠) مفحوص، بلغت قيمة الخطأ المعياري (٠.٩٠)، وعندما بلغ حجم العينة (٦٠٠) مفحوص، بلغت قيمة الخطأ المعياري (٠.٩٥)، وعندما بلغ حجم العينة (٢٠٠) مفحوص، بلغت قيمة الخطأ المعياري (٠.٩٠).

إن قيمة الخطأ المعياري يتناسب عكسيا مع حجم العينة، حيث أنه بزداد بزيادة حجم العينة، وتقل قيمته بنقصان حجم العينة، وهذا أمر طبيعي؛ إذ أن كمية المعلومات عند أي مستوى من مستويات القدرة تتناسب عكسيا مع الخطأ المعياري، فعند زيادة حجم العينة يقترب متوسط معلمة التخمين من الصفر وهذا يؤدي إلى التقليل من قيمة الخطأ المعياري، حيث أن قيمة الخطأ المعياري تقل كلما قلت قيمة التخمين، وقد يكون سبب ذلك أنه عند توليد البيانات باستخدام النموذج الثلاثي يتم أخذ معلمة التخمين بعين الاعتبار، مما يقلل من أثر التخمين وبالتالي التقليل من قيمة الخطأ المعياري والبواقي المعيارية.

لقد اتفقت نتائج هذه الدراسة مع نتائج دراسة (Mao, 2006)، والتي توصلت الى أن دقة تقدير الخطأ المعياري في المعادلة، والبواقي المعيارية كان أفضل في العينات ذات الحجم الكبير، وهذا ما توصلت اليه هذه الدراسة.

مناقشة النتائج المتعلقة بالسؤال الثاني: ما أثر صعوبة الاختبار وحجم العينة في البواقي المعيارية للمعادلة عند النقاط المختلفة على سلم الدرجات؟

لقد أظهرت النتائج أن قيم (RMSE) لطريقة (IRT) في معادلة درجات الاختبار الذي يتألف من (٦٠) فقره والمتشابه في مستوى صعوبة الفقرات قد تراوحت بين (٤.٩٩ -٤.٠١)، حيث بلغت اقل قيمة ل (RMSE) (٤.٠١)، في حين بلغت أعلى قيمة (٤.٩٩)، كذلك يظهر الجدول ان قيم (RMSE) للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تألفت العينة من (١٠٠٠) مفحوص، بلغت قيمة (RMSE) (٤.٠١)، وعندما بلغ حجم العينة (٦٠٠) مفحوص، بلغت قيمة (RMSE) (٤.١٢)، وعندما بلغ حجم العينة (٢٠٠) مفحوص، بلغت قيمة (RMSE) (٤.٩٩). كما أظهرت النتائج أن قيم (RMSE) لطريقة (IRT) في معادلة درجات الاختبار الذي يتألف من (٦٠) فقره والمختلف في مستوى صعوبة الفقرات قد تراوحت بين (٥.٠١ -١.٣١)، حيث بلغت اقل قيمة ل (RMSE) (١.٣١)، في حين بلغت أعلى قيمة (٥.٠١)، كذلك يظهر الجدول ان قيم (RMSE) للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تألفت العينة من (١٠٠٠) مفحوص، بلغت قيمة (RMSE) (١.٣١)، وعندما بلغ حجم العينة (٦٠٠) مفحوص، بلغت قيمة (RMSE) (١.٣٦)، وعندما بلغ حجم العينة (٢٠٠) مفحوص، بلغت قيمة (RMSE) (٥.٠١).

وعند مقارنة نتائج طريقة (ITR) في معادلة درجات الاختبار من خلال حجم العينة أظهرت نتائج الدراسة وتحت ظرف اختلاف حجم العينة، ان نتائج طريقة المعادلة باستخدام (IRT) قد أظهرت قيما منخفضة للخطأ المعياري وعبر الأحجام المختلفة للعينات. كما أظهرت النتائج أن حجم العينة يؤثر في قيمة الخطأ المعياري للمعادلة، فكلما زادت حجم العينة قلت قيمة الخطأ المعياري للمعادلة، وإذا قلت حجم العينة زادت قيمة الخطأ المعياري للمعادلة. ان الاختلاف في قيمة الخطأ المعياري للمعادلة باستخدام طريقة (IRT) يقل تدريجيا عندما يرتفع حجم العينة، إن مستوى صعوبة نماذج الاختبار لم يؤثر على حجم الخطأ المعياري للمعادلة بشكل كبير وخصوصا عند ارتفاع حجم العينة، كما اظهرت النتائج تقارب نتائج (RMSE) مع نتائج (SEE) عندما ترتفع حجم العينة. حيث تتأثر نتائج قيم (RMSE) وبدرجة كبيرة بحجم العينة، فالعينات الكبيرة تنتج قيما صغيرة ل (RMSE)، والعينات الصغيرة تنتج قيما كبيرة لقيم (RMSE).

وعند مقارنة نتائج طريقة (ITR) في معادلة درجات الاختبار من خلال حجم العينة وصعوبة الفقرات، فإن النتائج أظهرت قيما منخفضة للخطأ المعياري وعبر الاحجام المختلفة للعينات، فقيم الخطأ المعياري للاختبار الذي

يتكون من (٦٠) فقرة تنخفض عندما تختلف النماذج في مستوى صعوبتها، فقد بلغت قيمة الخطأ المعياري للعينة التي تكونت من (١٠٠٠) مفحوص والمتشابهة في مستوى الصعوبة (٣.٤٢)، في حين فقد بلغت قيمة الخطأ المعياري لنفس العينة التي تكونت من (١٠٠٠) مفحوص والمختلفة في مستوى الصعوبة(٠.٩٠)، كما بلغت قيمة الخطأ المعياري للعينة التي تكونت من (٦٠٠) مفحوص والمتشابهة في مستوى الصعوبة (٣.٤٧)، في حين بلغت قيمة الخطأ المعياري لنفس العينة التي تكونت من (٦٠٠) مفحوص والمختلفة في مستوى الصعوبة(٠.٩٥)، وكذلك بلغت قيمة الخطأ المعياري للعينة التي تكونت من (٢٠٠) مفحوص والمتشابهة في مستوى الصعوبة (٣.٦٣)، في حين بلغت قيمة الخطأ المعياري لنفس العينة التي تكونت من (٢٠٠) مفحوص والمختلفة في مستوى الصعوبة(٣.٣٨)، وهذا يعطي دلالة واضحة بان النماذج المتشابهة في صعوبتها تميل قيم الخطأ المعياري الى الانخفاض عندما تتشابه مستويات الصعوبة فيها، والنماذج المختلفة في صعوبتها تميل قيم الخطأ المعياري الى الارتفاع عندما تختلف مستويات الصعوبة فيها

أما نتائج (RMSE) تحت الظروف المختلفة لحجم العينة والتشابه في صعوبة النماذج، فقد اظهرت ان طريقة المعادلة باستخدام (IRT) اظهرت قيما منخفضة لقيم (RMSE) وعبر الاحجام المختلفة للعينات. كما أظهرت النتائج ان قيم (RMSE) للاختبار الذي يتكون من (٦٠) فقرة تنخفض عندما تختلف النماذج في مستوى صعوبتها، فقد بلغت قيمة (RMSE) للعينة التي تكونت من (١٠٠٠) مفحوص والمتشابهة في مستوى الصعوبة (٤.٠١)، في حين فقد بلغت قيمة (RMSE) لنفس العينة التي تكونت من (١٠٠٠) مفحوص والمختلفة في مستوى الصعوبة(١.٣١)، كما بلغت قيمة (RMSE) للعينة التي تكونت من (٦٠٠) مفحوص والمتشابهة في مستوى الصعوبة (٤.١٢)، في حين بلغت قيمة (RMSE) لنفس العينة التي تكونت من (٦٠٠) مفحوص والمختلفة في مستوى الصعوبة(١.٣٦)، وكذلك بلغت قيمة (RMSE) للعينة التي تكونت من (٢٠٠) مفحوص والمتشابهة في مستوى الصعوبة (٤.٩٩)، في حين بلغت قيمة (RMSE) لنفس العينة التي تكونت من (٢٠٠) مفحوص والمختلفة في مستوى الصعوبة(٥.٠١)، وهذا يعطي دلالة واضحة بان النماذج المختلفة في صعوبتها تميل قيم (RMSE) الى الانخفاض عندما تختلف مستويات الصعوبة فيها، والنماذج المتشابهة في صعوبتها تميل قيم (RMSE) الى الارتفاع عندما تتشابه مستويات الصعوبة فيها.

ان قيم RMSE وSEE المستخدم هنا ليست قيما معيارية (Standardized)، لأن فقرات الاختبارات لم تقسم باستخدام تطابق الانحراف المعياري للعلامات المشاهدة؛ ولأن الانحراف المعياري للدرجات الصحيحة يرتفع مع زيادة صعوبة الاختبار، فانه يتوقع أن القيم المعيارية لقيم RMSE سوف تظهر ميلا قليلا إلى الارتفاع مع ازدياد صعوبة الاختبار. كما ان الاختبار الصعب يعني احتمالية قليلة عند نقاط الدرجة، وبالتالي مفحوصين قليلين لكل درجة للعينات المركبة، وبالتالي فإن العدد القليل من المفحوصين عند كل درجة يقلل من دقة المعادلة عند كل النقاط.

لقد اتفقت نتائج هذه الدراسة مع نتائج دراسة (Mao, 2006)، والتي توصلت الى أن دقة تقدير الخطأ المعياري في المعادلة، والبواقي المعيارية كان أفضل في العينات ذات الحجم الكبير، وهذا ما توصلت اليه هذه الدراسة.

الاستنتاجات

خلصت الدراسة الى الاستنتاجات الآتية:

١- تميل قيم الخطأ المعياري في طريقة IRT إلى الانخفاض مع ارتفاع حجم العينة وخصوصاً، كذلك تميل قيم RMSE فيها إلى الانخفاض مع ازدياد حجم العينة.

٢- أظهرت نتائج هذه الدراسة أن أحجام العينات الكبيرة يوثر على معايير دقة

المعادلة، فحجم العينات الكبير يؤدي إلى تقليل الخطأ المعياري SEE، وتقليل قيم RMSE.

٣- بصورة عامة فإن الاختلاف في الخطأ المعياري للمعادلة وقيم RMSE يصبح أقل عندما ترتفع حجم العينة.

٤- أظهرت النتائج ان الخطأ المعياري للمعادلة وقيم (RMSE) للاختبار الذي يتكون من (٦٠) فقرة تنخفض عندما تختلف النماذج في مستوى صعوبتها، وترتفع عندما تتشابه النماذج في مستوى صعوبتها.

التوصيات

- استخدم الباحث لتقييم إجراءات المعادلة معياري الخطأ المعياري وجذر متوسط مربعات الفروق، لذلك يوصي الباحث باستخدام معايير أخرى مثل الصدق التقاطعي ومعيار الاهمية النسبية للمعادلة.
- استخدم بيانات حقيقية للوقوف على مقدار دقة نظرية استجابة الفقرة في معادلة درجات الاختبارات.

المراجع

الدوسري، راشد (٢٠٠٤). **القياس والتقويم التربوي الحديث: مبادئ وتطبيقات وقضايا معاصرة.** دار الفكر، عمان.

الصمادي، اسماعيل. (٢٠٠٦). **فاعلية طرق تصحيح اختبار الصواب – الخطأ المتعدد وتأثيرها على دقة معادلة الاختبار باستخدام نماذج النظرية الحديثة للقياس**، رسالة دكتوراه غير منشوره، جامعة اليرموك، عمان، الأردن.

Amanda, A. (2008). *A comparison of classical test theory and item response theory methods for equating Number-right scored to formula scored assessments* .Unpublished Doctoral Dissertation, University of Kansas, USA.

Baker, F. & Al-Karni, A. (1991). A comparison of two procedures for Computing IRT equating coefficients. ***Journal of Education Measurement,*** 28, 147-162.

Cook, L.L. & Eignor, D.R. (1991).An NCME instructional module on IRT equating methods. ***Journal of Education*** *Measurement: Issue and Practice, 10*, 37-45.

Hambelton, R. & Swaminthan, H. (1985). *Item Response Theory: Principles and Applications.* Boston, Kluwer: Nijhoff Publishing.

Hambelton, R. Swaminathan, H. Rogers, H. (1991). ***Fu****ndamentals Of item response theory.* New York: Sage Publications.

Huigin, H. (2004). *Investigation of (IRT)-based equating methods in The presence of outliers.* Unpublished Doctoral Dissertation University of Alberta, Canada.

Kolen, M. (1981). Comparison of traditional and item Response Theory methods for equating tests**.** *Journal of Educational Measurement, 18* (1), 1-11.

Kolen, M. (1988).Traditional equating methodology. *Educational Measurement: Issues and Practice, 7*, 29-36.

Kolen, M. & Brennan, R. (2004). *Test equating, scaling, and linking*, (2nd Ed). New York: Springer- Verlag.

Lord, D. (1980). *Applications of item Response Theory to Practical Testing Problems.* Hillsdale. N.J: Erlbaum.

Mao, X. (2006). *An investigation of the accuracy of the estimates of Standard errors for the kernel equating functions.* Unpublished Doctoral Dissertation, University of Iowa, USA.

Peterson, N. Kolen, M. & Hoover, H. (1989). Scaling, Norming and Equating**.** *Educational measurement, Washington D.C: American Council on Education:* 241-262.

Robert R. (2007). *A comparison of item response Theory true score equating and item response theory- Based local equating.* Unpublished Doctoral Dissertation, University of Massachusetts, USA.

YoungWoo, C. (2007). *Comparison of bootstrap standard errors of equating using (IRT) and equipercentile methods with polytomous-scored items under the common-item Nonequivalent- groups design.* Unpublished Doctoral Dissertation, University of Iowa.USA.

Yuki, N. (2008). *Comparison of parametric and Nonparametric (IRT) equating methods under the common- item nonequivalent groups design.* Unpublished Doctoral Dissertation, University of Iowa.USA.