201
المجلد 17 العدد 3 سبتمبر 2016

أثر حجم العينة وطول الاختبار في دقة معادلة درجات الاختبارات متعددة الحدود باستخدام طريقة كيرنيل

د. يوسف عبدالعاطي المحروق
قسم بحوث المناهج
وزارة التربية والتعليم – مملكة البحرين
Almahrouk2005@yahoo.com

أثر حجم العينة وطول الاختبار في دقة معادلة درجات الاختبارات متعددة الحدود باستخدام طريقة كيرنيل

د. يوسف عبدالعاطي المحروق
قسم بحوث المناهج
وزارة التربية والتعليم – مملكة البحرين

الملخص

هدفت هذه الدراسة إلى معرفة أثر حجم العينة وطول الاختبار في دقة معادلة درجات الاختبارات متعددة الحدود في تصميم الفقرات المشتركة باستخدام طريقة كيرنيل، وذلك من خلال دراسة المتغيرات الآتية: حجم العينة: وقد استخدمت ثلاثة مستويات: ٢٠٠، ٦٠٠، ١٠٠٠ وهذه الأحجام للعينات تُعَدُّ مناسبة لطرق المعادلة باستخدام تصميم المجموعات العشوائية، طول الاختبار: استخدمت ثلاثة مستويات لطول الاختبار: ٣٠-٤٠-٦٠ فقرة، تم التوصل إلى نتائج دقة المعادلة في ظل استخدام المتغيرات السابقة منفردة ومجتمعة، الاستماع، قراءة صوتية للكلمات، القاموس عرض القاموس المفصل.

ولمقارنة دقة المعادلة، تم توليد بيانات تجريبية باستخدام برمجية (Wingen2)، وذلك باستخدام متغيرين: طول الاختبار، وحجم العينة، وتم استخدام نظرية استجابة الفقرة لتوليد استجابات المفحوصين على فقرات الاختبار، حيث تم معادلة درجات الاختبارات باستخدام الدرجات الملاحظة في نظرية استجابة الفقرة كمعيار رئيس للمعادلة. أشارت النتائج أن حجم العينات الكبير يقلل من الخطأ المعياري للمعادلة ويقلل كذلك من البواقي المعيارية. كما أظهرت النتائج أن طول الاختبار يوثر في الخطأ المعياري للمعادلة، فالاختبار الطويل يعطي قيمًا كبرى للخطأ المعياري للمعادلة والبواقي المعيارية.

الكلمات المفتاحية: معادلة الاختبارات، تصميم الفقرات المشتركة، الاختبارات متعددة الحدود.

* تاريخ تسلم البحث: ٢٠١٤/١٠/٢١م
* تاريخ قبوله للنشر: ٢٠١٥/٢/٢م

The Effects of Sample Size and Test Length on the Accuracy of Equating Polynomial Scores using Kernel Method

Dr. Yousef A. Almahrouk
Curriculum Research Department
Ministry of Education - Bahrain

Abstract

This study explored the effects of sample size, test length on final equating results of Kernel Method when using the nonequivalent groups anchor test (NEAT) design. The equating method was evaluated using an equating criterion (SEE, RMSE). Standard error of equating differences between the criterion scores and equated scores, and root mean square error of the differences (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 also residuals. The results indicated that test length affects the standard error as the tall test gives large values for the standard error, and the standard residuals.

Keywords: equating test, NEAT design, Polynomial Scores.

أثر حجم العينة وطول الاختبار في دقة معادلة درجات الاختبارات متعددة الحدود باستخدام طريقة كيرنيل

د. يوسف عبدالعاطي المحروق
قسم بحوث المناهج
وزارة التربية والتعليم – مملكة البحرين

المقدمة

سيطرت نظرية القياس النفسي التربوي الكلاسيكية على الفكر التربوي، وظل العاملون بالقياس النفسي يستخدمون مبادئ هذه النظرية وأسسها 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, & Van der Hambelton, 1982) منها: اختلاف خصائص فقرات الاختبار باختلاف عينات المفحوصين المستخدمين في معايرة الفقرات، إذ إن 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 أقل من مستوى قدرة أفراد المجتمع.

كذلك من جوانب القصور التي تعاني منها النظرية الكلاسيكية أن الاختبارات في النظرية الكلاسيكية لا تزودنا بتقدير دقيق للمفحوصين من 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 أجل الوصول إلى قياس أكثر موضوعية، حاول علماء القياس المعاصرون الاستفادة من التقدم التكنولوجي في التوصل إلى طرق سيكومترية 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 (Tow Parameters) احداهما تتعلق بقدرة

الفرد والأخرى تتعلق بالفقرة التي يختبر بها، وبالتالي فإن 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 بدرجة عالية من الدقة والثبات.

لقد بين هامبلتون وسواميناثان (Hambelton & Swaminathan, 1991) مزايا رئيسه لنظرية استجابة الفقرة منها: أن تقدير قدرة الفرد يكون مستقلاً عن عينة الفقرات التي تطبق عليه، أي إن تقديرات القدرة للأفراد متحررة من خصائص الفقرات المستخدمة في تقدير القدرة (Item free). ومع افتراض وجود عدد كبير من الأفراد، يكون تقدير معالم الفقرات مستقلاً عن عينة الأفراد التي استخدمت في تقدير هذه المعالم (Sample free).

إن نظرية الاستجابة للفقرة تقوم على افتراضات أساسية ذكرها هامبلتون وسواميناثيان (Hambelton & 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 اللوجستية الأحادية والثنائية والثلاثية والرباعية.

إن النجاح في استخدام نماذج نظرية الاستجابة للفقرة يقوم على مجموعة من المتطلبات منها: أحادية البعد للاختبار، ومطابقة فقرات الاختبار للنموذج المستخدم، حجم العينة المستخدمة، وتوفر 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، وبنوك الأسئلة، والاختبارات التكيفية، وبناء الاختبارات وتحليلها وتقنينها، وكذلك معادلة الدرجات على صور الاختبار المختلفة.

ولضمان أمن الاختبارات تقوم العديد من المؤسسات التربوية التي تعنى ببناء الاختبارات وتصميمها بإعداد نماذج متعددة للاختبارات، وذلك من خلال تمثيل فقرات مختلفة في مواقع مختلفة من الاختبارات لتصبح هذه الاختبارات أكثر عدالة لدرجة مقبولة.

إن النماذج المتعددة للاختبارات تختلف غالبًا في صعوبتها، وتختلف كذلك في المفحوصين الذين ستطبَّقُ عليهم هذه النماذج، وهذا يؤدي إلى وجود أفضلية أو عدم أفضلية لدى المفحوصين في هذه الاختبارات؛ لأنهم تقدموا لنماذج سهلة أو صعبة من الاختبار، وحتى يتم

تعديل هذه الاختلافات في صعوبة النماذج، تستخدم المعادلة بحيث إن الدرجات على نماذج الاختبار المختلفة يمكن أن تستخدم بشكل تبادلي. (Kolen & Brenann, 2004).

إذا تمت عملية معادلة درجات الاختبارات بنجاح، فإن المفحوصين سيكون لهم نفس الدرجات المتوقعة بغض النظر عن أي نموذج تقدموا له، فالمعادلة إذن ستساعد بعد ذلك في تحقيق عدالة الاختبار، وإزالة آثار التدريب، ومقارنة أداء المفحوصين عبر نماذج الاختبار المختلفة. وحتى يتم تطبيق المعادلة، هناك أربع خطوات مهمة تؤخذ بعين الاعتبار وهي: (Kolen & Brenann, 2004):

١- اختيار تصميم جمع المعلومات.

٢- اختيار واحدة أو أكثر من التعريفات الإجرائية للمعادلة.

٣- اختيار طريقة التقدير الإحصائية المناسبة.

٤- تقييم نتائج المعادلة.

وعند الحديث عن معادلة الصور المختلفة للاختبارات توجد ثلاث قضايا مترابطة ومتداخلة يجب أخذها بعين الاعتبار وهي: (kolen, 1981)

١- الأساس النظري الذي يتم بموجبه التناظر والمقابلة بين العلامات في الصور المختلفة للاختبار، فهناك نظريتان رئيستان يتم من خلالهما اشتقاق طريقة المعادلة وهما: النظرية التقليدية والنظرية الحديثة في القياس.

٢- الحاجة إلى تصميم لجمع البيانات المعادلة، فهناك عده أنواع من التصاميم لجمع البيانات حيث تتدخل عدة اعتبارات عملية في اختيار التصميم المناسب.

٣ - الطرق الإحصائية التي ستستخدم في تقدير العلامة المناظرة، مثل هل سيتم استخدام الطريقة الخطية أم الطريقة المئينية؟

يتطلب القيام بإجراءات المعادلة وجود مكونين رئيسين هما: تصميم المعادلة وطريقة المعادلة، ويتضمن تصميم المعادلة عددًا من التصاميم: تصميم المجموعة المتكافئة (Equivalent Group) وتسمى أيضا المجموعة العشوائية (Random Group)، وتصميم المجموعة المفردة (Single Group Design)، وتصميم المجموعات العشوائية المتوازنة (Counter-Balanced Design)، تصميم المجموعات غير المتكافئة ذات الاختبار المشترك (Nonequivalent Group with Anchor Test) (NEAT)، وتصميم المجموعات المتكافئة ذات الاختبار المشترك (Equivalent Group with Anchor Test) (EAT).

يصنف المختصون في القياس والتقويم التربوي والنفسي طرق معادلة الاختبارات

إلى نوعين: الطرق التي تعتمد على النظرية الكلاسيكية في الاختبارات (Classical Test Theory)(CTT) التي منها طريقة كيرنيل في معادلة الدرجات على النظرية الكلاسيكية في الاختبارات، حيث تشير هذه الطريقة إلى مجموعة جديدة من طرق المعادلة اللامعلمية التي تستخدم التوزيع الطبيعي لمعادلة الدرجات من نموذج إلى آخر، التي تم تطويرها من قبل باول هولاند و دورثي وتاير (Dorthy, Thayer & Holland, 1987)، ولاحقًا تم تحديثها من قبل ألينا فون ديفر (Alina Von Davier, 2004)، حيث تعتمد طريقة كيرنيل على مجموعة مرنة من طرق المعادلة الشبيهة بالمعادلة المئينية التي تنتمي إلى الطرق الكلاسيكية في معالجة الدرجات المشاهدة، ويمكن أن تستخدم طريقة كيرنيل مع مختلف تصاميم جمع المعلومات ومن ضمنها التصاميم التي تعتمد على وجود الجذع المشترك. أما النوع الثاني فيُصنَّفُ ضمن الطرق التي تعتمد على النظرية الحديثة في القياس التربوي (نظرية الاستجابة للفقرة) (Item Response Theory IRT) .

عمليا، توجد هناك خمس خطوات لتطبيق طريقة كيرنيل في معادلة الاختبارات، والطريقة التي يتم فيها تنفيذ كل خطوة تختلف بالاعتماد على نوع الجذع المشترك الذي يستخدم هل هو داخلي (Internal) أو خارجي (External). وبشكل عام، يتم تطبيق طريقة كيرنيل في معادلة الاختبارات من خلال نظام يتكون من خمس خطوات متدرجة ومتتابعة هي:

١- مطابقة النماذج الخطية المتعددة لتوزيعات درجات الخام.

٢- تقدير الدرجات المحتملة للمجتمع المستهدف.

٣- استخدام توزيع كيرنيل لعمل اتصال بين العلامات المنفصلة في التوزيع.

٤- تطبيق المعادلة المئينية باستخدام التوزيع المتصل في الخطوة الثالثة.

٥- حساب الخطأ المعياري للمعادلة.

لقد أجريت العديد من الدراسات في مجال معادلة الدرجات، فقد أجرى (Livingston, 1993) والمشار إليها في (Akour, 2006) دراسة هدفت إلى فحص دقة معادلة كيرنيل مع أحجام عينات صغيره باستخدام تصميم المجموعات غير المتكافئة ذات الاختبار المشترك، تم استخدام استجابات أكثر من (٩٣٠٠٠) مفحوص على اختبار مكون من (١٠٠) فقرة؛ لتغطية توزيعات المجتمع كاملة. تم تطبيق المعادلة المئينية غير المهدة على توزيعات المجتمع للحصول على معايير المعادلة، بعد ذلك تم سحب عينات عشوائية من المفحوصين لتطبيق معادلة كيرنيل عليها باستخدام ثلاثة قيم لمعلم (h). وقد تم تكرار هذه العملية (٥) مرات بأحجام عينات مختلفة (٢٥-٥٠-١٠٠-٢٠٠). وقد أشارت النتائج إلى أن

معادلة كيرنيل كانت أكثر دقة في معادلة الدرجات مقارنة مع النتائج التي نتجت عن معادلة تشين المئينية للتوزيعات المهدة، كذلك كانت معادلة كيرنيل أكثر دقة مقارنة مع طريقة معادلة تشين المئينية للتوزيعات غير المهدة.

وأجرى (Wang, 2005) دراسة هدفت إلى مقارنة الطريقة اللوغاريتمية الخطية المستمرة مع طريقتين في المعادلة: طريقة كيرنيل وطرق المعادلة المئينية غير المهدة، تم استخدام تصميم المجموعات العشوائية في جمع بيانات اختبار حقيقي لمقارنة طرق المعادلة الثلاث. وقد أشارت النتائج إلى أن طريقة المعادلة اللوغاريتمية الخطية المستمرة أنتجت توزيعًا مستمرًا أكثر تمهيدًا، كذلك عندما تمت مقارنة الطريقتين: معادلة كيرنيل اللوغاريتمية (المستمرة) مع طريقة المعادلة المئينية غير المهدة وباستخدام البيانات نفسها، أشارت النتائج أن الطرق الثلاث أظهرت نتائج المعادلة نفسهاحتى في مقياس الدرجات.

وفي دراسة أجراها (Akour, 2006) للمقارنة بين ثلاث طرق لمعادلة الاختبارات باستخدام طرق المعادلة المئينية وهي: طريقة ما قبل التمهيد، وطريقة المعادلة الخطية المتعددة الحدود لما قبل التمهيد، وطريقة كوبيك سبلين بعد التمهيد، وبين طريقتين مختلفتين من طرق كيرنيل في المعادلة: (طريقة) كيرنيل، وطريقة المعادلة الخطية المستمرة (CLL). قام الباحث بجمع البيانات باستخدام تصميم المجموعات المتكافئة لتقدير معالم الفقرات باستخدام برمجية حاسوب للحصول على عينات عشوائية من مجتمع التوزيعات ومن ثم تطبيق إجراءات المعادلة على هذه العينات، حيث اعتمدت الدراسة على ثلاثة متغيرات رئيسة هي: حجم العينة ومستوى الصعوبة وطول الاختبار. طبق الباحث تقنيات نظرية IRT لتمهيد التوزيعات الملاحظة؛ وذلك ليلائم توزيعات المجتمع المستهدف، بعد ذلك تم تطبيق المعادلة المئينية على التوزيعات المهدة للحصول على علاقات المعادلة في المجتمع. وقد أشارت نتائج الدراسة إلى أن جميع طرق المعادلة كانت أكثر فاعلية من طرق ما قبل التمهيد.

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 معيارًا للمعادلة. وقد أشارت النتائج أن دقة تقدير الخطأ المعياري في المعادلة (SEE) كان أفضل في العينات ذات الحجم الكبير وذات إلى (bandwidth) الكبير، كذلك أشارت النتائج إلى أن الدرجات العالية من التمهيد تميل إلى إنتاج حجم أكبر من الأخطاء المعيارية للمعادلة مقارنة مع الدرجات المنخفضة من التمهيد.

وفي دراسة قام بها (Qu, 2007) لمعرفة أثر البيانات المركبة من مجموعتين تحت تأثير

اختلاف حجم العينات في دقة المعادلة. استخدم الباحث طريقتين في معادلة الدرجات وهي: طريقة كيرنيل في المعادلة التي تعتمد على المعادلة المئينية والطريقة التقليدية في المعادلة. طبق الباحث نوعين من البيانات: بيانات حقيقية، وبيانات مولدة باستخدام برامج حاسوبية، ثم قام الباحث بإيجاد كل من الخطأ المعياري للمعادلة (SEE)، والبواقي المعيارية، وتحيز المعادلة، والخطأ المعياري لاختلاف المعادلة ((SEED)) لكل طريقة من طرق المعادلة المستخدمة. وقد أشارت النتائج إلى أن طريقة المعادلة باستخدام البيانات المركبة من مجموعتين تكون أكثر حساسية للظروف المختلفة من البيانات غير المركبة.

أما (Amanda, 2008) فقد أجرت دراسة هدفت من خلالها 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، والطريقة غير المهدة والطريقة المئينية)، وأربع طرق من طرق معادلة الدرجات في نظرية استجابة الفقرة (النموذج الأحادي، والنموذج الثنائي، والنموذج الثلاثي، ونماذج الاختيار من متعدد).

تمت مقارنة الطرق السبعة المستخدمة في هذه الدراسة باستخدام بيانات حقيقية تم جمعها من خلال تطبيق اختبار SAT، وبيانات مولدة باستخدام الحاسوب. أشارت النتائج إلى 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 ما 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، وتطرقت بعض الدراسات إلى الحديث عن العوامل التي تؤثر في دقة معادلة الدرجات مثل تلك المتعلقة بصعوبة الفقرات، أو حجم العينات، أو طبيعة البيانات المستخدمة، كما عالجت بعض الدراسات أثر زيادة كل من عدد فقرات الاختبار وحجم العينة في دقة معادلة الدرجات.

وقد لاحظ الباحث أن بعض الدراسات عانت من بعض جوانب القصور – على الرغم من أن بعض هذه الجوانب تمت الإشارة إليها كمحددات في هذه الدراسة– مثل: اقتصار عينتها على فئة معينة من الفئات المختلفة لمجتمع الدراسة، إضافة إلى اقتصار بعض الدراسات على طريقة واحدة أو اثنتين من طرق معادلة درجات الاختبار المختلفة. كذلك لاحظ الباحث أيضًا أن الدراسات العربية – رغم محدودية عددها – لم تتطرق الى طريقة كيرنيل في معادلة درجات

الاختبارات - في حدود علم الباحث- حيث ركزت معظمها على معادلة الدرجات باستخدام المئينات أو طرق نظرية استجابة الفقرة.

ولعل هذه الدراسة تتميز عن الدراسات السابقة في كونها تناولت طريقة جديدة من طرق المعادلة لم يتم دراستها في الأدب التربوي في الوطن العربي - في حدود علم الباحث - كذلك تتميز هذه الدراسة بتناولها لمتغيرات عديدة مثل حجم العينة، وعدد فقرات الاختبار لمعرفة أثر هذه المتغيرات في دقة معادلة درجات الاختبارات.

مشكلة الدراسة

عندما يتم استخدام الاختبار وتطبيقه بشكل متكرر، فإن 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 التي تم بناؤها بالاعتماد على الخصائص نفسها مثل المحتوى، ومستوى الصعوبة نفسه، إلا أن نماذج الاختبار لا تكون متكافئة بالضبط؛ ولهذا السبب فإن 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 الدرجات نفسها بغض النظر عن أي اختبار يتم التقدم له.

من هنا، هدفت هذه الدراسة إلى معرفة أثر حجم العينة وطول الاختبار في دقة معادلة درجات الاختبارات متعددة الحدود في تصميم الفقرات المشتركة باستخدام طريقة كيرنيل واختيرت هذه الطريقة كموضوع لهذه الدراسة كونها من الطرق الأقل شيوعًا واستخدامًا في معادلة الاختبارات في الوطن العربي.

وبعبارة أخرى فإن الغرض الأساس من الدراسة هو التحقق من دقة معادلة درجات الاختبارات ذات الجذع المشترك للمجموعات المتكافئة باستخدام طريقة كيرنيل في معادلة درجات الاختبارات متعددة الحدود تحت ظروف اختلاف حجم العينة وطول الاختبار.

ويمكن صياغة مشكلة الدراسة بالتساؤل الآتي:

ما أثر حجم العينة وطول الاختبار في دقة معادلة درجات الاختبارات متعددة الحدود في تصميم الفقرات المشتركة باستخدام طريقة كيرنيل؟

أسئلة الدراسة

وبالتحديد فإن هذه الدراسة تحاول الإجابة عن السؤالين الآتيين:

١- ما أثر حجم العينة وطول الاختبار في الخطأ المعياري للمعادلة (Standard Error of Equating-SEE) عند النقاط المختلفة على سلم الدرجات؟

٢- ما أثر حجم العينة وطول الاختبار في البواقي المعيارية للمعادلة (Root Mean Standard Error of Equating-RMSE) عند النقاط المختلفة على سلم الدرجات؟

أهميَّة الدراسة

تكتسب الدراسة أهميتها من خلال استخدامها تصميم فقرات الجذع المشترك في مقارنة دقة معادلة الاختبارات باستخدام طريقة كيرنيل تحت ظروف: البيانات متعددة الحدود، والفقرات المشتركة، وحجم العينة، وطول الاختبار. ويمكن تلخيص أهمية الدراسة في الجوانب الآتية:

١- الإسهام في إلقاء الضوء على مدى دقة طريقة كيرنيل في معادلة درجات الاختبارات، بحيث تسهم في إعطاء صورة واضحة للمتخصصين في مجال بناء الاختبارات لاختيار طريقة المعادلة التي تتواءم مع طبيعة الاختبار الذي يتم بناؤه؛ وذلك للحصول على نتائج تمتاز بالشمولية والدقة.

٢- تبرز أهمية الدراسة في أن معظم الدراسات السابقة كانت تتم باختيار طريقة المئينات أو نظرية استجابة الفقرة او طريقة المعادلة الخطية، بينما في هذه الدراسة سيتم استخدام طريقة كيرنيل للمعادلة للوقوف على مقدار دقتها في معادلة درجات الاختبارات.

٣- التحقق من النتائج المحتملة بالإجابة عن الأسئلة التي طرحتها الدراسة، أي كيف يمكن استخدام طريقة كيرنيل لمعادلة درجات الاختبارات متعددة الحدود.

أهداف الدراسة

١- توضيح طرق جديدة لمصممي الاختبارات ومحلليها في وزارات التربية والتعليم والجامعات المهتمين بموضوع إيجاد صور متكافئة للاختبارات.

٢- 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 كيرنيل.

- 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 في أحد الاختبارات إذا علمنا درجته في الاختبار الآخر

تصميم الفقرات المشتركة: هو اختبار مكون من مجموعة من الفقرات المشتركة بين اختبارات الصفوف المختلفة.

دقة المعادلة: هو أسلوب إحصائي يستخدم للتأكد من مدى فاعلية المعادلة باستخدام اختبار الجذع المشترك.

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

الطريقة والإجراءات

اعتمدت هذه الدراسة على توليد البيانات باستخدام طريقة المونت كارلو للإجابة عن أسئلة الدراسة، لذلك تضمن هذا الجزء من البحث المواضيع الرئيسة الآتية: جمع البيانات واشتملت على (توليد البيانات، وإجراءات توليد البيانات، والعوامل التي أخذ بها لتوليد البيانات، والخطوات العامة المستخدمة في توليد البيانات)، وإجراءات جمع البيانات، وتصميم الدراسة، ومتغيرات الدراسة، والبرمجيات المستخدمة، وأخيرا التحليل الإحصائي.

تصميم الدراسة

اعتمدت هذه الدراسة على اختبار الجذع المشترك، حيث تم جمع المعلومات بالاعتماد على تصميم المجموعات المتكافئة ذات الجذع المشترك، وهذا النوع من تصاميم جمع المعلومات يسمح لنا بمعادلة نماذج الاختبارات (X, Y) باستخدام فقرات الجذع المشترك، لذلك فان الفقرات التي تم توليدها تتميز بتعددية الاستجابة في اختبارات الجذع المشترك، بحيث تأخذ الإجابة الصحيحة رقم (١) والإجابة الخاطئة رقم (٠)، لذلك فإن الإجابات الصحيحة قورنت بالدرجات المعادلة في الاختبارات المستخدمة؛ لتحديد أثر حجم العينة وطول الاختبار في دقة معادلة درجات الاختبارات.

جمع البيانات:

أولا: توليد البيانات:

استخدمت طريقة المونت كارلو (Monte Carlo) الكمبيوتر كوسيلة مساعدة في توليد البيانات، وذلك لإعطاء الدليل للمشاكل التي لا يمكن حلها رياضيًا مثل: عندما يكون توزيع العينات غير معروف، أو عندما تكون الفرضية الصفرية غير صحيحة.

ثانيا: العوامل التي أخذ بها لتوليد البيانات:

لمعرفة أثر حجم العينة وطول الاختبار في دقة معادلة الدرجات باستخدام طريقة كيرنيل المستخدمة في هذه الدراسة، تم أخذ بعض العوامل في عين الاعتبار عند القيام بالمقارنة، ومن العوامل التي حددت إجراءات المقارنة:

١- **حجم العينة:** واستخدم ثلاثة مستويات لحجم العينة وهي (٢٠٠، ٦٠٠، ١٠٠٠) وهذه الأحجام للعينات تعتبر مناسبة لطرق المعادلة باستخدام تصميم المجموعات العشوائية.

٢- **طول الاختبار:** واستخدم ثلاثة مستويات لطول الاختبار وهي ٣٠-٤٠-٦٠ فقرة. وهنا تم استخدام معالم صعوبة الفقرات لتصنيف جميع النماذج على مستوى الصعوبة.

ثالثا: إجراءات توليد البيانات:

تم استخدام برنامج (Wingen 2) لتوليد عينات عشوائية تمثل مجتمع التوزيعات ولإيجاد معيار العلاقة بين طرق المعادلة. أما الخطوات العامة المستخدمة في عملية توليد البيانات فهي كما يأتي:

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

٢- توليد البيانات وفقاً لنموذج نظرية استجابة الفقرة المطلوب.

٣- تقدير المعالم مثل: معلمة الصعوبة، والتمييز، والتخمين، ومعلمة القدرة باستخدام إحدى البرامج الخاصة وذلك بالاعتماد على استجابات الفقرة التي تم توليدها.

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

٥- مقارنة النتائج التي تقيس تأثير المتغيرات المراد دراستها.

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

رابعا: خطوات توليد البيانات المستخدمة في هذه الدراسة:

اعتمدت هذه الدراسة على توليد البيانات باستخدام المحاكاة حيث تم استخدام إجراءات

المعادلة نفسها في المستويات الثلاثة لحجم العينة لكل متغير (طول الاختبار، وحجم 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.

إجراءات جمع البيانات

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

متغيرات الدراسة

اشتملت هذه الدراسة على المتغيرات المستقلة الرئيسية الآتية:

١- حجم العينة: تم استخدام ثلاثة مستويات لحجم العينة وهي: ٢٠٠، ٦٠٠، ١٠٠٠ وهذه الأحجام للعينات تُعَدُّ مناسبة لطرق المعادلة باستخدام تصميم المجموعات العشوائية.

٢- طول الاختبار وتم استخدام ثلاثة مستويات لطول الاختبار وهي ٣٠-٤٠-٦٠ فقرة.

أما المتغيرات التابعة فقد شكلت معيار الحكم على دقة معادلة درجات الصور المختلفة للاختبارات، وذلك بحساب الخطأ المعياري للمعادلة (SEE) والجذر التربيعي لمتوسط مربعات الخطأ (RMSE).

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).

واستخدم الباحث هذا البرنامج في هذه الدراسة لتوليد العينات من الاستجابات المتعددة لعينات تحاكي عينات المجتمع الأصلي يتراوح حجمها من (٢٠٠) إلى (١٠٠٠) مفحوص تحت افتراض التوزيع الطبيعي، واختبارات يتراوح طولها من (٣٠) فقرة إلى (٦٠) فقرة تحت افتراض التوزيع الطبيعي (السوي) لصعوبة الفقرات والتوزيع المنتظم لمعلمة التمييز والتخمين.

ولعل التوليد للعينات يعالج العيوب والمشكلات التي تواجه بعض الباحثين في ضبط الموقف الاختباري، الأمر الذي يمكنهم من الحصول على إجابات دقيقة مقارنة بالأدوات البحثية الحقيقية التي تطبق على الأفراد إذ تعاني من بعض العيوب والمشكلات مثل: عدم الجدية والتخمين وعدم إكمال الإجابة عن الاختبار. وقد أجريت العديد من الدراسات لمقارنة نتائج عينات من المجتمع الأصلي وأشارت النتائج إلى أن ٩٥٪ من النتائج كانت متطابقة. وهذا مؤشر على مصداقية النتائج التي نحصل عليها في مثل هذا النوع من الدراسات (Harwell, 1997).

منهجية المعادلة

١- مطابقة النماذج الخطية المتعددة لتوزيعات الدرجات الخام.

٢- تقدير الدرجات المحتملة للمجتمع المستهدف.

٣- استخدام توزيع كيرنيل لعمل اتصال بين العلامات المنفصلة في التوزيع.

٤- تطبيق المعادلة المئينية باستخدام التوزيع المتصل في الخطوة الثالثة.

٥- حساب الخطأ المعياري للمعادلة.

مقاييس المعادلة (إجراءات المعادلة):

لكل مستوى من طول الاختبار وتشابه مستويات الصعوبة في النماذج كان مقياس المعادلة الذي تم اعتماده 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 خطوات المعادلة وتقييمها. وقدتم تطبيق طريقة المعادلة باستخدام برمجية (Equating Recipes).

نتائج الدراسة

هدفت هذه الدراسة إلى معرفة أثر حجم العينة وطول الاختبار في دقة معادلة درجات الاختبارات متعددة الحدود في تصميم الفقرات المشتركة باستخدام طريقة كيرنيل واختيرت هذه الطريقة كموضوع لهذه الدراسة كونها من الطرق الأقل شيوعًا واستخدامًا في معادلة الاختبارات في الوطن العربي، حيث تم التوصل إلى نتائج دقة المعادلة في ظل استخدام السابقة منفردة ومجتمعة، وقد تم تقدير دقة المعادلة باستخدام معياريّ الخطأ المعياري للمعادلة والجذر التربيعي لمتوسط مربعات الخطأ، وفيما يأتي عرض لأهم نتائج هذه الدراسة، وذلك بتناول نتائج كل متغير مستقل على حدة:

أولاً: النتائج المتعلقة بالسؤال الأول

نص هذا السؤال على: ما أثر حجم العينة وطول الاختبار في الخطأ المعياري للمعادلة عند النقاط المختلفة على سلم الدرجات؟

أ- النتائج المتعلقة بنماذج الاختبار الذي يتألف من ٣٠ فقرة:

تم توليد نموذجين للاختبار من (٣٠) فقرة متشابهين في مستوى الصعوبة (X, Y) بعد ذلك تم توليد العينات العشوائية وبأحجام (٢٠٠-٦٠٠-١٠٠٠) من مجتمع التوزيع، طبقت طريقة (كيرنيل) في المعادلة المستخدمة في هذه الدراسة على هذه العينات العشوائية؛ لتقدير خطأ المعادلة وجذر متوسط مربعات الفروق، وفيما يأتي توضيح لهذه النتائج:

جدول (١)
قيم الخطأ المعياري للمعادلة لطريقة كيرنيل في معادلة الدرجات للاختبار الذي يتألف من ٣٠ فقرة

حجم العينة	٢٠٠	٦٠٠	١٠٠٠
الخطأ المعياري	٢,٧١	٢,٧٠	٢,٤٧

يظهر الجدول (١) أن قيم الخطأ المعياري للمعادلة لطريقة كيرنيل في معادلة الدرجات

للاختبار الذي يتألف من ٣٠ فقرة قد تراوحت بين (٢,٤٧ -٢,٧١)، حيث بلغت أقل قيمة للخطأ المعياري (٢,٤٧)، في حين بلغت أعلى قيمة (٢,٧١)، كذلك يظهر الجدول أن قيمة الخطأ المعياري للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تكونت العينة من (١٠٠٠) مفحوص، بلغت قيمة الخطأ المعياري (٢,٤٧)، وعندما بلغ حجم العينة (٦٠٠) مفحوص، بلغت قيمة الخطأ المعياري (٢,٧٠)، وعندما بلغ حجم العينة (٢٠٠) مفحوص، بلغت قيمة الخطأ المعياري (٢,٧١). ويوضح الشكل (١) قيم الخطأ المعياري للمعادلة لطريقة كيرنيل على سلم الدرجات وعبر اختلاف حجم العينات.

الشكل (١)
الخطأ المعياري للمعادلة لطريقة كيرنيل في معادلة الاختبارات للاختبار الذي يتألف من (٣٠) فقرة

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

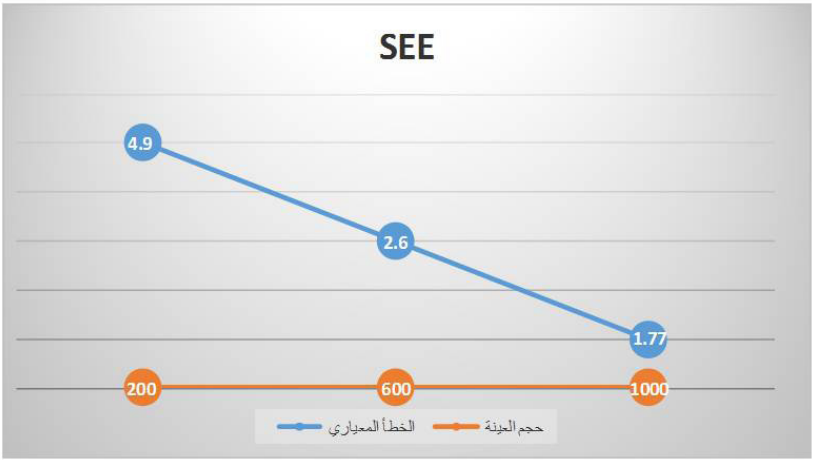
ب - النتائج المتعلقة بنماذج الاختبار الذي يتألف من ٤٠ فقرة:

تم توليد نموذجين للاختبار من (٤٠) فقرة متشابهين في مستوى الصعوبة (X, Y) بعد ذلك تم توليد العينات العشوائية وبأحجام (٢٠٠-٦٠٠-١٠٠٠) من مجتمع التوزيع، ثم طبقت طريقة (كيرنيل) في المعادلة على هذه العينات العشوائية لتقدير خطا المعادلة وجذر متوسط مربعات الفروق، وفيما يأتي توضيح لهذه النتائج:

جدول (٢)
الخطأ المعياري للمعادلة لطريقة كيرنيل في معادلة الدرجات للاختبار الذي يتألف من (٤٠) فقرة

حجم العينة	٢٠٠	٦٠٠	١٠٠٠
الخطأ المعياري	٤,٩٠	٢,٦٠	١,٧٧

يظهر الجدول (٢) أن قيم الخطأ المعياري للمعادلة لطريقة كيرنيل في معادلة الدرجات للاختبار الذي يتألف من ٤٠ فقرة قد تراوحت بين (١,٧٧ -٤,٩٠)، حيث بلغت أقل قيمة للخطأ المعياري (١,٧٧)، في حين بلغت أعلى قيمة (٤,٩٠)، كذلك يظهر الجدول أن قيمة الخطأ المعياري للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تألفت العينة من (١٠٠٠) مفحوص، بلغت قيمة الخطأ المعياري (١,٧٧)، وعندما بلغ حجم العينة (٦٠٠) مفحوص، بلغت قيمة الخطأ المعياري (٢,٦٠)، وعندما بلغ حجم العينة (٢٠٠) مفحوص، بلغت قيمة الخطأ المعياري (٤,٩٠)، ويوضح الشكل رقم (٢) قيم الخطأ المعياري للمعادلة باستخدام طريقة كيرنيل على سلم الدرجات وعبر الأحجام المختلفة للعينات (٢٠٠-٦٠٠-١٠٠٠) عندما تتألف نماذج الاختبار من (٤٠) فقرة.



الشكل (٢)
الخطأ المعياري للمعادلة للاختبار الذي يتألف من (٤٠) فقرة.

يظهر من الشكل السابق أن أقل قيمة للخطأ المعياري للمعادلة (SEE) تم الحصول عليها عندما كان حجم العينة (٦٠٠)، في حين بلغت أعلى قيمة للخطأ المعياري عندما كان حجم العينة (٢٠٠).

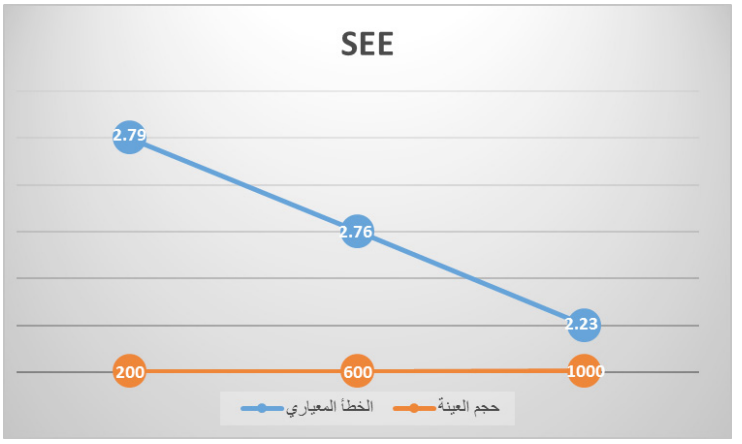
ج-النتائج المتعلقة بنماذج الاختبار الذي يتألف من (٦٠) فقرة:

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 كيرنيل في معادلة درجات الاختبارات؛ لتقدير خطأ المعادلة وجذر متوسطات مربعات الفروق وفيما يأتي توضيح لهذه النتائج:

جدول (٣)
الخطأ المعياري للمعادلة لطريقة كيرنيل في معادلة درجات الاختبارات للاختبار الذي يتألف من (٦٠) فقرة

حجم العينة	٢٠٠	٦٠٠	١٠٠٠
الخطأ المعياري	٢,٧٩	٢,٧٦	٢,٢٣

يظهر الجدول (٣) أن قيم الخطأ المعياري للمعادلة لطريقة كيرنيل في معادلة الدرجات للاختبار الذي يتألف من ٦٠ فقرة قد تراوحت بين (٢,٢٣ -٢,٧٩)، حيث بلغت أقل قيمة للخطأ المعياري (٢,٢٣)، في حين بلغت أعلى قيمة (٢,٧٩)، كذلك يظهر الجدول أن قيمة الخطأ المعياري للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تألفت العينة من (١٠٠٠) مفحوص، بلغت قيمة الخطأ المعياري (٢,٢٣)، وعندما بلغ حجم العينة (٦٠٠) مفحوص، بلغت قيمة الخطأ المعياري (٢,٧٦)، وعندما بلغ حجم العينة (٢٠٠) مفحوص، بلغت قيمة الخطأ المعياري (٢,٧٩)، ويوضح الشكل رقم (٣) قيم الخطأ المعياري للمعادلة باستخدام طريقة كيرنيل على سلم الدرجات وعبر الأحجام المختلفة للعينات (٢٠٠-٦٠٠-١٠٠٠) عندما تتألف نماذج الاختبار من (٦٠) فقرة.



الشكل (٣)
الخطأ المعياري للمعادلة لطريقة كيرنيل في معادلة درجات الاختبارات للاختبار الذي يتألف من (٦٠) فقرة

يوضح الشكل السابق أن طريقة كيرنيل عبر الأحجام المختلفة للعينات كانت قيمة الخطأ المعياري للمعادلة فيها أقل قيمة، ويلاحظ من الشكل السابق أن قيمة الخطأ المعياري للمعادلة تنخفض مع ازدياد حجم العينة.

ثانيا: النتائج المتعلقة بالسؤال الأول

نص هذا السؤال على: ما أثر حجم العينة وطول الاختبار في البواقي المعيارية للمعادلة (RMSE) عند النقاط المختلفة على سلم الدرجات؟

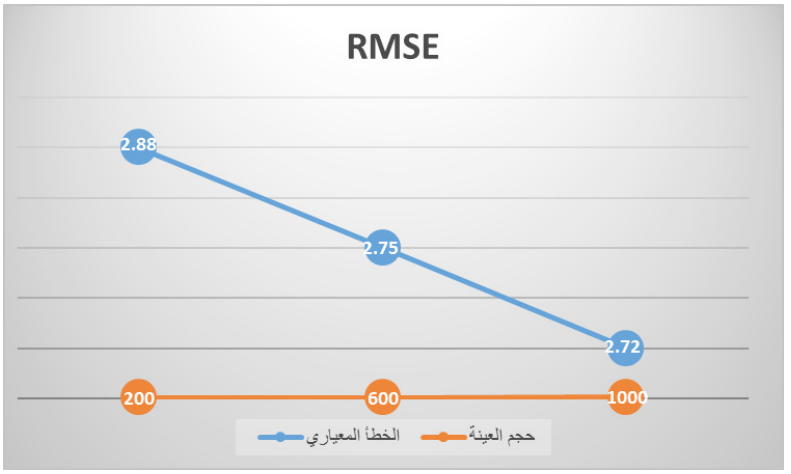
أ- النتائج المتعلقة بنماذج الاختبار الذي يتألف من ٣٠ فقرة:

تم توليد نموذجين للاختبار من (٣٠) فقرة متشابهين في مستوى الصعوبة (X, Y) بعد ذلك تم توليد العينات العشوائية وبأحجام (٢٠٠-٦٠٠-١٠٠٠) من مجتمع التوزيع، طبقت طريقة (كيرنيل) في المعادلة المستخدمة في هذه الدراسة على هذه العينات العشوائية؛ لتقدير البواقي المعيارية، وفيما يأتي توضيح لهذه النتائج:

جدول (٤)
قيم (RMSE) لطريقة كيرنيل في معادلة الاختبارات للاختبار الذي يتألف من ٣٠ فقرة

حجم العينة	٢٠٠	٦٠٠	١٠٠٠
قيم (RMSE)	٢,٨٨	٢,٧٥	٢,٧٢

يظهر الجدول (٤) أن قيم (RMSE) لطريقة كيرنيل في معادلة الدرجات للاختبار الذي يتألف من ٣٠ فقرة قد تراوحت بين (٢,٧٢ -٢,٨٨)، حيث بلغت أقل قيمة لـ (RMSE) (٢,٧٢)، في حين بلغت أعلى قيمة (٢,٨٨)، كذلك يظهر الجدول أن قيم (RMSE) للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تألفت العينة من (١٠٠٠) مفحوص، بلغت قيمة (RMSE) (٢,٧٢)، وعندما بلغ حجم العينة (٦٠٠) مفحوص، بلغت قيمة (RMSE) (٢,٧٥)، وعندما بلغ حجم العينة (٢٠٠) مفحوص، بلغت قيمة (RMSE) (٢,٨٨). ويوضح الشكل (٤) قيم (RMSE) لطريقة كيرنيل في معادلة الدرجات على سلم الدرجات وعبر اختلاف احجام العينات للاختبار الذي يتألف من ٣٠ فقرة.



الشكل (٤)
قيم RMSE لطريقة كيرنيل في معادلة الدرجات للاختبار الذي يتألف من (٣٠) فقرة

يوضح الشكل السابق أن طريقة كيرنيل وعبر الأحجام المختلفة للعينات كانت قيمة (RMSE) فيها أقل قيمة عندما كانت حجم العينة كبيرة (١٠٠٠ مفحوص)، في حين كانت أعلى قيمة عند استخدام حجم عينة صغير (٢٠٠ مفحوص)، ويلاحظ من الشكل السابق أن قيمة (RMSE) تنخفض مع ازدياد حجم العينة.

ب - النتائج المتعلقة بنماذج الاختبار الذي يتألف من ٤٠ فقرة:

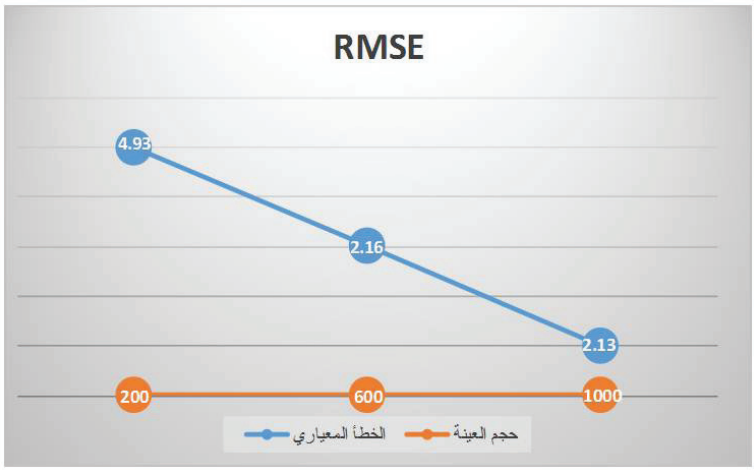
تم توليد نموذجين للاختبار من (٤٠) فقرة متشابهين في مستوى الصعوبة (X, Y) بعد ذلك تم توليد العينات العشوائية وبأحجام (٢٠٠-٦٠٠-١٠٠٠) من مجتمع التوزيع، ثم طبقت طريقة (كيرنيل) في المعادلة على هذه العينات العشوائية لتقدير البواقي المعيارية، وفيما يأتي توضيح لهذه النتائج:

جدول (٥)
قيم (RMSE) لطريقة كيرنيل في معادلة درجات الاختبارات للاختبار الذي يتألف من ٤٠ فقرة

حجم العينة	٢٠٠	٦٠٠	١٠٠٠
قيم (RMSE)	٤,٩٣	٢,١٦	٢,١٣

يظهر الجدول (٥) أن قيم (RMSE) لطريقة كيرنيل في معادلة الدرجات للاختبار الذي يتألف من ٤٠ فقرة قد تراوحت بين (٢,١٣ -٤,٩٣)، حيث بلغت أقل قيمة للخطأ

المعياري (٢,١٣)، في حين بلغت أعلى قيمة (٤,٩٣)، كذلك يظهر الجدول أن قيم (RMSE) للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تألفت العينة من (١٠٠٠) مفحوص، بلغت قيمة (RMSE) (٢,١٣)، وعندما بلغ حجم العينة (٦٠٠) مفحوص، بلغت قيمة (RMSE) (٢,١٦)، وعندما بلغ حجم العينة (٢٠٠) مفحوص، بلغت قيمة (RMSE) (٤,٩٣). ويوضح الشكل (٥) قيم (RMSE) باستخدام طريقة كيرنيل على سلم الدرجات وعبر الأحجام المختلفة (٢٠٠-٦٠٠-١٠٠٠).



شكل (٥)
قيم RMSE للمعادلة لطريقة كيرنيل في معادلة درجات الاختبارات للاختبار الذي يتألف من (٤٠) فقرة

يلاحظ من الشكل السابق أن أقل قيمة من قيم (RMSE) تم الحصول عليها عند استخدام حجم عينة (١٠٠٠)، في حين بلغت أعلى قيمة من قيم (RMSE) عند استخدام حجم عينة (٢٠٠). ويلاحظ من الجدول السابق أن قيم (RMSE) تقل بازدياد حجم العينة والعكس صحيح.

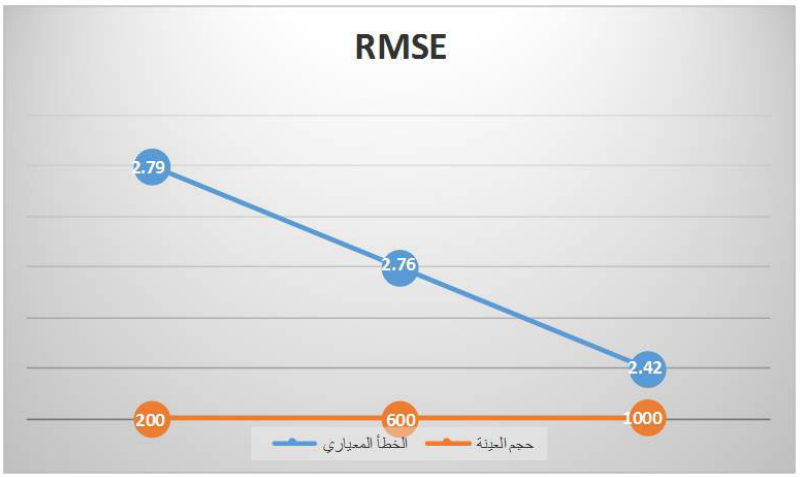
ج-النتائج المتعلقة بنماذج الاختبار الذي يتألف من (٦٠) فقرة:

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 كيرنيل في معادلة درجات الاختبارات؛ لتقدير البواقي المعيارية، وفيما يأتي توضيح لهذه النتائج:

جدول (٦)
قيم (RMSE)لطريقة كيرنيل في معادلة درجات الاختبارات
للاختبار الذي يتألف من (٦٠) فقرة

حجم العينة	٢٠٠	٦٠٠	١٠٠٠
قيم RMSE	٢,٧٩	٢,٧٦	٢,٤٢

يظهر الجدول (٦) أن قيم (RMSE) لطريقة كيرنيل في معادلة الدرجات للاختبار الذي يتألف من ٦٠ فقرة قد تراوحت بين (٢,٤٢ -٢,٧٩) ، حيث بلغت أقل قيمة للخطأ المعياري (٢,٤٢)، في حين بلغت أعلى قيمة (٢,٧٩)، كذلك يظهر الجدول أن قيم (RMSE) للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تألفت العينة من (١٠٠٠) مفحوص، بلغت قيمة (RMSE) (٢,٤٢)، وعندما بلغ حجم العينة (٦٠٠) مفحوص، بلغت قيمة (RMSE) (٢,٧٦)، وعندما بلغ حجم العينة (٢٠٠) مفحوص، بلغت قيمة (RMSE) (٢,٧٩)، ويظهر الجدول أن قيم (RMSE) تتأثر بحجم العينة، فكلما كبر حجم العينة قلت قيمة (RMSE)، وإذا نقص حجم العينة ارتفعت قيم (RMSE). ويوضح الشكل (٦) قيم (RMSE) عند استخدام طريقة كيرنيل في معادلة درجات الاختبارات على سلم الدرجات وعبر الأحجام المختلفة للعينات (٢٠٠-٦٠٠-١٠٠٠):



شكل (٦)
قيم RMSE للمعادلة لطريقة كيرنيل في معادلة درجات الاختبارات
للاختبار الذي يتألف من (٦٠) فقرة

يظهر الشكل (٦) أن قيم (RMSE) تتأثر بحجم العينة، فكلما كبر حجم العينة قلت قيمة (RMSE)، وإذا نقص حجم العينة ارتفعت قيم (RMSE).

مناقشة النتائج والتوصيات:

هدفت هذه الدراسة إلى معرفة أثر حجم العينة وطول الاختبار في دقة معادلة درجات الاختبارات متعددة الحدود في تصميم الفقرات المشتركة باستخدام طريقة كيرنيل، وذلك من خلال الإجابة عن السؤالين الآتيين:

١- ما أثر حجم العينة وطول الاختبار في الخطأ المعياري للمعادلة (SEE) عند النقاط المختلفة على سلم الدرجات؟

٢- ما أثر حجم العينة وطول الاختبار في البواقي المعيارية للمعادلة (RMSE) عند النقاط المختلفة على سلم الدرجات؟

أولا: مناقشة النتائج المتعلقة بالسؤال الأول

ما أثر حجم العينة وطول الاختبار في الخطأ المعياري للمعادلة عند النقاط المختلفة على سلم الدرجات؟

أظهرت النتائج أن قيم الخطأ المعياري للمعادلة لطريقة كيرنيل في معادلة الدرجات للاختبار الذي يتألف من ٣٠ فقرة قد تراوحت بين (٢,٤٧ -٢,٧١)، حيث بلغت أقل قيمة للخطأ المعياري (٢,٤٧)، في حين بلغت أعلى قيمة (٢,٧١)، كذلك يظهر الجدول أن قيمة الخطأ المعياري للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تكونت العينة من (١٠٠٠) مفحوص، بلغت قيمة الخطأ المعياري (٢,٤٧)، وعندما بلغ حجم العينة (٦٠٠) مفحوص، بلغت قيمة الخطأ المعياري (٢,٧٠)، وعندما بلغ حجم العينة (٢٠٠) مفحوص، بلغت قيمة الخطأ المعياري (٢,٧١). كما أظهرت النتائج أن قيم الخطأ المعياري للمعادلة لطريقة كيرنيل في معادلة الدرجات للاختبار الذي يتألف من ٤٠ فقرة قد تراوحت بين (١,٧٧ -٤,٩٠)، حيث بلغت أقل قيمة للخطأ المعياري (١,٧٧)، في حين بلغت أعلى قيمة (٤,٩٠)، كذلك يظهر الجدول أن قيمة الخطأ المعياري للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تألفت العينة من (١٠٠٠) مفحوص، بلغت قيمة الخطأ المعياري (١,٧٧)، وعندما بلغ حجم العينة (٦٠٠) مفحوص، بلغت قيمة الخطأ المعياري (٢,٦٠)، وعندما بلغ حجم العينة (٢٠٠) مفحوص، بلغت قيمة الخطأ المعياري (٤,٩٠)، كما أظهرت النتائج أن قيم الخطأ المعياري للمعادلة لطريقة كيرنيل في معادلة الدرجات للاختبار الذي يتألف من ٦٠ فقرة قد تراوحت بين (٢,٢٣ -٢,٧٩)، حيث بلغت أقل قيمة للخطأ المعياري (٢,٢٣)، في حين بلغت أعلى قيمة (٢,٧٩)، كذلك يظهر الجدول أن قيمة الخطأ المعياري للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تألفت العينة من (١٠٠٠) مفحوص، بلغت قيمة الخطأ المعياري (٢,٢٣)، وعندما بلغ حجم

العينة (٦٠٠) مفحوص، بلغت قيمة الخطأ المعياري (٢,٧٦)، وعندما بلغ حجم العينة (٢٠٠) مفحوص، بلغت قيمة الخطأ المعياري (٢,٧٩). أظهرت النتائج أن طول الاختبار يوثر في الخطأ المعياري للمعادلة، فالاختبار الطويل يعطي قيمًا كبيرة للخطأ المعياري للمعادلة، في حين أن تقليل طول الاختبار من (٦٠) إلى (٤٠) يؤدي إلى تخفيض كبير للخطأ المعياري للمعادلة مقارنة مع التخفيض عندما يكون طول الاختبار يتناقص من (٤٠) إلى (٣٠). فعلى سبيل المثال عندما كان الخطأ المعياري (٢,٢٣) للاختبار الذي يتكون من (٦٠) فقرة وعند حجم عينة (١٠٠٠) مفحوص، فإن هذا الخطأ المعياري انخفض إلى (١,٧٧) للاختبار الذي يتكون من (٤٠) فقرة وبنفس حجم العينة.

إن قيمة الخطأ المعياري يتناسب عكسيًا مع حجم العينة، حيث إنه يزداد بزيادة حجم العينة، وتقل قميته بنقصان حجم العينة، وهذا أمر طبيعي؛ إذ أن كمية المعلومات عند اي مستوى من مستويات القدرة تتناسب عكسيًا مع الخطأ المعياري، فعند زيادة حجم العينة يقترب متوسط معلمة التخمين من الصفر وهذا يؤدي إلى التقليل من قيمة الخطأ المعياري، حيث إن 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)، التي توصلت إلى أن دقة تقدير الخطأ المعياري في المعادلة، والبواقي المعيارية كان أفضل في العينات ذات الحجم الكبير، وهذا ما توصلت إليه هذه الدراسة.

ثانيا: مناقشة النتائج المتعلقة بالسؤال الثاني

ما أثر حجم العينة وطول الاختبار في البواقي المعيارية للمعادلة (RMSE) عند النقاط المختلفة على سلم الدرجات؟

أظهرت النتائج أن قيم (RMSE) لطريقة كيرنيل في معادلة الدرجات للاختبار الذي يتألف من ٣٠ فقرة قد تراوحت بين (٢,٧٢ -٢,٨٨)، حيث بلغت أقل قيمة لـ (RMSE) (٢,٧٢)، في حين بلغت أعلى قيمة (٢,٨٨)، فقد أظهرت النتائج أن قيم (RMSE) للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تألفت العينة من (١٠٠٠) مفحوص، بلغت قيمة (RMSE) (٢,٧٢)، وعندما بلغ حجم العينة (٦٠٠) مفحوص، بلغت قيمة (RMSE) (٢,٧٥)، وعندما بلغ حجم العينة (٢٠٠) مفحوص، بلغت قيمة (RMSE) (٢,٨٨). كما

أظهرت النتائج أن قيم (RMSE) للاختبار الذي يتألف من ٤٠ فقرة قد تراوحت بين (٢,١٣ -٤,٩٣)، حيث بلغت أقل قيمة (٢,١٣)، في حين بلغت أعلى قيمة (٤,٩٣)، كما أظهرت النتائج أن قيم (RMSE) للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تألفت العينة من (١٠٠٠) مفحوص، بلغت قيمة (RMSE) (٢,١٣)، وعندما بلغ حجم العينة (٦٠٠) مفحوص، بلغت قيمة (RMSE) (٢,١٦)، وعندما بلغ حجم العينة (٢٠٠) مفحوص، بلغت قيمة (RMSE) (٤,٩٣). كما أظهرت النتائج أن قيم (RMSE) لطريقة كيرنيل في معادلة الدرجات للاختبار الذي يتألف من ٦٠ فقرة قد تراوحت بين (٢,٤٢ -٢,٧٩) ، حيث بلغت أقل قيمة للخطأ المعياري (٢,٤٢)، في حين بلغت أعلى قيمة (٢,٧٩). إن قيم (RMSE) للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تألفت العينة من (١٠٠٠) مفحوص بلغت قيمة (RMSE) (٢,٤٢)، وعندما بلغ حجم العينة (٦٠٠) مفحوص، بلغت قيمة (RMSE) (٢,٧٦) ، وعندما بلغ حجم العينة (٢٠٠) مفحوص، بلغت قيمة (RMSE) (٢,٧٩)، لقد أظهرت النتائج أن قيم (RMSE) تتأثر بحجم العينة، فكلما كبر حجم العينة قلت قيمة (RMSE)، وإذا نقص حجم العينة ارتفعت قيم (RMSE) .

وعند مقارنة نتائج قيم (RMSE) للمعادلة من خلال طول الاختبار، أظهرت النتائج ان طول الاختبار يوثر في قيم (RMSE)، فالاختبار الطويل يعطي قيمًا كبيرة لـ (RMSE)، في حين أن تقليل طول الاختبار من (٦٠) إلى (٤٠) يؤدي إلى تخفيض كبير لقيم (RMSE) مقارنة مع التخفيض عندما يكون طول الاختبار يتناقص من (٤٠) إلى (٣٠)، فالاختبار الطويل يعطي قيمًا كبيرةً لـ (RMSE)، فعند تقليل طول الاختبار من (٦٠) إلى (٤٠) فقرة، أدى هذا إلى انخفاض كبير لقيم (RMSE) مقارنة مع الانخفاض عندما كان طول الاختبار يتناقص من (٤٠) إلى (٣٠). فعلى سبيل المثال عندما كانت قيمة (RMSE) (٢,٤٢) للاختبار الذي يتكون من (٦٠) فقرة وعند حجم عينة (١٠٠٠) مفحوص، كان الخطأ المعياري قد انخفض إلى (٢,١٣) للاختبار الذي يتكون من (٤٠) فقرة وبنفس حجم العينة.

إن الاختلاف في قيم (RMSE) للمعادلة يصبح أقل عندما ينخفض طول الاختبار. وهذه يمكن تفسيرها في عاملين رئيسين:

أ- إن قيم RMSE المستخدمة هنا ليست قيمًا معيارية (Standardized)، لأن فقرات الاختبارات لم تقسم باستخدام تطابق الانحراف المعياري للعلامات المشاهدة؛ ولأن الانحراف المعياري للدرجات الصحيحة يرتفع مع زيادة طول الاختبار، وهنا يتوقع أنّ القيم المعيارية لقيم (RMSE) سوف تظهر ميلاً قليلاً إلى الارتفاع مع ازدياد طول الاختبار.

ب- إن الاختبار الطويل يعني احتمالية عالية 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. وعند مقارنة نتائج البواقي المعيارية للمعادلة وعبر الاحجام المختلفة للعينات وتحت ظرف اختلاف حجم العينة وطول الاختبار، فإن النتائج أظهرت تقارب نتائج الخطأ المعياري للمعادلة (SEE) مع نتائج (RMSE) عندما يرتفع حجم العينة ويقل طول الاختبار. حيث أظهرت النتائج أن قيم الخطأ المعياري وقيم (RMSE) تتأثر وبدرجة كبيرة بحجم العينة وطول الاختبار.

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

الاستنتاجات

خلصت الدراسة إلى الاستنتاجات الآتية:

١- إن طريقة كيرينل تميل قيم الخطأ المعياري فيها إلى الانخفاض مع ارتفاع حجم العينة وخصوصاً لنماذج الاختبار من (٤٠) – (٦٠) فقرة، كذلك الحال بالنسبة لقيم (RMSE) فيها إذ تميل إلى الانخفاض مع ازدياد حجم العينة.

٢- أظهرت نتائج هذه الدراسة أن أحجام العينات الكبيرة يوثر في 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 يرتفع حجم العينة.

٤- الاختبار الطويل يؤدي إلى ازدياد قيم الخطأ المعياري و(RMSE).

التوصيات والمقترحات

١- يوصي الباحث بتطبيق بيانات حقيقية متعددة الاستجابة، وذلك للوقوف على مقدار الاختلاف في المعادلة بين البيانات الحقيقية والبيانات التجريبية.

٢- استخدام طرق أخرى للمعادلة للوقوف على مدى دقتها في معادلة الدرجات.

٣-استخدام تصميم آخر غير تصميم الفقرات المشتركة؛ للوقوف على مقدار الاختلاف في دقة المعادلة لنفس الطريقة المستخدمة في هذه الدراسة.

المراجع

Akour, M. (2006). *A comparison of various equipercentile and Kernel equating methods under the random groups*. Unpublished Doctoral Dissertation University of Iowa, USA.

Alina, A. Paul, W. Holland, & Dorothy T. (2004).*The kernel method of test equating*. New York: Springer – Verlag.

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.

Hambelton, R & swaminthan, H. (1991). *Item Response Theory: Principles and Applications*. Boston, Kluwer: Nijhoff Publishing.

Hambleton, R& Van der & Linden, W. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement; 6,* 373-378.

Hambleton, R & Han, K. (2007). *User manual for winGen: Windows Software that generates (IRT) model parameters and item Response*. New York: Holt Rinehart and Weston.

Harwell, M. (1997) .Analyzing the results of Monte Carlo studies in item response theory. Applied Psychological Measurement, 57(2), 266-280.

Holland, W. Thayer, D. Tayer. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions*. (Technical Report 87- 79), Princeton, NJ: Educational Testing Service.

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

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

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.

Qu, Y. (2007). T*he Effect of Weighting in Kernel Equating Using Counter-Balanced Designs*. Unpublished Doctoral Dissertation, Michigan State University, USA.

Wang, T. (2005a). *An alternative continuation method to the Kernel Method in von Davier, Holland and Thayer's (2004) test equating Framework*. (CASMA Research Report 11). Iowa City: Center for Advanced Studies in Measurement and assessment.