دراسات، العلوم التربوية، المجلّد 46، العدد 1، ملحق 3، 2019

فاعلية طريقة العلامات المشاهدة وكيرنيل في معادلة درجات الاختبارات

يوسف عبد العاطي محمد المحروق *

ملخص

هدفت هذه الدراسة إلى الكشف عن فاعلية طريقة العلامات المشاهدة وكيرنيل في معادلة درجات الاختبارات، من خلال دراسة المتغيرات: حجم العينة، والاستماع قراءة صوتية للكلمات القاموس – عرض القاموس المفصل طول الاختبار، تم توليد بيانات تجريبية باستخدام نظرية استجابة الفقرة لتوليد استجابات المفحوصين على فقرات الاختبار، أشارت النتائج إلى أن حجم العينات الكبير يقلل من الخطأ المعياري للمعادلة، ويقلل من البواقي المعيارية. كذلك أظهرت النتائج أن طول الاختبار يوثر في الخطأ المعياري؛ فالاختبار الطويل يعطي قيما كبيرة للخطأ المعياري والبواقي المعيارية. كما أظهرت النتائج أن طريقة المعادلة باستخدام الدرجات الملاحظة كانت أكثر دقة من طريقة كيرنيل.

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

المقدمة

اهتم علماء النفس ومنذ ظهور حركة القياس النفسى والتربوي بتحقيق صدق وثبات الاختبارات والمقاييس النفسية، سعياً منهم لتحقيق أعلى درجة من الموضوعية فى هذه الأدوات، عند استخدامها فى عملية القياس؛ لذلك فقد وظفوا مباديء نظرية القياس النفسي التربوي الكلاسيكية في بناء الاختبارات، وظل العاملون بالقياس النفسي يستخدمون مبادئ وأسس هذه النظرية في بناء الاختبارات والمقاييس بأشكالها المختلفة وتفسير الدرجات المتحققة عليها لفترة طويلة من الوقت.

إن بناء الاختبارات في المجالات التربوية والنفسية بعد من أهم التطبيقات التي ركزت عليها النظرية الكلاسيكية في القياس؛ من أجل تزويد صاحب القرار او المعلم بمعلومات يمكن استخدامها في اتخاذ قرارات تتعلق بالتعيين او الترقية او إعطاء الطالب تقديرات تتعلق بتحصيله الاكاديمي. (Hambelton and linden, 1985).

ولضمان امن الاختبارات تقوم العديد من المؤسسات التربوية التي تعنى في بناء وتصميم الاختبارات بإعداد نماذج متعددة للاختبارات، من خلال تمثيل فقرات مختلفة في مواقع مختلفة من الاختبارات لتصبح هذه الاختبارات أكثر عدالة لدرجة مقبولة، إلا أن النماذج المتعددة للاختبارات تختلف غالبا في صعوبتها، وتختلف كذلك في المفحوصين الذين سيطبق 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 and Brenann, 2004)

ويعد مجال معادلة الاختبارات من المجالات التطبيقية الهامة في القياس التربوي؛ ففي الكثير من البرامج الاختبارية تنشا الحاجة إلى استخدام عدة صور من الاختبار الواحد لضمان السرية، كما أن الصور المتعددة ضرورية في اختبارات القبول وفي التقويم البنائي للطالب حيث يلزم تطبيق اختبارات دورية ومتعددة اثناء الفصل الدراسي ومقارنة العلامات التي يحصل عليها عبر الزمن (Kolen and Brenann, 2004)، لذلك تقوم العديد من المؤسسات التربوية التي تعنى في قياس تحصيل الطلاب بإعداد نماذج متعددة للاختبارات؛ ضمانا لامن هذه الاختبارات، وذلك من خلال تضمين فقرات مختلفة في مواقع مختلفة في هذه الاختبارات لتصبح هذه الاختبارات تمتاز بالعدالة لدرجة مقبولة، بحيث إن العلامة على أي صورة يكون لها نفس الدلالة ونفس القيمة القياسية لو تحققت نفسها على صورة أخرى، ويقال عندئذ ان هناك تكافؤا تاماً بين صور الاختبار الواحد، لكن نادراً ما يتحقق مثل هذا التكافؤ التام بين صور أعدت عن نفس الاختبار، وعندها يصبح من الضروري معادلة هذه الصور بتحويل نظام الوحدات في احدى الصور الى نظام الوحدات في الصورة الأخرى حتى تكون العلامات المشتقة من الصورتين بعد عملية التحويل متكافئة تماماً (المدانات، 2008).

---

* وزارة التربية والتعليم، مملكة البحرين. تاريخ استلام البحث 2016/01/14، وتاريخ قبوله 2017/01/13.

عمليا قلما نحصل على نموذجين متكافئين تماما في المستوى وفي مدى الصعوبة لنفس الاختبار، لذا يصبح من الضروري ان نعادل هذه النماذج؛ بمعنى ان نقوم بتحويل العلامات في أحد النماذج إلى ما يناظرها او ما يكافئها في النموذج الآخر، وتتطلب معظم البرامج الاختبارية أن يكون لدينا نماذج متكافئة من نفس الاختبار، ولإجراء المعادلة بين نموذجين (صورتين) لنفس الاختبار فإننا نطبق النموذجين على عينة من الأفراد – حسب التصميم المناسب الذي تم اختياره– ثم نجري التحليلات الإحصائية المناسبة التي تمكننا من تحويل العلامات من نموذج الى آخر عندما يطبق على أفراد أحد النموذجين (Angoff. 1971).

إذا تمت عملية معادلة درجات الاختبارات بنجاح، فان 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 and Brenann, 2004):

1- اختيار تصميم جمع المعلومات.
2- اختيار واحدة أو أكثر من التعريفات الإجرائية للمعادلة.
3- اختيار طريقة التقدير الإحصائية المناسبة.
4- تقييم نتائج المعادلة.

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

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

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

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

يصنف المختصون في القياس والتقويم التربوي والنفسي طرق معادلة الاختبارات إلى نوعين: الطرق التي تعتمد على النظرية الكلاسيكية في الاختبارات (CTT) Classical Test Theory التي منها: طريقة المئينات، وطريقة كيرنيل في معادلة الدرجات، 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 Holland, Dorthy,Thayer (1987) ولاحقا تم تحديثها من قبل ألينا فون ديفر (Alina Von Davier, 2004).

تقوم طريقة كيرنيل على مجموعة مرنة من طرق المعادلة الشبيهة بالمعادلة المئينية وتنتمي 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 (Alina Von Daveir,et,al. 2004).

يوجد هناك طريقتان رياضيتان تستخدمان لحساب معاملات معادلة كيرنيل:

1- معادلة تشين Chain equating-CE-وفي معادلة تشين تواجه أولا الربط بين نموذج (x) مع الجذع المشترك A، ثم يتبع بربط A مع نموذج (Y).

2- معادلة ما بعد المطابقة (post stratification-PSE) حيث يتم تقدير التوزيع الهامشي لنموذج(x) ونموذج (Y) في المجتمع المستهدف ثم بعد ذلك تتم إجراءات المعادلة. (Alina Von Daveir,et,al. 2004)

عمليا، يوجد 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) ام خارجي (External). وبشكل عام، يتم تطبيق طريقة كيرنيل في معادلة الاختبارات من خلال نظام يتكون من خمس خطوات متدرجة ومتتابعة وهي:

الخطوة الأولى: قبل التمهيد Pre-smoothing

إن عملية ما قبل التمهيد في طريقة كيرنيل لها هدفان، فهي تزودنا بتوزيع الدرجات التي هي متطلب رئيسي لعمليات المعادلة المتبقية، ثم أنها تكون المصفوفة التي سوف تستخدم لاحقا لحساب الخطأ المعياري للمعادلة. وخلال هذه الخطوة يتم تقدير احتمالات الدرجة في تصميم المعادلة المستخدم، وهذه الاحتمالات تحدد بمطابقة النموذج الإحصائي للعلامات الخام، وهذا

النموذج الإحصائي الذي يستخدم يختلف باختلاف التصميم المستخدم.

الخطوة الثانية: تقدير احتمالات الدرجة (Estimation of score probabilities)

في هذه الخطوة من خطوات كيرنيل تكون توزيعات احتمالية الدرجة تعتمد على المجتمع المستهدف (T) التي تقدر باستخدام توزيعات الدرجات الممهدة التي يتم الحصول عليها من خطوة ما قبل التمهيد.

الخطوة الثالثة: الاستمرارية:Continuation

في هذه الخطوة يتم تحديد شكل توزيع كيرنيل ومتوسطه (الدرجات المنفصلة) ويترك التباين لتوزيع كيرنيل دون تحديد، لكن يتم التحكم بتباين كل توزيعات كيرنيل من خلال استخدام احد المعالم ويسمى (bandwidths, *hX* and *hY*).

إن شكل توزيع كيرنيل يتأثر وبشكل كبير باختيار معلم آل (bandwidth) فإذا كانت قيمة (h) صغيرة على سبيل المثال (0.3) فان تركز توزيعات كيرنيل عند كل نقطة درجة ستكون ضعيفة مع قليل من التداخل، ويكون التوزيع المستمر الناتج من ربط كل التوزيعات الفردية لكير نيل ذات شكل حاد. أما إذا كانت قيمة (h) كبيرة على سبيل المثال (1.0) فان تركز توزيعات كيرنيل عند كل نقطة درجة ستكون كبيرة مع كثير من التداخل، ويكون التوزيع المستمر الناتج من ربط كل التوزيعات الفردية لكير نيل تقريبا ممهدا.(Achour, 2006)

حتى يتم اختيار قيمة مناسبة ل (h) اقترحت الين فون ديفر (AlinaVonDaveir,et,al. 2004 اختيار القيمة التي تقلل من مجموع مربع الاختلافات بين التكرارات المنفصلة والتكثيفات المتطابقة للتوزيع المستمر. إن اختيار قيمة مناسبة ل (h) تعتمد على الباحثين وتتحكم بالكيفية التي يسمح للمشاهدات أن تبتعد عن نقطة المصدر في توزيع (θ). كذلك قيمة(h) يجب أن تقلل من متوسط مربع الخطأ (MSE) حيث إن (MSE) يستعمل بالتعاقب بين تقدير الانحراف (التحيز) والتباين، لذلك فان كل نموذج اختبار له(bandwidth) الخاص به فعلى سبيل المثال فان النموذج (x) له (bandwidth) الخاص به ونموذج (y) له (bandwidth) الخاص به hx,hy على التوالي.

إن القيمة الأفضل ل (h) تحدد من خلال الانتشار والخطأ، ويستطيع أي باحث اختيار قيم متعددة من (h) ويشاهد دقة التمهيد التي تحدث من خلال القيم المختلفة ل(h).

لقد أوصت الين فون ديفر (Alina Von Daveir,et,al. 2004)باستخدام صيغة رياضية لحساب القيمة الأفضل ل (h)، ويتم الحصول على penalty of 1 من خلال المعادلة الآتية:

$$PEN_1(h_x) = \sum_j (\hat{r}_j - \hat{f}_{hx}(x_j))^2.$$

Where $f_{hx}(x_j)$ is the height of the density function at the score x_j.

كذلك يتم الحصول على penalty of 2 من خلال المعادلة الآتية:

$$PEN_2(h_x) = \sum_j A_j(1 - B_j),$$

Where A_j=1 if $f`_{hx}(x) < 0$ a little to the left of $x_j$, and $B_j = 0$ if $f`_{hx}(x) > 0$ a little to the right of x_j.

ويمكن ربط القاعدتين السابقتين باستخدام (Z)

$$PEN_1(h_x) + K \times PEN_2(h_x)$$

لقد وجد أن استخدام pen الذي يقلل قيم *hx* يكون غالبا اكبر مرتين أو ثلاث مرات من (hx=0.33)، لقد وجد أن (z) إذا كانت 1 تكون مفيدة عندما يكون هناك فجوات في التوزيع التي تتطلب عملية التمهيد (Alina Von Daveir,et,al. 2004).

الخطوة الرابعة: المعادلة equating

بعد أن يتم تكثيف الدرجات تكون الخطوة التالية هو حساب معادلة كير نيل وتتم هذه المعادلة بالصيغة الآتية لمعادلة (x) إلى

:(y)

$$e_y(x) = G^{-1}(F(x))$$

ولمعادلة نموذج (y) إلى (x)

$$e_y(y) = G^{-1}(F(x)).$$

بعد أن تتم معادلة الدرجات تكون الخطوة التالية هو تحديد تأثير المعادلة من خلال التوزيع الأصلي للدرجات المنفصلة لـ (x)و (y). وقد اقترحت الين فون ديفر (Alina Von Daveir,et,al. 2004) طريقتين محتملتين لإنجاز هذه المهمة: الأولى هي مقارنة إحصائيات التوزيعين ex(x) و y، وهذه الإحصائيات تقارن من خلال استخدام نسبة الخطأ النسبي the Percent Relative Error (PRE)،

$$\mu_p(e_Y(\vec{X})) = \sum (e_Y(x_j))^p r_j \text{ and } \mu_p(\vec{Y}) = \sum (y_k)^p s_k,$$

ونسبة الخطأ النسبي (PRE) ل (X)و (Y) يتم إيجاده باستخدام الصيغة الآتية:

المتجهات ȳو $\bar{x}$ هي الدرجات من x، y على التوالي.

الطريقة الثانية لتقييم إثر المعادلة هو مقارنة توزيعات التكرار التراكمي. وحتى يتم تحديد الدرجة التي يكون فيها fe(x)(y) تختلف عن G (y) هو حساب الاختلاف:

Fe(x) (y) - G(y

الخطوة الخامسة: حساب الخطأ المعياري للمعادلة (Calculating the (SEE)

إن الخطأ المعياري للمعادلة (SEE) وانحراف الخطأ المعياري للمعادلة يمكن إيجاده لمعادلة كيرنيل، حيث إن الخطأ المعياري للمعادلة يحدد درجة عدم التأكد في تقدير وظيفة المعادلة في حين أن (SEED) هو اختلاف بين الخطأ المعياري لطريقتين من طرق المعادلة، ويمكن حساب الخطأ المعياري لمعادلة $\bar{X}$ إلى ȳ كما يأتي:

$$SEE\ \ y_{(CE)}(X) = \sqrt{\left[SEE_{\ y}(e_A(x)\right] + \left[\bar{e}_y(e_A(x)SEE_{\ A}(x)\right]^2}$$

ويجب أن يلاحظ أن إيجاد الخطأ المعياري للمعادلة هو أكثر ملاءمة للعينات الكبيرة ويمكن أن يحسب في حالة العينات الصغيرة (Alina Von Daveir,et,al. 2004)

أما النوع الثاني من طرق المعادلة فيُصنَّفُ ضمن الطرق التي تعتمد على النظرية الحديثة في القياس التربوي (نظرية الاستجابة للفقرة) Item Response Theory (IRT) مثل طريقة معادلة العلامات الحقيقية وطريقة معادلة العلامات المشاهدة(kolen and Brennan,2004)، وقد تتضمن بعض تصاميم المعادلة استخدام جذع مشترك أي مجموعة من الفقرات المشتركة بين النموذجين أو الصورتين اللتين يتم معادلتهما، لذا قد يستخدم عدد متغير من فقرات الجذع المشتركة، كما قد يستخدم عينات مختلفة في الحجم أو النوع، فقد نختار عينات متكافئة على أساس انتقائها عشوائيا، وقد نختار في مرة أخرى عينات غير متكافئة بشكل غير عشوائي. وللتحقق من فاعلية المعادلة قد تستخدم معايير عديدة منها المؤشرات التلخيصية، والخطأ المعياري للمعادلة، ومحك الاتساق، والمعادلة الدائرية أو معادلة الاختبار لنفسه ومحك الاستقرار(Hua Gao, 2004).

ويشير (Baker and Alkarni, 1991, p 147) إلى أنه من الإسهامات الكبيرة للنظرية الحديثة في القياس التربوي والنفسي -التي اصطلح على تسميتها بنظرية الاستجابة للفقرة (Item Response Theory -IRT) في الممارسة التربوية، قدرتها على وضع عدة اختبارات ومجموعات من المفحوصين على تدريج مشترك common scale في عملية القياس؛ وإمكانية استخدامها في المعادلة الأفقية والرأسية للاختبار.

إن مجموعة الافتراضات التي تقوم عليها نظرية الاستجابة للفقرة، 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، التي 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. ومن هنا، تقوم نظرية الاستجابة لمفردة الاختبار بوصف هذه العلاقة بواسطة دالة تعتمد على مجموعة من الافتراضات، هي: أحادية البُعْدْ (أحادية السمة) unidimensionality 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 الذي 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).

لقد أوضح كل من هامبلتون وسواميناثان (Hambleton and Swaminthan,1985)، وكذلك كولن وبرينان (Kolen and Brennan, 2004)، الخطوات الضرورية لمعادلة الاختبارات بواسطة النظرية الحديثة في القياس التربوي والنفسي؛ وهي كالآتي:

- اختيار التصميم المناسب لمعادلة الاختبار مع الأخذ بعين الاعتبار خصائص مجموعة المفحوصين وطبيعة الاختبارات المراد معادلتها.
- اختيار 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.
- اختيار التدريج المناسب لوضع درجات الاختبار. أي هل تُكْتَبُ الدرجات كدرجات خام (درجات مُشاَهَدَة)، 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 حدٍ كبير القيام بهذه العملية يدوياً، وخاصة في هذه النظرية.

ويوضح كوك وايجنور (Cook and Eignor,1991) الآليات الأساسية لعملية معادلة الاختبارات في هذه النظرية، التي يمكن إيجازها في الآتي:

- اختيار 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 3000 مفحوص. وإذا كانت هناك فقرات مشتركة بين صورتي الاختبار على هيئة اختبار مشترك Anchor Test، فيجب أن تعكس هذه الفقرات المحتوى والخصائص الإحصائية لصورتي الاختبار؛ وأن لا تقل نسبة فقرات الاختبار المشترك عن 20% 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 إن منحنيات خاصية الفقرة Item Characteristic Curves (ICC) 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 Intercept، 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 استخراج 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:

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

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

$\hat{T}_x$ = الدرجة الحقيقية المقدرة للصورة الأولى من الاختبار.

$\hat{T}_y$ = الدرجة الحقيقية المقدرة للصورة الثانية من الاختبار.

$\hat{P}_j(\theta)$ = الدالة المقدِّرة لاستجابة الفقرة في الفقرات للصورة الأولى للاختبار.

$\hat{P}_j(\theta)$ = الدالة المقدِّرة 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 هذا 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، 2004)

أما معادلة الاختبارات باستخدام طريقة الدرجات المشاهَدَة (الدرجات الخام) Observed-score Equating فتقوم 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 $P_1(\theta) = P(\theta)$، فإن التكرار النسبي للدرجات الصحيحة (X) للمفحوص (g) يمكن حسابه رياضياً بالمعادلة الآتية ضمن توزيع ذي الحدين:

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

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

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

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

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

3-. تكرار الخطوة رقم (2) للاختبار الثاني.

4. 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.

وعلى الرغم من وجود النماذج المتعددة للاختبارات التي تم بناؤها بالاعتماد على نفس الخصائص (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 المعادلة بشكل مثالي بغض النظر عن الطريقة المستخدمة في المعادلة، فان المفحوصين سوف يحصلون على نفس الدرجات بغض النظر عن أي اختبار يتم التقدم له. (Kolen and Brenann, 2004) فالمعادلة هي إجراء إحصائي يتم فيه تعديل الاختلاف في مستوى الصعوبة من نموذج إلى آخر، ونتيجة لهذا التعريف ظهر جدل واسع بين علماء القياس حول ما هو نوع الإجراء الإحصائي المميز أو الظروف التي ستكون مقنعة إذا تم إجراء المعادلة؟

مشكلة الدراسة وأسئلتها

تعتمد موضوعية وصدق نتائج الاختبارات على دقة الأساليب التى استخدمت فى بنائها واختيار فقراتها وتفسير نتائجها، وكذلك فى وصفها للقدرة التى يقيسها الاختبار، لذلك فإن الأمر يتطلب ضرورة استخدام اساليب حديثة فى القياس والتى أثبتت البحوث التجريبية أنها تحقق الدقة والموضوعية المنشودة فى القياس التربوي والنفسي، وهذا الامر دفع الباحث لاختيار طريقة كيرنيل و طريقة العلامات المشاهدة في النظرية الحديثة في القياس للكشف عن مدى دقتهما في معادلة درجات الاختبارات متعددة الاستجابة. وبعبارة أخرى فان الغرض الأساسي من الدراسة هو التحقق من دقة معادلة درجات الاختبارات ذات الجذع المشترك للمجموعات المتكافئة باستخدام طريقة كيرنيل وطريقة العلامات المشاهدة في النظرية الحديثة في القياس. وبالتحديد فان هذه الدراسة حاولت الإجابة عن السؤالين الآتيين:

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

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

واختارت هذه الدراسة طريقة كيرنيل وطريقة العلامات المشاهدة في النظرية الحديثة في القياس باعتبارهما أكثر الطرق شيوعاً في المعادلة؛ وذلك للكشف عن دقة معادلة الدرجات في الاختبارات مستخدمين تصميم المجموعات العشوائية ذات الفقرات المشتركة.

أهمية الدراسة

تكتسب الدراسة أهميتها من خلال استخدامها طريقة كيرنيل و طريقة العلامات المشاهدة في النظرية الحديثة في القياس؛ للكشف عن دقة معادلة درجات الاختبارات تحت ظروف: البيانات متعددة الاستجابة، الفقرات المشتركة، حجم العينة، طول الاختبار، ويمكن تلخيص أهمية الدراسة في الجوانب الآتية:

الأهمية النظرية:

- الإسهام في إلقاء الضوء على فاعلية طريقة كيرنيل و طريقة العلامات المشاهدة في النظرية الحديثة في القياس في معادلة درجات الاختبارات متعددة الحدود والمستخدمة في هذه الدراسة، بحيث تسهم في إعطاء صورة واضحة للمتخصصين في مجال بناء الاختبارات لاختيار طريقة المعادلة التي تتواءم مع طبيعة الاختبار الذي يتم بناؤه للحصول على نتائج تمتاز بالشمولية والدقة.
- التحقق من النتائج المحتملة بالإجابة عن الأسئلة التي طرحتها الدراسة، أي كيف يمكن استخدام الطرق المختلفة لمعادلة درجات الاختبارات.

الأهمية العملية:

- يمكن أن تسهم هذه الدراسة في توضيح طرق جديدة لمصممي ومحللي الاختبارات في وزارات التربية والتعليم والجامعات العربية المهتمين بموضوع إيجاد صور متكافئة للاختبارات.
- تعريف المختصين ومحللي الاختبارات على بعض البرامج الحاسوبية التي تساعدهم في معادلة درجات الصور المختلفة للاختبارات من خلال توفير معلومات وإرشادات للمهتمين في بناء الاختبارات حول كيفية توظيف البرامج الحاسوبية المستخدمة في معادلة الدرجات وتطبيقها بكل يسر وسهولة وخاصة برمجية (Equating Receips).

التعريفات الإجرائية

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.

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

البيانات متعددة الاستجابة: وهي فقرات اختباريه تكون الاستجابة عليها متعددة (1-2-3-4-5) حيث يكلف المفحوص بتحديد الاستجابة التي يراها مناسبة.

الدراسات السابقة

في مجال الحديث عن الدراسات السابقة لم يجد الباحث دراسات سابقة – في حدود علمه- وظفت الطريقتين مجتمعتين في معادلة درجات الاختبارات ؛ لذلك تم تناول الدراسات التي تناولت كل طريقة بصورة منفصلة، ومن الدراسات التي استخدمت طرق النظرية الحديثة في معادلة درجات الاختبارات فقد أجرى أيوب (1994) دراسة هدفت إلى المقارنة بين أربع طرق لمعادلة الاختبارات وهي الطريقة الخطية و الطريقة المئينية المنبثقتان عن النظرية التقليدية في القياس ونماذج النظرية الحديثة في القياس أحادي المعلمة وثنائي المعلمة عندما يكون التصميم (أ) مجموعات متطابقة أو عشوائية، وعندما يكون التصميم(ب) مجموعات غير متكافئة مع فقرات مشتركة. لتحقيق هدف الدراسة تم بناء ثلاثة اختبارات بصورتين (أ) و(ب) لمادة الرياضيات للصفوف الرابع والخامس والسادس.أشارت نتائج المعادلة الأفقية أن نماذج النظرية الحديثة في القياس كانت أكثر فاعلية من طريقتي المعادلة الخطية.كذلك أشارت نتائج المعادلة العمودية إلى أن الطريقة المئينية كانت أكثر فاعلية من الطرق الأخرى المستخدمة في الدراسة يليها النموذج الثنائي المعلمة ثم النموذج أحادي المعلمة.

كما اجرى لي وكولن وفرزبي (Lee, Kolen, Frezby, 2001) دراسة هدفت إلى المقارنة بين معادلة اختبارين أحدهما ثنائي التدريج والآخر متعدد التدريج وفق نماذج النظرية الحديثة والنظرية الكلاسيكية في المعادلة، حيث تم استخدام طرق المعادلة التقليدية الآتية لمعادلة الاختبار الثنائي التدريج وهي (الأوساط الخطية والمئينية)، بينما تم استخدام طرق المعادلة وفق النظرية الحديثة وهي طريقتي المعادلة التاليتين (GRM) و(NM) في معادلة الاختبار المتعدد التدريج وتم استخدام محك الجذر ألتربيعي للأوساط غير الموزونة للحكم على فاعلية الطريقة المستخدمة في المعادلة. أظهرت النتائج تشابهاً بين طرق معادلة العلامات الحقيقية والملاحظة في الاختبار المتعدد التدريج وفق نماذج النظرية الحديثة، والطرق الكلاسيكية المستخدمة في المعادلة (الأوساط الخطية والمئينية) في الاختبار ثنائي التدريج للعلامات الحقيقية.

كما أجرى الصمادي (2006) 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 (35) فقرة لكل منها أربعة بدائل من نوع فقرات الصواب والخطأ المتعدد في مبحث الرياضيات. تكونت عينة الدراسة من (873) طالبا وطالبة من الصف الأول الثانوي موزعين على عشرة مدارس تشمل تسعا وعشرين شعبة في تخصصات الأدبي والإدارة المعلوماتية والعلمية للعام الدراسي (2005/2004). وقد أظهرت النتائج أن طريقة التصحيح الرابعة –وهي إعطاء الطالب علامة واحدة لكل بديل تم الإجابة عليه بشكل صحيح – التي تعتمد على مراعاة المعرفة الجزئية، كانت الأكثر دقة في قياس قدرات الأفراد، وتعطي معلومات أكثر عن الاختبار، كما كانت الأكثر فاعلية في معادلة الاختبار.

كما أجرى يونغ وو (Young Woo,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=250-500-1000- 1680) للاختبارات الخمسة وتم حسابه باستخدام Bootstrap، وبأحجام عينات(500-1000-1680) للاختبارات الثلاثة. وقد أشارت النتائج بصورة عامة أن طريقة المعايرة المتلاقية أظهرت أخطاء معيارية وبواقي معيارية اقل من طريقة المعايرة المنفصلة، كذلك أظهرت النتائج أن معادلة الدرجات المشاهدة أنتجت أخطاء معادلة وبواقي معيارية اقل من معادلة الدرجات الحقيقية.

وأجرت آماندا (Amanda, 2008) دراسة هدفت من خلالها إلى مقارنة طرق المعادلة في النظرية الكلاسيكية ونظرية استجابة

الفقرة باستخدام تقييم عدد الدرجات الصحيحة أو إحراز الصيغة. 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 التي تم جمعها تحت تعليمات SAT وبيانات مولدة، فالبيانات الحقيقية استخدمت فيها بيانات إحراز الصيغة، بعد ذلك تم استخدام نماذج الاختيار من متعدد حتى يكمل الاستجابات المحذوفة للبيانات المولدة ليتبع تعليمات إحراز الدرجات الصحيحة. في البيانات المولدة تم تطبيق نموذج الاختيار من متعدد لتوليد البيانات بحيث تمثل كلا التطبيقين. أشارت النتائج أن الطريقة التي أنتجت اقل قيمة للتحيز في البيانات الحقيقية والمولدة هي طريقة توكر الخطية.

أما المدانات (2012) فقد أجرى دراسة هدفت إلى مقارنة فاعلية طريقتي معادلة العلامات الحقيقية والمشاهدة في معادلة الاختبارات باستخدام جذع مشترك ومجموعات غير متكافئة، حيث قام الباحث ببناء صورتين متكافئتين لاختبار في الفيزياء عدد فقرات كل منهما (20) فقرة، بالإضافة إلى (10) فقرات استخدمت كاختبار جذع مشترك. وقد تكونت عينة الدراسة من مجموعتين من الطلبة تم اختيارهما عشوائيا، حيث تقدمت المجموعة الأولى للصورة الأولى من الاختبار في الفصل الدراسي الأول، وتقدمت المجموعة الثانية للصورة الثانية من الاختبار في الفصل الدراسي الثاني وعدت مع المجموعة الأولى مجموعتين غير متكافئتين. تم استخدام طريقتين للمعادلة تتبعان النظرية الحديثة في القياس وهما طريقة معادلة العلامات الحقيقية وطريقة معادلة العلامات المشاهدة وأربعة أحجام من فقرات الجذع المشترك. أظهرت النتائج وجود فروق دالة احصائيا في متوسطات القيم المعادلة بطريقة المعادلة بالعلامات الحقيقية وطريقة المعادلة بالعلامات المشاهدة عندما كانت عدد فقرات الجذع المشترك (4،7،10) لصالح طريقة معادلة العلامات المشاهدة، ولكن الفروق لم تكن ذات دلالة عندما كان الجذع المشترك فقرة واحدة، كما أشارت النتائج إلى عدم وجود أثر لعدد فقرات الجذع المشترك باستخدام أي من الطريقتين.

أما في مجال الحديث عن الدراسات التي تناولت طريقة كيرنيل في معادلة درجات الاختبارات، فقد أجرى وانج (Wang, 2005) دراسة هدفت إلى مقارنة الطريقة اللوغارتميه الخطية المستمرة مع طريقتين في المعادلة: طريقة كيرنيل وطرق المعادلة المئينيه غير الممهدة، وقد تم استخدام تصميم المجموعات العشوائية في جمع بيانات اختبار حقيقي لمقارنة طرق المعادلة الثلاث. أشارت النتائج أن طريقة المعادلة اللوغارتمية الخطية المستمرة أنتجت توزيعا مستمرا أكثر تمهيدا، كذلك عندما تمت مقارنة الطريقتين: معادلة كير نيل اللوغارتميه (المستمرة) مع طريقة المعادلة المئينيه غير الممهدة وباستخدام نفس البيانات، أشارت النتائج أن الطرق الثلاث أظهرت نفس نتائج المعادلة حتى في مقياس الدرجات.

أما ماو (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 اكبر من الأخطاء المعيارية للمعادلة مقارنة مع الدرجات المنخفضة من التمهيد.

وأجرى آكور (Akour, 2006) دراسة هدفت إلى المقارنة بين ثلاث طرق لمعادلة الاختبارات باستخدام طرق المعادلة المئينيه وهي: طريقة ما قبل التمهيد، طريقة 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). وقد 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 وهي: حجم العينة ومستوى الصعوبة وطول الاختبار. أشارت نتائج الدراسة أن جميع طرق المعادلة كانت أكثر فاعلية من طرق قبل التمهيد.

كما أجرى كيو (Qu, 2007) دراسة هدفت إلى معرفة اثر البيانات المركبة من مجموعتين تحت تأثير اختلاف حجم العينات على دقة المعادلة. استخدم الباحث طريقتين في معادلة الدرجات وهي: طريقة كيرنيل في المعادلة التي تعتمد على المعادلة المئينيه والطريقة التقليدية في المعادلة. وقد طبق الباحث نوعين من البيانات: بيانات حقيقية، وبيانات مولده باستخدام برامج حاسوبيه، ثم قام الباحث بإيجاد كل من الخطأ المعياري للمعادلة (((SEE))، البواقي المعيارية، تحيز المعادلة، والخطأ المعياري

لاختلاف المعادلة ((SEED)) لكل طريقة من طرق المعادلة المستخدمة. أشارت النتائج أن طريقة المعادلة باستخدام البيانات المركبة من مجموعتين تكون أكثر حساسية للظروف المختلفة من البيانات غير المركبة.

كما أجرى المحروق (2011) دراسة هدفت إلى مقارنة طرق كيرنيل والمئينات وطرق نظرية استجابة الفقرة عند استخدام تصميم الفقرات المشتركة في دقة معادلة درجات الاختبارات متعددة الحدود تحت ظروف: البيانات متعددة الحدود، الفقرات المشتركة، حجم العينة، طول الاختبار، وتجانس مستوى الصعوبة. ولمقارنة طرق المعادلة الثلاث، تم توليد بيانات تجريبية باستخدام برمجية (Wingen2)، وذلك باستخدام ثلاثة متغيرات: طول الاختبار، حجم العينة، والتشابه في مستويات الصعوبة، 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، كذلك تم استخدام الخطأ المعياري للمعادلة والبواقي المعيارية لتقييم نتائج المعادلة. وقد أشارت النتائج إلى أن طريقة المعادلة باستخدام نظرية استجابة الفقرة كانت أكثر الطرق دقة، ثم تلتها طريقة كيرنيل، وعند مختلف الظروف، كذلك أشارت النتائج أن حجم العينات الكبير يقلل من الخطأ المعياري للمعادلة ويقلل كذلك من جذر متوسط مربعات الخطأ.

كما أجرى الدويري (2012) دراسة هدفت إلى مقارنة طرق كيرنيل وطريقة المئينات المتساوية لمعادلة صورتي اختبار في ضوء شكل التوزيع لبيانات مولدة، حيث تم توليد البيانات باستخدام برمجية (WINGEN)، حيث تم توليد ستة أزواج من نماذج الاختبارات بواقع (40) فقرة لكل نموذج حيث تم حساب معامل التمييز والصعوبة والتخمين لكل فقرة حسب النظرية الحديثة في القياس ووفقا للنموذج اللوجستي ثلاثي المعلمة، وقد تم معادلة درجات الاختبارات باستخدام الدرجات المشاهدة كمعيار لحساب معاملات الدقة المستخدمة في الدراسة وهي: الخطأ المعياري للمعادلة والتحيز والجذر التربيعي لمتوسط مربعات الأخطاء. وقد أشارت النتائج أنه لا يوجد فرق دال في متوسط قيم الخطا المعياري بين الطريقتين.

التعقيب على الدراسات السابقة:

من مجمل الدراسات السابقة يمكن القول أن هذه الدراسات تناولت معادلة الدرجات من جوانب متعددة، فمنها من تناول معادلة الدرجات باستخدام الاختبارات المشتركة الداخلية أو الخارجية مع التصاميم المختلفة لجمع المعلومات. ومنها ما تطرقت 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 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 من الفئات المختلفة، إضافة إلى اقتصار بعض الدراسات على طريقة واحدة أو اثنتين من طرق معادلة درجات الاختبار المختلفة. وقد لاحظ الباحث أيضاً أن الدراسات العربية رغم محدودية عددها لم تشمل مقارنة طريقتي الدرجات المشاهدة وطريقة كيرنيل باستخدام بيانات متعددة الاستجابة، حيث ركزت معظمها على معادلة درجات الاختبارات باستخدام بيانات ثنائية الاستجابة. ولعل هذه الدراسة تتميز عن الدراسات السابقة في كونها تناولت معادلة البيانات متعددة الاستجابة، كذلك تتميز هذه الدراسة بتناولها لمتغيرات عديدة مثل حجم العينة، وطول فقرات الاختبار للوقوف على أثرها في دقة معادلة الدرجات متعددة الحدود.

الطريقة والإجراءات

هدفت هذه الدراسة إلى الكشف عن فاعلية طريقة الدرجات المشاهدة وكيرنيل في معادلة درجات الاختبارات، وفيما يلي وصفا للمنهجية المتبعة في هذه الدراسة.

أولاً: تصميم الدراسة ومتغيراتها:

تكونت متغيرات الدراسة المستقلة من:1- طول الاختبار2- حجم العينة.

طول الاختبار: وقصد به عدد الفقرات في الاختبار الواحد بما فيها فقرات الجذع المشترك، وتكون من مستويين: 1- 30 فقرة. 2- 60 فقرة. **حجم العينة:** ويقصد به عدد الأفراد في العينة الواحدة، وتكونت من ثلاثة مستويات: 1- 200. 2- 600. 3- 1000.. وفي ضوء المتغيرات،كان تصميم الدراسة (2 X 3) وبعدد خلايا (6) خلايا، والجدول (1) يبين التصميم وخلاياه:

الجدول (1)

خلايا التصميم متضمنا المتغيرات المستقلة في الدراسة وعدد مرات التكرار في كل خلية

طول الاختبار	الدرجات المشاهدة			كيرنيل		
	SEE		RMSE	SEE		RMSE
	200	600	1000	200	600	1000
30	√	√	√	√	√	√
60	√	√	√	√	√	√

ثانيا: المتغيرات التابعة:

وهو معيار الحكم على دقة معادلة درجات الصور المختلفة للاختبار وهو الخطأ المعياري للمعادلة، والجذر التربيعي لمتوسط مجموع مربعات إنحرافات الخطأ.ويحسب الخطأ المعياري للمعادلة بالمعادلة الآتية:

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

حيث: n = عدد مرات توليد البيانات. y_k = تمثل الدرجات على النموذج y.

$e_x(y_k)$ =هي الدرجة المعادلة على النموذج x من خلال النموذج y_k.

$\hat{e}_x(y_k)$ = هي متوسط درجات المعادلة للدرجة y_k عبر مرات التكرارات.

أما طريقة 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 إن: K = عدد المفحوصين. j = رمز المفحوص. n_s = عدد العينات والمتمثلة في عدد مرات توليد البيانات. n_e = عدد المفحوصين في العينة الواحدة. y_j = الدرجة على الصورة الأولى للمفحوص j في ضوء درجته على الصورة الثانية. E_j = الدرجة الحقيقية المتوقعة للمفحوص على الصورة الأولى.

توليد بيانات الاستجابة على الفقرات ومعادلة الدرجات:

تم توليد البيانات ومعادلة الدرجات وتقييم دقة المعادلة وفق الخطوات الآتية:

1- توليد قدرات الأفراد:ولدت θ والممثلة لقدرات أفراد كل عينة باستخدام برمجية (WINGEN2)، حيث ولدت θ عندما يكون التوزيع توزيعا سويا بمتوسط صفر وانحراف معياري واحد.

2- **تقدير معالم الفقرات:**تم تقدير معالم الفقرات باستخدام نموذج الاستجابات المتعددة، أما معالم الفقرات التي تم تمثيلها فهي (الصعوبة والتمييز)، ولتحقيق هذا الغرض استخدمت برمجية WINGEN2.

3- **اختيار الفقرات:** تم تحديد فقرات كل صورة من صور الاختبار، كما حددت فقرات الجذع المشترك وبنسبة الثلث من عدد فقرات الاختبار الكلي.

4- **توليد استجابات المفحوصين.** ولدت استجابات كل مفحوص من إفراد العينة (200، 600، 1000) وذلك وفق مستويات كل متغير من المتغيرات المتمثلة في التصميم.حيث استخدمت قدرة المفحوص، كما تم توليدها سابقا ومعالم كل فقرة في نموذج الاستجابات المتعدد لتحديد احتمالية الإجابة عن الفقرة واستخدمت نفس البرمجية لتوليد الاستجابات.وبنفس الإجراءات تم توليد استجابات للمفحوصين على الصورة الثانية، وبالتالي تكون لكل مفحوص من إفراد العينة (200، 600، 1000) درجة كلية على الصورة الثانية من الاختبار ودرجة على فقرات الجذع المشترك.

5- تكرار الخطوة السابقة بتوليد الاستجابات للمفحوصين باستخدام معالم الفقرات ومعالم قدرات المفحوصين 200 مرة في كل خلية من خلايا التصميم.

كررت الخطوة السابقة وذلك باستخدام نفس البرمجية حيث كان يتم في كل مرة توليد قدرات الأفراد على التوزيع المطلوب، وباستخدام نفس معالم الفقرات لتحديد استجابات المفحوصين وفق النموذج المتعدد الاستجابات.

6- تخزين استجابات المفحوص في كل عينة وفي كل مرة تكرر فيها للصورة الأولى والثانية في الاختبار.

7- تم تحويل الملفات إلى ملفات نصية لصورتي الاختبار الأولى والثانية؛ وذلك من اجل التعامل مع البيانات، وتم إجراء المعادلة وفق الخطوات الآتية:

* تم حساب درجة كل مفحوص على الصورة الأولى من الاختبار ككل، ودرجته على فقرات الجذع المشترك.
* تم حساب درجة كل مفحوص على الصورة الثانية من الاختبار ككل، ودرجته على فقرات الجذع المشترك.
* تم حساب المتوسط الحسابي والانحراف المعياري لأفراد العينة،وذلك للدرجة الكلية وللدرجة على فقرات الجذع المشترك في الصورة الأولى، وكذلك الأمر في الصورة الثانية.

8- تم إنشاء ملف يحتوي على قيم (X) وما يقابلها من قيم (Ẏ) لكل مفحوص من المفحوصين، حيث يتم وفق هذا الملف حساب معياري تقدير دقة المعادلة والبواقي المعيارية.

9-حساب SEE,RMSE.

نتائج الدراسة

هدفت هذه الدراسة إلى الكشف عن فاعلية طريقة الدرجات المشاهدة وكيرنيل في معادلة درجات الاختبارات، وذلك في ظل ظروف تجريبية مختلفة وفيما ياتي عرضا لنتائج الدراسة وذلك بتناول كل متغير مستقل على حده، ثم تناول المتغيرات المستقلة مجتمعة:

أولاً: النتائج المتعلقة بالسؤال الأول: ما أثر حجم العينة وطول الاختبار في الخطأ المعياري للمعادلة عند النقاط المختلفة على سلم الدرجات؟

أ: النتائج المتعلقة بنماذج الاختبار الذي يتألف من 30 فقرة:

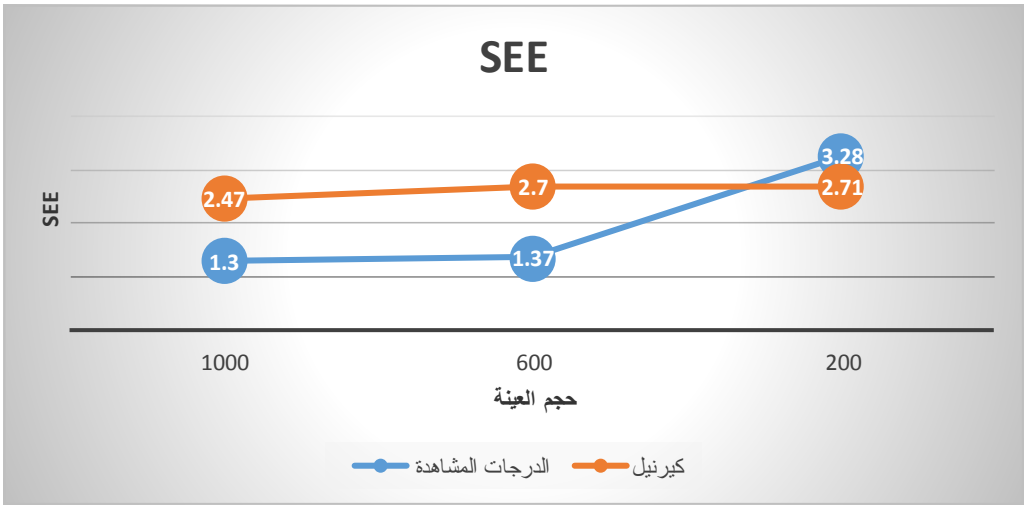
الخطأ المعياري للمعادلة: يوضح الجدول (2) قيم الخطا المعياري للمعادلة لطريقة الدرجات المشاهدة وكيرنيل في معادلة الدرجات.

الجدول (2)

الخطأ المعياري للمعادلة لطريقة الدرجات المشاهدة وكيرنيل في معادلة الدرجات للاختبار الذي يتالف من (30) فقرة

SEE	حجم العينة	200	600	1000
	الدرجات المشاهدة	3.28	1.37	1.30
	كيرنيل	2.71	2.70	2.47

يُظهر الجدول (2) أن قيم الخطأ المعياري للمعادلة لطريقة الدرجات المشاهدة في معادلة الدرجات للاختبار الذي يتألف من 30 فقرة قد تراوحت بين (1.30 -3.28)، حيث بلغت اقل قيمة للخطأ المعياري (1.30)، في حين بلغت أعلى قيمة (3.28)، كذلك يظهر الجدول ان قيمة الخطأ المعياري للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تكونت العينة من (1000) مفحوص، بلغت قيمة الخطأ المعياري (1.30)، وعندما بلغ حجم العينة (600) مفحوص، بلغت قيمة الخطأ المعياري (1.37)، وعندما بلغ حجم العينة (200) مفحوص، بلغت قيمة الخطأ المعياري (3.28). كما يظهر الجدول (2) أن قيم الخطأ المعياري للمعادلة لطريقة كيرنيل في معادلة الدرجات للاختبار الذي يتألف من 30 فقرة قد تراوحت بين (2.47-2.71)، حيث بلغت اقل قيمة للخطأ المعياري (2.47)، في حين بلغت أعلى قيمة (4.900)، كذلك يظهر الجدول ان قيمة الخطأ المعياري للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تكونت العينة من (1000) مفحوص، بلغت قيمة الخطأ المعياري (2.47)، وعندما بلغ حجم العينة (600) مفحوص، بلغت قيمة الخطأ المعياري (2.70)، وعندما بلغ حجم العينة (200) مفحوص، بلغت قيمة الخطأ المعياري (2.71). ويوضح الشكل رقم (1) قيم الخطأ المعياري للمعادلة لطرق المعادلة على سلم الدرجات وعبر اختلاف حجم العينات.



الشكل (1)

الخطا المعياري للمعادلة لطريقة الدرجات المشاهدة والمئينات في معادلة درجات الاختبار الذي يتألف من (30) فقرة

يوضح الشكل (1) أن طريقة الدرجات المشاهدة وعبر الأحجام المختلفة للعينات، كانت قيمة الخطأ المعياري للمعادلة فيها أقل قيمة، كما يلاحظ من الشكل (1) ان قيمة الخطا المعياري للمعادلة تنخفض مع ازدياد حجم العينة، وان المنحنيات تميل الى الانخفاض وتقترب من الصفرعندما ترتفع حجم العينة، وهذا يشير إلى أن ارتفاع حجم العينة يقلل من قيمة الخطأ المعياري للمعادلة (SEE).

ب- النتائج المتعلقة بنماذج الاختبار الذي يتألف من 60 فقرة:

يوضح الجدول (3) قيم الخطا المعياري للمعادله باستخدام طريقة الدرجات المشاهدة وكيرنيل:

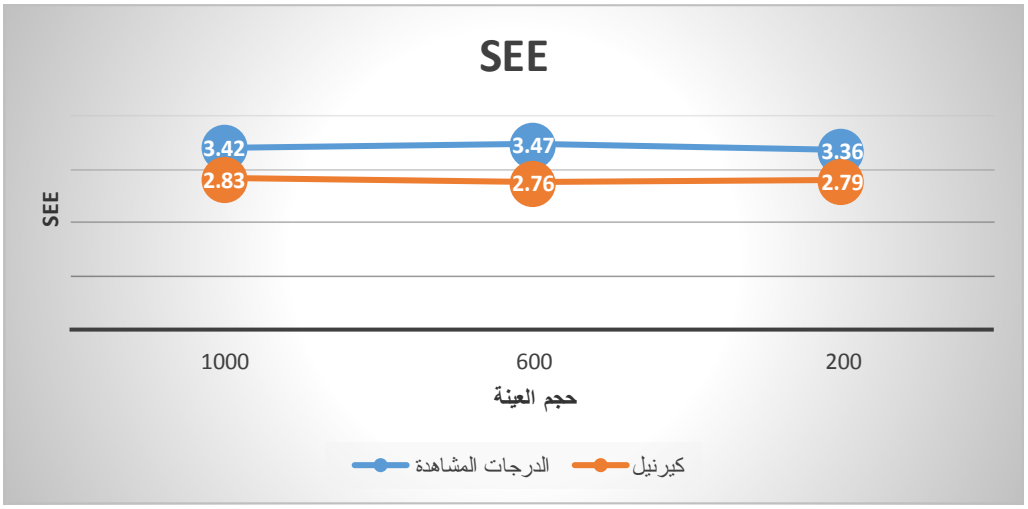
الجدول (3)

الخطأ المعياري للمعادلة لطريقة الدرجات المشاهدة وكيرنيل في معادلة درجات الاختبار الذي يتألف من (60) فقرة

SEE	حجم العينة	200	600	1000
	الدرجات المشاهدة	3.36	3.47	3.42
	كيرنيل	2.79	2.76	2.83

يظهر الجدول (3) أن قيم الخطأ المعياري للمعادلة لطريقة الدرجات المشاهدة في معادلة الدرجات للاختبار الذي يتألف من 60 فقرة قد تراوحت بين (3.36 -3.47)، حيث بلغت اقل قيمة للخطأ المعياري (3.36)، في حين بلغت أعلى قيمة (3.47)، كذلك يظهر الجدول ان قيمة الخطأ المعياري للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تكونت العينة من (1000) مفحوص، بلغت قيمة الخطأ المعياري (3.42)، وعندما بلغ حجم العينة (600) مفحوص، بلغت قيمة الخطأ المعياري (3.47)، وعندما بلغ حجم العينة (200) مفحوص، بلغت قيمة الخطأ المعياري (3.36). كما يظهر الجدول (3) أن قيم الخطأ المعياري للمعادلة لطريقة كيرنيل في معادلة الدرجات للاختبار الذي يتألف من 60 فقرة قد تراوحت بين (2.76 -2.83)، حيث بلغت اقل قيمة للخطأ المعياري (2.76)، في حين بلغت أعلى قيمة (2.83)، كذلك يظهر الجدول ان قيمة الخطأ المعياري للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تكونت العينة من (1000) مفحوص، بلغت قيمة الخطأ المعياري (2.83)، وعندما بلغ حجم العينة (600) مفحوص، بلغت قيمة الخطأ المعياري (2.76)، وعندما بلغ حجم العينة (200) مفحوص، بلغت قيمة الخطأ المعياري (2.79).

ويوضح الشكل رقم (2) قيم الخطأ المعياري للمعادلة لطرق المعادلة على سلم الدرجات وعبر اختلاف حجم العينات:



الشكل (2)

الخطأ المعياري للمعادلة لطريقة الدرجات المشاهدة وكيرنيل في معادلة الدرجات للاختبار الذي يتألف من (60) فقرة

يوضح الشكل (2) أن طريقة المئينات وعبر الأحجام المختلفة للعينات كانت قيمة الخطأ المعياري للمعادلة فيها اقل قيمة، ويلاحظ من الشكل (2) أن قيمة الخطأ المعياري للمعادلة تنخفض مع ازدياد حجم العينة.

ثانياً: النتائج المتعلقة بالسؤال الثاني: 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 30 فقرة:

يوضح الجدول (4) قيم جذر متوسط مربعات الفروق (RMSE) لطريقة الدرجات المشاهدة في وكيرنيل معادلة الدرجات.

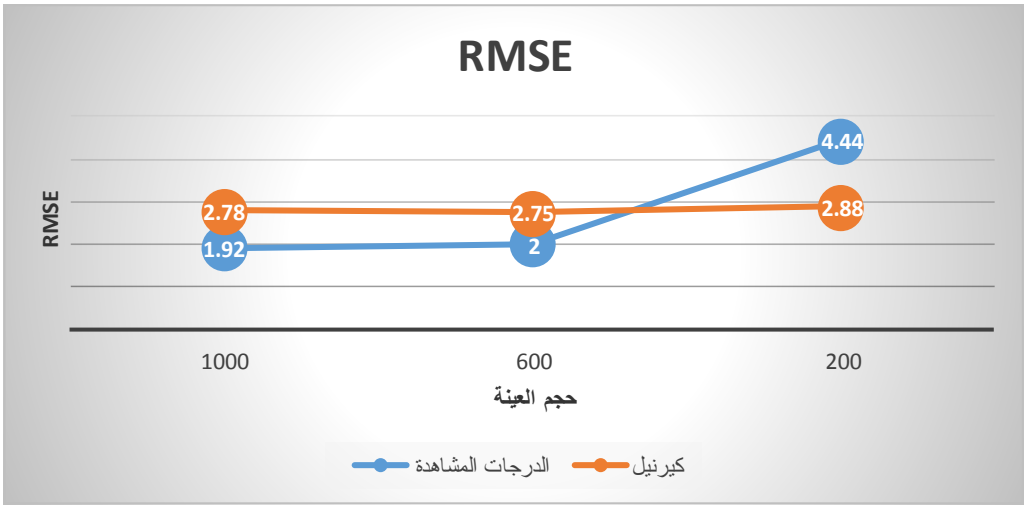
الجدول (4)

قيم (RMSE) لطرق المعادلة لطريقة الدرجات المشاهدة وكيرنيل في معادلة الدرجات الاختبار الذي يتألف من 30 فقرة

RMSE	حجم العينة	200	600	1000
	الدرجات المشاهدة	4.44	2.00	1.92
	كيرنيل	2.88	2.75	2.78

يظهر الجدول (4) أن قيم (RMSE) للمعادلة لطريقة الدرجات المشاهدة في معادلة الدرجات للاختبار الذي يتألف من 30 فقرة قد تراوحت بين (1.92 -4.44)، حيث بلغت اقل قيمة ل (RMSE) (1.92)، في حين بلغت أعلى قيمة (4.44)، كذلك يظهر الجدول ان قيم (RMSE)للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تكونت العينة من (1000) مفحوص، بلغت قيمة (RMSE) (1.92)، وعندما بلغ حجم العينة (600) مفحوص، بلغت قيمة (RMSE) (2.00)، وعندما بلغ حجم العينة (200) مفحوص، بلغت قيمة (RMSE) (4.44).

كما يظهر الجدول (4) أن قيم (RMSE)للمعادلة لطريقة كيرنيل في معادلة الدرجات للاختبار الذي يتألف من 30 فقرة قد تراوحت بين (2.75 -2.88)، حيث بلغت اقل قيمة ل (RMSE) (2.75)، في حين بلغت أعلى قيمة (4.93)، كذلك يظهر الجدول ان قيم (RMSE)للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تكونت العينة من (1000) مفحوص، بلغت قيمة (RMSE) (2.78)، وعندما بلغ حجم العينة (600) مفحوص، بلغت قيمة (RMSE) (2.75)، وعندما بلغ حجم العينة (200) مفحوص، بلغت قيمة (RMSE) (2.88). ويوضح الشكل رقم (3) قيم (RMSE) لطريقة الدرجات الملاحظة في معادلة الدرجات على سلم الدرجات وعبر اختلاف احجام العينات.



الشكل (3)

قيم RMSE للمعادلة لطريقة الدرجات المشاهدة وكيرنيل في معادلة الدرجات للاختبار الذي يتألف من (30) فقرة

يوضح الشكل (3) أن طريقة الدرجات المشاهدة وعبر الاحجام المختلفة للعينات كانت قيمة (RMSE) فيها أقل قيمة، في حين كانت أعلى قيمة عند استخدام طريقة كيرنيل في المعادلة وذلك عبر الاحجام المختلفة للعينات، ويلاحظ من الشكل (3) أن قيمة (RMSE)تنخفض مع ازدياد حجم العينة.

ب–النتائج المتعلقة بنماذج الاختبار الذي يتألف من 60 فقرة:

لنماذج الاختبار الذي يتالف من (60) فقره يوضح الجدول (5) قيم (RMSE) لطرق المعادلة:

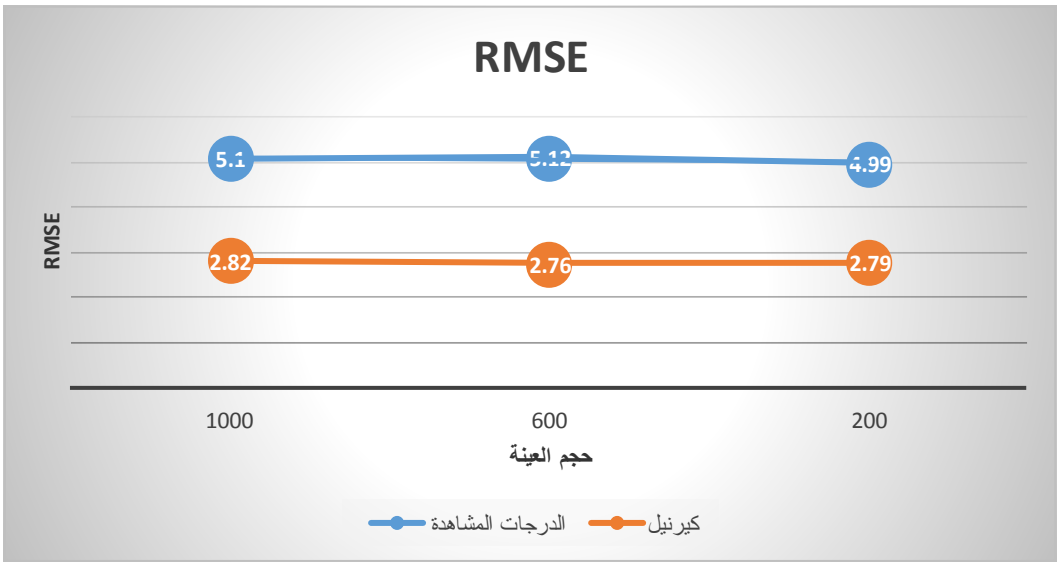
الجدول (5)

قيم (RMSE) لطريقة الدرجات المشاهدة وكيرنيل في معادلة الدرجات للاختبار الذي يتألف من (60) فقرة

RMSE	حجم العينة	200	600	1000
	الدرجات المشاهدة	4.99	5.12	5.10
	كيرنيل	2.79	2.76	2.82

يظهر الجدول (5) أن قيم (RMSE) للمعادلة لطريقة الدرجات المشاهدة في معادلة الدرجات للاختبار الذي يتألف من 60 فقرة قد تراوحت بين (4.99 –5.12)، حيث بلغت اقل قيمة ل (RMSE) (4.99)، في حين بلغت أعلى قيمة (5.12). كذلك يظهر الجدول (5) أن قيم (RMSE) للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تكونت العينة من (1000) مفحوص، بلغت قيمة (RMSE) (5.10)، وعندما بلغ حجم العينة (600) مفحوص، بلغت قيمة (RMSE) (5.12)، وعندما بلغ حجم العينة (200) مفحوص، بلغت قيمة (RMSE) (4.99).

كما يظهر الجدول (5) أن قيم (RMSE) للمعادلة لطريقة كيرنيل في معادلة الدرجات للاختبار الذي يتألف من 60 فقرة قد تراوحت بين (2.76 –2.82)، حيث بلغت اقل قيمة ل (RMSE) (2.76)، في حين بلغت أعلى قيمة (2.82)، كذلك يظهر الجدول ان قيم (RMSE) للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تكونت العينة من (1000) مفحوص، بلغت قيمة (RMSE) (2.82)، وعندما بلغ حجم العينة (600) مفحوص، بلغت قيمة (RMSE) (2.76)، وعندما بلغ حجم العينة (200) مفحوص، بلغت قيمة (RMSE) (2.79).و يوضح الشكل رقم (3) قيم (RMSE) لطريقة(IRT) في معادلة الدرجات على سلم الدرجات وعبر اختلاف احجام العينات، ويوضح الشكل رقم (4) قيم (RMSE)عند استخدام طريقة الدرجات المشاهدة وكيرنيل في معادلة الدرجات على سلم الدرجات وعبر الاحجام المختلفه للعينات (200–600–1000).



الشكل (4)

قيم RMSE للمعادلة لطريقة الدرجات المشاهدة وكيرنيل في معادلة الدرجات للاختبار الذي يتألف من (60) فقرة

يظهر الشكل رقم (4) أن طريقة الدرجات المشاهدة كانت اكثر دقه في معادلة الدرجات وعبر الأحجام المختلفه للعينات، بينما كانت طريقة كيرنيل أقل الطرق دقه في معادلة الدرجات وعبر الاحجام المختلفه للعينات، ويظهر الجدول ان قيم (RMSE) تتاثر بحجم العينه فكلما كبر حجم العينه قلة قيمة (RMSE) واذا نقص حجم العينة ارتفعت قيم (RMSE).

مناقشة النتائج والتوصيات

أولاً: مناقشة النتائج المتعلقة بالسؤال الأول: ما أثر حجم العينة وطول الاختبار في الخطأ المعياري للمعادلة عند النقاط المختلفة على سلم الدرجات؟

فيما يتعلق بنتائج الخطأ المعياري، أظهرت النتائج أن قيم الخطأ المعياري للمعادلة لطريقة الدرجات المشاهدة في معادلة الدرجات للاختبار الذي يتألف من 30 فقرة قد تراوحت بين (1.30 –3.28)، حيث بلغت اقل قيمة للخطأ المعياري (1.30)، في حين بلغت أعلى قيمة (3.28)، كما ان قيمة الخطأ المعياري للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تكونت العينة من (1000) مفحوص، بلغت قيمة الخطأ المعياري (1.30)، وعندما بلغ حجم العينة (600) مفحوص، بلغت قيمة الخطأ المعياري (1.37)، وعندما بلغ حجم العينة (200) مفحوص، بلغت قيمة الخطأ المعياري (3.28). وأظهرت النتائج أن قيم الخطأ المعياري للمعادلة لطريقة كيرنيل في معادلة الدرجات للاختبار الذي يتألف من 30 فقرة قد تراوحت بين (2.47–2.71)، حيث بلغت اقل قيمة للخطأ المعياري (2.47)، في حين بلغت أعلى قيمة (4.900)، كذلك يظهر الجدول ان قيمة الخطأ المعياري للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تكونت العينة من (1000) مفحوص، بلغت قيمة الخطأ المعياري (2.47)، وعندما بلغ حجم العينة (600) مفحوص، بلغت قيمة الخطأ المعياري (2.70)، وعندما بلغ حجم العينة (200) مفحوص، بلغت قيمة الخطأ المعياري (2.71).

كما أظهرت النتائج المتعلقة أن قيم الخطأ المعياري للمعادلة لطريقة الدرجات المشاهدة في معادلة الدرجات للاختبار الذي يتألف من 60 فقرة قد تراوحت بين (3.36 –3.47)، حيث بلغت اقل قيمة للخطأ المعياري (3.36)، في حين بلغت أعلى قيمة (3.47)، كذلك يظهر الجدول ان قيمة الخطأ المعياري للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تكونت العينة من (1000) مفحوص، بلغت قيمة الخطأ المعياري (3.42)، وعندما بلغ حجم العينة (600) مفحوص، بلغت قيمة الخطأ المعياري (3.47)، وعندما بلغ حجم العينة (200) مفحوص، بلغت قيمة الخطأ المعياري (3.36). كما أظهرت النتائج أن قيم الخطأ المعياري للمعادلة لطريقة كيرنيل في معادلة الدرجات للاختبار الذي يتألف من 60 فقرة قد تراوحت بين (2.76 –2.83)، حيث بلغت اقل قيمة للخطأ المعياري (2.76)، في حين بلغت أعلى قيمة (2.83)، كذلك يظهر الجدول ان قيمة الخطأ المعياري للمعادلة تنخفض مع

ازدياد حجم العينة، فعندما تكونت العينة من (1000) مفحوص، بلغت قيمة الخطأ المعياري (2.83)، وعندما بلغ حجم العينة (600) مفحوص، بلغت قيمة الخطأ المعياري (2.76)، وعندما بلغ حجم العينة (200) مفحوص، بلغت قيمة الخطأ المعياري (2.79).

إن قيمة الخطأ المعياري يتناسب عكسيا مع حجم العينة، حيث إنه بزداد بزيادة حجم العينة، وتقل قميته بنقصان حجم العينة، وهذا أمر طبيعي؛ إذ إن كمية المعلومات عند اي 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 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 (الصمادي، 2006)، التي توصلت الى أن دقة تقدير الخطأ المعياري في المعادلة، والبواقي المعيارية كان أفضل في العينات ذات الحجم الكبير، وهذا ما توصلت اليه هذه الدراسة.

ثانياً: 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) للمعادلة لطريقة الدرجات المشاهدة في معادلة الدرجات للاختبار الذي يتألف من 30 فقرة قد تراوحت بين (1.92–4.44)، حيث بلغت اقل قيمة ل (RMSE) (1.92)، في حين بلغت أعلى قيمة (4.44)، كذلك يظهر الجدول ان قيم (RMSE)للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تكونت العينة من (1000) مفحوص، بلغت قيمة (RMSE) (1.92)، وعندما بلغ حجم العينة (600) مفحوص، بلغت قيمة (RMSE) (2.00)، وعندما بلغ حجم العينة (200) مفحوص، بلغت قيمة (RMSE) (4.44).

كما أظهرت النتائج أن قيم (RMSE) للمعادلة لطريقة كيرنيل في معادلة الدرجات للاختبار الذي يتألف من 30 فقرة قد تراوحت بين (2.75–2.88)، حيث بلغت اقل قيمة ل (RMSE) (2.75)، في حين بلغت أعلى قيمة (4.937)، كما ان قيم (RMSE)للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تكونت العينة من (1000) مفحوص، بلغت قيمة (RMSE) (2.78)، وعندما بلغ حجم العينة (600) مفحوص، بلغت قيمة (RMSE) (2.75)، وعندما بلغ حجم العينة (200) مفحوص، بلغت قيمة (RMSE) (2.88).أما فيما يتعلق بنتائج (RMSE).

وقد أظهرت النتائج أن قيم(RMSE) للمعادلة لطريقة الدرجات المشاهدة في معادلة الدرجات للاختبار الذي يتألف من 60 فقرة قد تراوحت بين (4.99–5.12)، حيث بلغت اقل قيمة ل (RMSE) (4.99)، في حين بلغت أعلى قيمة (5.12). كذلك يظهر الجدول ان قيم (RMSE)للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تكونت العينة من (1000) مفحوص، بلغت قيمة (RMSE) (5.10)، وعندما بلغ حجم العينة (600) مفحوص، بلغت قيمة (RMSE) (5.12)، وعندما بلغ حجم العينة (200) مفحوص، بلغت قيمة (RMSE) (4.99).كما يظهر الجدول (5) أن قيم (RMSE)للمعادلة لطريقة كيرنيل في معادلة الدرجات للاختبار الذي يتألف من 60 فقرة قد تراوحت بين (2.76–2.82)، حيث بلغت اقل قيمة ل (RMSE) (2.76)، في حين بلغت أعلى قيمة (2.82)، كذلك يظهر الجدول ان قيم (RMSE) للمعادلة تنخفض مع ازدياد حجم العينة، فعندما تكونت العينة من (1000) مفحوص، بلغت قيمة (RMSE) (2.82)، وعندما بلغ حجم العينة (600) مفحوص، بلغت قيمة (RMSE) (2.76)، وعندما بلغ حجم العينة (200) مفحوص، بلغت قيمة (RMSE) (2.79) ويوضح الشكل رقم (3) قيم (RMSE) لطريقة (IRT) في معادلة الدرجات على سلم الدرجات وعبر اختلاف احجام العينات.

يتضح من النتائج السابقة أن قيمة الخطا المعياري والبواقي المعيارية تتناقص تدريجيا بتزايد القدرة حتى تصل إلى أقصى قيمة ممكنة لها عندما تكون قيمة القدرة صفرا لوجيتا، وهي تقابل متوسط الصعوبة للمفردات صفر لوجيت، إذ تكون قيمة الخطأ المعياري والبواقي المعيارية صغيرة كلما اقترب مستوى القدرة للأفراد من مستوى الصعوبة للمفردات، وبالتالي تزداد كمية المعلومات بنقصان الخطأ المعياري وبالتالي البواقي المعيارية.

مقارنة نتائج طرق المعادلة من خلال حجم العينة:

الخطأ المعياري للمعادلة: أظهرت نتائج الدراسة تحت ظرف طول الاختبار واختلاف حجم العينة، ان نتائج طريقة المعادلة باستخدام الدرجات الملاحظة قد أظهرت قيما منخفضة للخطأ المعياري مقارنة مع طريقة المعادلة باستخدام كيرنيل وعبر ألاحجام 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.

(RMSE): تحت كل ظرف من طول الاختبار واختلاف حجم العينة، فان النتائج اظهرت تقارب نتائج (RMSE) مع نتائج (SEE) عندما ترتفع حجم العينة. لقد اظهرت النتائج ان نتائج قيم (RMSE) لكل طريقة يتاثر وبدرجة كبيرة بحجم العينة، فالعينات الكبيرة تنتج قيما صغيرة ل (RMSE)، والعينات الصغيرة تنتج قيما كبيرة لقيم (RMSE)، كما اظهرت النتائج ان قيم (RMSE) تقل في الطرق المختلفة مع ازدياد حجم العينات. كذلك اظهرت قيم (RMSE) الناتجة عن تطبيق طرق المعادلة،ان طريقة الدرجات المشاهدة في معادلة الدرجات كانت اكثر دقة في معادلة الدرجات على سلم الدرجات وعبر الاحجام المختلفة للعينات من طريقة كيرنيل.

إن قيمة الخطأ المعياري والبواقي المعيارية تناسب عكسيا مع حجم العينة، حيث إنه بزداد بزيادة حجم العينة، وتقل قميته بنقصان حجم العينة، وهذا أمر طبيعي؛ إذ إن كمية المعلومات عند اي مستوى من مستويات القدرة تتناسب عكسيا مع الخطأ المعياري والبواقي المعيارية، فعند زيادة حجم العينة يقترب متوسط معلمة التخمين من الصفر وهذا يؤدي إلى التقليل من قيمة الخطأ المعياري والبواقي المعيارية، حيث إن قيمة الخطأ المعياري تقل كلما قلت قمية 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.

مقارنة نتائج طرق المعادلة من خلال طول الاختبار:

تحت الظروف المختلفة لحجم العينة، أظهرت النتائج أن المعادلة باستخدام طريقة الدرجات المشاهدة اظهرت قيما منخفضة للخطا المعياري مقارنة مع طريقة كيرنيل وعبر الاحجام المختلفة للعينات. كما أظهرت النتائج أن طول الاختبار يوثر في الخطأ المعياري للمعادلة، فالاختبار الطويل يعطي قيما كبيرة للخطا المعياري للمعادلة، على حين ان تقليل طول الاختبار من (60) الى (30) يؤدي الى تخفيض كبير للخطا المعياري للمعادلة. فعلى سبيل المثال في طريقة الدرجات المشاهدة عندما كان الخطا المعياري (3,42) للاختبار الذي يتكون من (60) فقرة وعند حجم عينة (1000)، فإن هذا الخطأ المعياري ينخفض الى (1.30) للاختبار الذي يتكون من (30) فقرة وبنفس حجم العينة. إن الاختلاف في قيم الخطأ المعياري للمعادلة يصبح أقل عندما ينخفض طول الاختبار، أن طريقة الدرجات المشاهدة كانت أكثر دقة في هذه الدراسة مقارنة مع طريقة كيرنيل عند استخدام معيار الخطأ المعياري للمعادلة.

وفيما يتعلق بنتائج قيم (RMSE)، فإن نتائج قيم (RMSE) تشابه مع نتائج (SEE)للطرق المستخدمة في هذه الدراسة وعبر الظروف المختلفة من حجم العينة، حيث دلت نتائج قيم (RMSE) الناتجة عن استخدام طرق المعادلة ان طريقة الدرجات المشاهدة اظهرت قيما منخفضة ل (RMSE) مقارنة مع طريقة المئينات. وان زيادة طول الاختبار يؤدي الى ارتفاع قيم (RMSE). فالاختلاف في قيم (RMSE) يقل مع زيادة طول الاختبار.

وعليه، وبناء على النتائج السابقة المتعلقة بالخطا المعياري والبواقي المعيارية وعبر الاختلاف في احجام العينات وطول الاختبار تعتبر طريقة الدرجات المشاهدة اكثر دقة في معادلة الدرجات مقارنة مع طريقة كيرنيل في معادلة الدرجات على سلم الدرجات وعبر الاحجام المختلفة للعينات.

ويمكن تفسيرهذه النتيجة في عاملين رئيسين:

أ–ان قيم RMSE، SEE المستخدمة هنا ليست قيما معيارية (Standardized)، لأن فقرات الاختبارات لم تقسم باستخدام تطابق الانحراف المعياري للعلامات المشاهدة؛ ولأن الانحراف المعياري للدرجات الصحيحة يرتفع مع زيادة طول الاختبار، فانه يتوقع أن القيم المعيارية لقيم RMSE،SEE سوف تظهر ميلا قليلا إلى الارتفاع مع ازدياد طول الاختبار.

ب–ان 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. وهذه النتيجة لا تتفق مع نتائج دراسة (الصمادي، 2006) التي أشارت الى ان طول الاختبار يؤدي الى نقصان قيمة الخطا المعياري.

وفي مجال مقارنة نتائج هذه الدراسة مع الدراسات الأخرى، فقد إتفقت نتائج هذه الدراسة مع دراسة يونغ وو (Youngwoo, 2007) التي أشارت إلى ان طريقة IRT كانت أكثر دقة في معادلة الدرجات من طرق المئنات. في حين نجد أن نتائج هذه الدراسة قد اختلفت مع نتائج دراسة أيوب (1994) التي أشارت إلى أن طريقة المئينات كانت أكثر دقة في معادلة الدرجات من طريقة IRT، كما اختلفت هذه الدراسة مع نتائج دراسة أماندا (Amanda, 2008) التي أظهرت أن طريقة الميئنات في المعادلة كانت أكثر

دقة من طريقة الدرجات المشاهدة.

التوصيات:

1.استخدمت هذه الدراسة نموذج الإستجابات المتعدد لتوليد البيانات لإجراء المعادلة، لذلك يوصي الباحث باستخدام نماذج أخرى من نماذج نظرية استجابة الفقرة للقيام بالمعادلة.

2.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.

3.يوصي الباحث بتطبيق طريقة المئينات و IRT على بيانات حقيقية متعددة الاستجابة وذلك للوقوف على مقدار الإختلاف في المعادلة بين البيانات الحقيقية والبيانات التجريبية.

المراجع

أولاً: المراجع العربية:

أيوب، ح. (1994)، المقارنة بين أربع طرق للمعادلة عندما يكون تصميم من مجموعات متكافئة وغير متكافئة، أطروحة دكتوراه غير منشورة، الجامعة الأردنية،عمان، الأردن.

الدوسري، ر. (2004)، القياس والتقويم التربوي الحديث: مباديء وتطبيقات وقضايا معاصرة. دار الفكر، عمان.

الدويري، م. (2012)، مقارنة طرق كيرنيل وطريقة المئينات المتساوية لمعادلة صورتي اختبار في ضوء شكل التوزيع لبيانات مولدة، أطروحة دكتوراه غير منشوره، جامعة اليرموك، عمان، الأردن.

الصمادي، إ. (2006)، أثر طريقة اختيار الفقرات في اختبار الجذع المشترك على دقة معادلة اختبار متعدد المستوى في الرياضيات للمرحلة الأساسية في الأردن، أطروحة دكتوراه غير منشورة، جامعة عمان العربية، عمان، الأردن.

المحروق، ي. (2011)، مقارنة طرق كيرنيل والمئينات وطرق نظرية استجابة الفقرة عند استخدام تصميم الفقرات المشتركة في دقة معادلة درجات الاختبارات متعددة الحدود، أطروحة دكتوراه غير منشوره، الجامعة الأردنية، عمان، الأردن.

المدانات، ر. (2008)، أثر طريقة المعادلة باستخدام جذع مشترك وعدد فقراته وحجم العينة على القيم المعادلة والخطأ في المعادلة بين صورتي اختبار في الفيزياء، أطروحة دكتوراه غير منشورة، جامعة عمان العربية، عمان، الأردن.

المدانات، ر. (2012)، مقارنة فاعلية طريقتي معادلة العلامات الحقيقية والمشاهدة في معادلة الاختبارات باستخدام جذع مشترك ومجموعات غير متكافئة، مجلة العلوم التربوية والنفسية، المجلد 13 العدد (2)، جامعة البحرين.

ثانياً: المراجع الأجنبية:

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

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

Alina A, Von Davier, Paul W. Holland, Dorothy T.Tayer. (2004).The Kernel method of test equating. New York: Springer – Verlag.

Angoff,W.H. (1971). Scales, norms, and equivalent scores.InR L.Thorndike (Ed), Educational measurement (2nd ed.) (pp.508-600).Washington: American Council on Education.

Baker, F.B, and Al-Karni, A.(1991).A comparison of two procedures for computing IRT equating coefficients. Journal of Educationa Measurement, 28, 147-162.

Cook, L.L & Eignor, D.R. (1991). An NCME instructional module on IRTequating methods.

Educational Measurement: Issue and Practice, 10,37-45.

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

Hambelton, R.K., Swaminathan, H., and Rogers, H.J. (1991). Fundamentals of item response theory. New York: Sage Publications.

Hambleton, Linden. (1985). Advances in Item Response Theory and Applications: An Introduction. Applied Psychological Measurement; 6, 373-378.

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

Hue, G. (2004). The effect of different anchor tests on the accuracy of test equating for test adaptation. Unpublished Doctoral Dissertation, Ohio University, USA.

Kolen, M. 1 and Brennan, R. L. (2004). Test equating, scaling, and linking and practices, (2nd ed). New York: Springer- Verlag.

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

Lee, G.Hill, Kolen, M. J., Frisbee, D. and Ankenmann, R. (2001). Comparison of dichotomous and polytomous item Response models in equating scores from tests composed of test lets. Applied Psychological Measurement, 25 (4), 357-372.

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, Yanxuan. (2007). The Effect Of Weighting In Kernel Equating Using Counter-Balanced Designs. Unpublished Doctoral Dissertation, Michigan State University, USA.

Wang, T. (2005a). An alternative continuization 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.

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

Effectiveness of Observed Scores and Kernel Method in Equating Test Scores

Yousef A. Almahrouk *

ABSTRACT

This study explored the Effectiveness of observed scores and kernel method in equating test scores through studying variables of sample siz and the length of the test. This study used Simulation data. The results indicated that the large sample size reduced the standard error of the equating and reduces residuals. Theys also indicated that the test length effected the standard error, the long test gave large standard error, and the residuals standard. It was also revealed that the the methods of equating using observed scores was more accurate than Kernel method.

Keywords: Equating test, NEAT design, Polynomial Scores.

* Ministry of Education, Kingdom of Bahrain. Received on 14/01/2016 and Accepted for Publication on 13/01/2017.