كلية التربية
College of Education
QATAR UNIVERSITY جامعة قطر

دار نشر جامعة قطر
Qatar University Press

OPEN ACCESS
Submitted: 8/1/2019
Accepted: 27/3/2019

معادلة كتيبات اختبار الدراسة الدولية لقياس مهارات القراءة (PIRLS 2011) بسلطنة عمان باستخدام نظرية الاستجابة للمفردة

أمل بنت خميس بن عبدالله الزعابي
وزارة التربية والتعليم، سلطنة عمان
amal.zaabe@moe.om

راشد بن سيف المحرزي
جامعة السلطان قابوس، سلطنة عمان
mehrzi@squ.edu.om

عبد الحميد حسن
جامعة السلطان قابوس، سلطنة عمان
abhshk@squ.edu.om

ملخص

هدفت الدراسة إلى المقارنة بين طريقة معادلة الدرجات الحقيقية، وطريقة معادلة الدرجات المشاهدة في إجراء معادلة كُتيبات الاختبارات المكونة من فقرات مختلطة: (ثنائية الاستجابة ومتعددة الاستجابة)، باستخدام تصميم المجموعة الواحدة، وتصميم المجموعات غير المتكافئة ذات المفردات المشتركة. استخدمت الدراسة اختبار قياس مهارات التقدم في القراءة (PIRLS) لعام 2011، والذي يتكون من 146 فقرة اختبارية (منها 74 فقرة ثنائية الاستجابة، و72 فقرة متعددة الاستجابة)، تم تقسيمها إلى 13 كُتيبًا بحيث يشمل كل كُتيب على نص أدبي ونص علمي، ثم تم توزيعها على 13 مجموعة، من طلبة الصف الرابع الأساسي بسلطنة عمان، بعينة إجمالية بلغت 10394 طالبًا وطالبة. تم استخدام برنامج MULTILOG 7.03، لتقدير معالم المفردات الاختبارية، وقدرات الأفراد، باستخدام نموذج الاستجابة المتدرجة، ووضعها على تدريج مشترك، ثم تم تطبيق طريقتين للمعادلة تتبعان النظرية الحديثة في القياس: (طريقة معادلة الدرجات الحقيقية، وطريقة معادلة الدرجات المشاهدة)، باستخدام برنامج Poly Equate (V0.5). أوضحت النتائج وجود فروق بين الإحصاءات الوصفية لكتيبات اختبار PIRLS 2011 قبل المعادلة، وتقاربها بعد إجراء المعادلة بالطريقتين. وأوضحت نتائج اختبار "ت" للعينات المرتبطة وجود فروق ذات دلالة إحصائية عند مستوى دلالة (0.05)، بين متوسطات القيم المعادلة بطريقتي الدرجات الحقيقية، والدرجات المشاهدة لصالح المعادلة بالدرجات المشاهدة في جميع الكتيبات، ما عدا الكُتيب (10)، والتي كانت لصالح المعادلة بالدرجات الحقيقية، والكتيبات (4)، و (5)، والتي أظهرت عدم وجود فروق دالة إحصائيًا بين طريقتي المعادلة.

الكلمات المفتاحية: اختبار PIRLS، معادلة الاختبار، معادلة الدرجات الحقيقية، معادلة الدرجات المشاهدة، سلطنة عمان

للاقتباس: الزعابي أ.، المحرزي ر. وحسن ع. الحـ.، "معادلة كتيبات اختبار الدراسة الدولية لقياس مهارات القراءة (PIRLS 2011) بسلطنة عمان باستخدام نظرية الاستجابة للمفردة"، مجلة العلوم التربوية، العدد 15، 2020

https://doi.org/10.29117/jes.2020.0022

دار نشر جامعة قطر
Qatar University Press

كلية التربية
College of Education
QATAR UNIVERSITY جامعة قطر

OPEN ACCESS
Submitted: 8/1/2019
Accepted: 27/3/2019

Equating Test Forms for Progress in International Reading Literacy Study (PIRLS 2011) in Sultanate of Oman Using Item Response Theory

Amal khamis Abdullah Alzaabi
Ministry of Education, Sultanate of Oman
amal.zaabe@moe.om

Rashid Saif Al-mehrzi
Sultan Qaboos University, Sultanate of Oman
mehrzi@squ.edu.om

Abdulhameed Hassan
Sultan Qaboos University, Sultanate of Oman
abhshk@squ.edu.om

Abstract

The study aims to compare between true score of mixed item formats (dichotomous and polytomous items), using single group design and nonequivalent group with common items design. The study used PIRLS test, which consists of 146 items, (74 multiple-choice items and 72 short answer items), distributed over 13 booklets. Each booklet has common items with some other booklets and some uncommon items, as well as two passages: reading for literary experience and reading to acquire and use information. The 13 booklets were distributed to a random sample of 10394 students from grade 4 in the Sultanate of Oman. MULTILOG 7.03 software was used to estimate item and ability parameters using the graded response model and Poly Equate (V0.5) software. The 13 booklets were equated using true score equating and observed score equating. The results showed differences in the descriptive characteristics for the 13 booklets before equating and reduction in these differences after equating with both equating methods. Paired samples t-test showed statistical significant differences between the two equating methods with all booklets towards the observed score equating except with booklet 10 where there are differences towards true score equating and booklets 4 and 5 where there is no statistical difference between the two equating methods.

Keywords: PIRLS test; Test equating; True score equating; Observed score equating; Oman

للاقتباس: الزعابي أ.، المحرزي ر. وحسن ع. الح.، "معادلة كتيبات اختبار الدراسة الدولية لقياس مهارات القراءة (PIRLS 2011) بسلطنة عمان باستخدام نظرية الاستجابة للمفردة"، مجلة العلوم التربوية، العدد 15، 2020

https://doi.org/10.29117/jes.2020.0022

مقدمة

للاختبارات أهمية كبيرة في حياة الطالب، حيث تلعب دورًا أساسيًا في جميع مراحل حياته التعليمية، وتعد الوسيلة الأساسية في تقويم العملية التعليمية التربوية؛ وذلك لأنها تساعد على اتخاذ العديد من القرارات، التي من شانها أن تحدد مستقبل هذا الطالب، (عودة وعبيدات، 2013). لذلك لا بد من بناء هذه الاختبارات بإشراف من مختصين في القياس، وأن تقيس فعلًا ما أعدت لقياسه (Branberg, 2010)؛ نظرًا لأهمية استخدام نتائج هذه الاختبارات في التعرف إلى مدى تحقيق الطالب للأهداف التعليمية، وجوانب القوة والقصور لديه، وتقديم معلومات دقيقة عن مستواه.

فهناك توجه عالمي إلى تنويع البرامج الاختبارية، والسماح للطالب بإعادة الاختبار أكثر من مرة، وكذلك استخدام الاختبارات في التقويم البنائي، الذي يقوم على عقد اختبارات دورية أثناء العام الدراسي للطالب، ومقارنة درجاته التي يحصل عليها عبر الزمن. وتتطلب هذه البرامج الاختبارية توافر نُسخ ونُسخ متعددة من الاختبار، تكون مناسبة للأهداف التربوية المعاصرة، وتقلل من القصور الملاحظ في أدوات القياس المستخدمة في تقويم التحصيل الدراسي لدى الطلاب، (الوليلي، 2005).

كما أن تحقيق السرية في نُسخ الاختبارات، تتطلب أن تتكون هذه النماذج من مفردات مختلفة، مما يجعل هذه الاختبارات غير قابلة للمقارنة؛ لذلك من الضروري أن تتصدى وزارات التربية والتعليم في الدول العربية لهذه المشكلة، عن طريق تحليل الامتحانات العامة، واجراء معادلة للنُّسخ المختلفة، ووضعها في بنوك أسئلة، حيث إن عمليات المعادلة تسمح بمقارنة الدرجات التي نحصل عليها من نُسخ الاختبارات مختلفة (Battauz, 2015).

وعلى الرغم من أن موضوع معادلة الاختبارات (Test Equating)، بدأ الاهتمام به متأخرًا من قبل المختصين في القياس والتقويم خلال السنوات المنصرمة، لكونه يشكل جزءًا هامًا من مهماتهم البحثية والمهنية، إلا أنهم بحثوا فيه بشكل مستفيض اعترافًا بأهميته في بناء الاختبارات. ويمكن أن يُعزى هذا الاهتمام المتزايد بموضوع معادلة الاختبارات إلى ثلاثة تطورات مهمة خلال الخمس عشرة سنة الماضية، لخصها الدوسري (2001): حيث يتعلق التطور الأول بزيادة عدد برامج الاختبارات وتنوعها، مما أدى إلى ضرورة معادلة الاختبارات التي تقيس السمة نفسها، ويتعلق التطور الثاني بمطوري الاختبارات، الذين أدركوا أهمية معادلتها لمعالجة الكثير من القضايا الفنية التي يثيرها النقاد عند تحليل درجاتها، أما التطور الثالث فيتعلق بمسألة المساءلة والمحاسبة في التربية، وقضية قدرة الاختبارات على العدل في قياس السمة لمجموعة من الأفراد المتباينة في الجنس والأصل العرقي.

وعلى الرغم من الاعتماد على مفاهيم النظرية الكلاسيكية، واستخدام الطرق والأساليب المتعلقة بها لسنوات طويلة، في تحقيق المعادلة بين الدرجات الناتجة من النُّسخ (الصور) الاختبارية؛ إلا أنه

تبين قصور هذه النظرية في مواجهة الكثير من المشاكل السيكومترية المتعلقة بهذه العملية، حيث أشار الوليلي (2005)، وجو لين (Ju Lin, 2008)، إلى أنه لا يمكن المقارنة بين مجموعتين من الأفراد على اختبارين مختلفين عن طريق الدرجات الخام، كما أن استخدام الدرجات المعيارية (Z scores)، للتغلب على هذه المشكلة لا تقدم علاجًا ناجعًا لها، حيث يتطلب ذلك اشتقاق عينتين متكافئتين من المجتمع نفسه، وهذا لا يحدث في الواقع.

لذلك تضافرت الجهود البحثية للتغلب على تلك المشكلات، وقد أسفرت الجهود عن إحداث اتجاهات معاصرة في القياس والتقويم التربوي، من بينها نظرية الاستجابة للمفردة، والتي استخدمت العديد من الأساليب والطرق لإجراء عمليات المعادلة، والتي تفوقت على نظرية القياس التقليدية؛ حيث قدمت قواعد جديدة للباحثين عن معادلة الأدوات والقياسات. كما قدمت العديد من الطرق، للتأكد من تكافؤ قياس الأدوات المختلفة المستخدمة في المقارنة بين الأفراد؛ وتتميز بأنها لا تختبر فقط شروط المعادلة، وإنما تقدم أيضًا أسباب عدم تحقق المعادلة، وهل يرجع إلى معاملات الصعوبة أم إلى القدرة التمييزية للمفردة؟ كما تقدم مؤشرات يمكن من خلالها التحقق من المعادلة على مستوى الفقرة، أو على مستوى الأبعاد التي يتكون منها الاختبار، (محمد، 2006).

أسهمت معادلة الاختبارات عن طريق نظرية الاستجابة للمفردة، في حل الكثير من المشكلات التي عجزت عن حلها النظرية التقليدية في الاختبارات، حيث يشير "بيكر"، والقرني (Baker & Alkarni, 1991)، إلى أنه من الإسهامات الكبيرة للنظرية الحديثة في القياس قدرتها على وضع عدة اختبارات، وعدّة مفحوصين على تدريج مشترك "Common Scale" في عملية القياس، وإمكانية استخدامها في المعادلة الأفقية والرأسية للاختبار. كما أن من أهم ميزات استخدام النظرية الحديثة في مجال معادلة الاختبارات، احتمال تقليل التحيز، أو التذبذب لميزان الدرجات الذي يمكن أن يحدث عندما تستخدم الأساليب الكلاسيكية، في إجراء معادلة الاختبارات عبر الزمن، وبخاصة إذا كانت العينة، التي تم تطبيق نسختي الاختبار عليها، تم اختيارها بطريقة غير عشوائية، (علام، 2005).

وتتوفر ثلاث طرق رئيسة لمعادلة الاختبارات باستخدام نظرية الاستجابة للمفردة ذكرها الدوسري (2001) و" Hambleton, Swaminathan & Rogers (1991)"، وهي على النحو التالي:

1- معادلة الاختبار باستخدام درجات القدرة (السمة) (Ability-Score Equating)، في هذه الطريقة نحتاج إلى معادلة درجات القدرة أثناء معايرة الاختبارات (تقدير معلمات المفردات كالصعوبة والتمييز)، وينتج عن ذلك علاقة خطية بين تدريجات القدرة في نُسخ الاختبار بعد تقدير معلماتها، وتقدير معلمات المفردات.

2- معادلة الاختبارات باستخدام الدرجات الحقيقية (True-Score Equating)، في النظرية الحديثة في القياس النفسي والتربوي، تصاغ الدرجة الحقيقية للمفحوص رياضيًا على أنها مجموع الدرجة المتوقعة للطالب على جميع مفردات الاختبار، $\sum_{i=1}^{n} p(\theta)$. فلو افترضنا أن نُسختي الاختبار تقيس السمة نفسها، وقد تم وضع معالم الفقرات لكلا النُّسختين على التدريج نفسه، فإن الدرجة الحقيقية (*Tx*) للطالب في الاختبار (*X*)، تكون هي: (*Ty*) في الاختبار (*Y*) ويجب أن يكون تدريج القياسات للدرجات مستقلًا عن مجموعة المفحوصين في الاختبار، وأيضًا يكون تدريج قياس درجات القدرة مستقلًا عن عدد مفردات الاختبارين.

3- معادلة الاختبارات باستخدام الدرجات المشاهدة (الدرجات الخام) (Observed Score Equating)، تقوم فكرة معادلة الاختبار بهذه الطريقة على التنبؤ بالتوزيع النظري للدرجات الخام للاختبار عن طريق بناء التوزيع التكراري للدرجات الخام للاختبار لكل مفحوص حسب قدرته، كما أنها لا تتطلب أن تكون نماذج الاختبارات المراد معادلتها قد بنيت بمواصفات متشابهة، لذلك فهي أكثر استخدامًا في هذه العمليات، (Kolen & Brennan, 2004).

كما أن هناك عدة تصاميم رئيسة لجمع بيانات عملية المعادلة باستخدام نظرية الاستجابة للمفردة، منها:

1- تصميم المجموعة الواحدة (Single Group Design)، ويتم فيه تطبيق نُسخ الاختبارات المراد إجراء معادلتها على نفس الأفراد.

2- تصميم المجموعات المتكافئة (Equivalent Group Design)، حيث يتم تطبيق نُسختي الاختبار المراد معادلتهما من خلال توزيع الأفراد بنُسخة عشوائية على مجموعتين، بحيث تكون المجموعتان متكافئتين في السمة المُقاسة بالاختبار، ويتم تطبيق النُّسخة الأولى على المجموعة الأولى، والنُّسخة الثانية على المجموعة الثانية.

3- تصميم المجموعات العشوائية المتوازية، حيث يتم تطبيق نُسختي الاختبار المراد معادلتهما من خلال توزيع الأفراد بنُسخة عشوائية على مجموعتين، ويتم تطبيق النُّسخة الأولى على المجموعة الأولى، والنُّسخة الثانية على المجموعة الثانية، ولكن يتم بعد ذلك عكس نُسختي الاختبار على المجموعتين. والفكرة في هذا التصميم ضمان أن العوامل مثل: التعب، والتعلم السابق، والخبرة، والممارسة، يكون لها نفس التأثير في نُسختي الاختبار (Kabasakal & Kelecioglu, 2015 و Livingston, 2004).

4- تصميم المجموعات غير المتكافئة ذات المفردات المشتركة (Common (Items Non-Equivalent Groups Design، حيث يتم تطبيق نُسختي الاختبار المراد معادلتهما على مجموعتين، ويتم تطبيق

النُّسخة الأولى على المجموعة الأولى، والنُّسخة الثانية على المجموعة الثانية، ومن ثم يتم تطبيق اختبار مشترك على المجموعتين بنفس الوقت والترتيب. وقد يكون الاختبار المشترك داخليًا (مجموعة المفردات موجودة في النُّسختين، وهي جزء لا يتجزأ من نُسختي الاختبار)، أو خارجيًا (اختبار مستقل عن نُسختي الاختبار الأصلي، ويعطى خارج وقت الاختبار الأصلي).

أشار تشين، وليفينجستون، وهولاند (Chen, Livingston & Holland, 2011)، الوارد في المحرزي (2015)، أن تصميم المجموعات غير المتكافئة ذات المفردات المشتركة، يتيح عددًا من التسهيلات العملية في تطبيق نماذج الاختبار على الأفراد، فهو يسمح بتقديم نُسخ مختلفة من الاختبار في كل يوم من أيام التطبيق، ولا يتطلب من الطالب أن يقدّم أكثر من نُسخة واحدة، سواء في التوقيت نفسه أم في توقيت مختلف، ويساعد هذا التصميم على التقليل من احتمالية تسرب مفردات الاختبار، وتقليل الآثار المترتبة على حدوث نوع من التسرب في حالة حصوله. كما يشترط بناء مفردات مشتركة (Common Items)، أو اختبار رابط (Anchor Test) بين نُسخ الاختبار، والتي يشترط فيها أن تمثل محتوى الاختبار الكلي، وكذلك أن يمثل نسبة 20٪ إلى 40٪ من عدد المفردات الكلي في الاختبار، وأن تتكافأ خصائصها السيكومترية مع الخصائص السيكومترية للاختبار الكلي.

لقد أصبح مجال معادلة الاختبارات باستخدام النظرية الحديثة في القياس محط اهتمام العديد من الباحثين، للمزايا التي توفرها هذه النظرية، ولقدرتها على معالجة جوانب القصور التي كانت في النظرية الكلاسيكية، وتمكنها من تقدير المعالم والقدرات بدقة كبيرة؛ لذلك أجريت العديد من الدراسات حول معادلة الاختبارات باستخدام نظرية الاستجابة للمفردة، منها: دراسة أيوب (1994)، التي هدفت إلى إجراء معادلة الاختبارات بأربع طرق منها استخدام طريقتي نموذج "راش"، والنموذج ثنائي المعالم في نظرية الاستجابة للمفردة، وباستخدام تصميم المجموعات غير المتكافئة ذات المفردات المشتركة، حيث قامت الدراسة بتصميم ثلاثة اختبارات للصفوف: الرابع والخامس والسادس، لكل اختبار أربعون فقرة، تم تطبيقها على عينتين، تم اختيارهما بطريقة عشوائية بلغت العينة الأولى (1390)، والعينة الثانية (1412)، وأشارت النتائج إلى أن المعادلة باستخدام نماذج الاستجابة للمفردة كانت أكثر فاعلية من بقية الطرق الأخرى المستخدمة في إجراء عملية المعادلة.

وتناولت دراسة طيفور (2007)، مقارنة نماذج نظرية الاستجابة للمفردة: (أحادي المعلم، أو ثنائي المعلم، أو ثلاثي المعلم) في معادلة الاختبارات، وهدفت إلى معرفة أنسب هذه النماذج في معادلة الاختبارات عند تطبيق ثلاثة تصميمات للمعادلة: (تصميم المفردات المشتركة، وتصميم المجموعة الواحدة، وتصميم المجموعات المتكافئة)، باستخدام اختبارين في مادة الجبر والإحصاء، يتكون كل منهما من 33 فقرة بينهما 8 مفردات مشتركة، تم تطبيق الدراسة على عينة مكونة من (1346) طالبًا

وطالبة من طلاب الصف الأول الإعدادي. وأظهرت نتائج الدراسة تشابه درجات معادلة الاختبارين بين النماذج الثلاثة باستخدام تصميم المجموعة الواحدة، بينما أظهر النموذج أحادي المعلم دقة أعلى عند معادلة الاختبارين باستخدام تصميم المفردات المشتركة، وتصميم المجموعات المتكافئة.

كما هدفت دراسة الوليلي (2005) إلى المقارنة بين دقة معادلة الاختبارات، في ضوء نظريتي القياس الحديثة والكلاسيكية، وذلك من خلال المقارنة بين نتائج اختبارين: أحدهما سهل، والآخر صعب، تم اشتقاق مفرداتهما من اختبار مادة الرياضيات، وتم تطبيقه على عينة من تلاميذ الصف الرابع الابتدائي، بلغت (418) تلميذًا. وقد تم استخدام طريقة المعادلة المئينية في نظرية القياس الكلاسيكية، واستخدام نموذج "راش" في نظرية الاستجابة للمفردة، عند تطبيق تصميم المفردات المشتركة، وتصميم المجموعة الواحدة. بعد تحليل البيانات باستخدام برنامج (RUMM)، أظهرت نتائج الدراسة أن تصميم المفردات المشتركة يعطي درجات معادلة الاختبارات أكثر دقة من تصميم المجموعة الواحدة، سواء في حالة استخدام نظرية الاستجابة للمفردة أم طريقة المعادلة المئينية.

وهدفت دراسة المدانات (2012) إلى مقارنة فاعلية طريقتي: معادلة الدرجات الحقيقية، والدرجات المشاهدة في معادلة نُسختين لاختبار الفيزياء: (تتكون كل من النُّسختين من 20 فقرة بالإضافة إلى 10 مفردات استخدمت كمفردات مشتركة). تم تطبيق الاختبار على عينتين بلغت كل منها (199) طالبًا. وأظهرت النتائج عدم وجود فروق ذات دلالة في نتائج معادلة نُسختي اختبار الفيزياء، بين طريقتي معادلة الدرجات الحقيقية والدرجات المشاهدة.

وهدفت دراسة الشمري، والشريفين (2015)، إلى معادلة درجات نُسخ مختلفة من اختبار القدرات المعرفية العامة، (مكوّن من 96 فقرة ثنائية التدريج)، موزعة على ثلاثة نماذج، والمطبق على طلبة الثانوية العامة في المملكة العربية السعودية؛ وذلك باستخدام طريقتي المعادلة الخطية، والمعادلة المئينية التابعة للنظرية الكلاسيكية، واستخدام نموذج "راش" التابع لنظرية الاستجابة للمفردة، بتصميم المجموعات العشوائية المتكافئة، على عينة بلغت (4500) طالبًا وطالبة، بحيث كان عددهم (1500) طالبًا وطالبة لكل نموذج اختباري. استخدمت الدراسة برمجيات (SPSS - BILOG-MG)، لتحليل البيانات، واستخدمت معيارين للحكم على فاعلية الطرق المستخدمة في المعادلة: الصدق التقاطعي، والخطأ المعياري للمعادلة. وأشارت نتائج الدراسة أن نموذج "راش"، كان أكثر فاعلية في عملية معادلة الاختبار من طريقتي المعادلة الخطية والمئينية، وفقًا لمعيار الصدق التقاطعي، والخطأ المعياري.

ومن الدراسات الأجنبية التي تناولت معادلة الاختبارات باستخدام نظرية الاستجابة للمفردة، دراسة ليستز، ويانغ (Li, Lissitz & Yang, 1999)، والتي درست فعالية طرق المعادلة في الاختبارات المكونة من مفردات ثنائية الاستجابة، ومتعددة الاستجابة. استخدمت الدراسة نموذج الاستجابة

المتدرجة لاختبار ذي مفردات مشتركة، في مهارتي: القراءة والكتابة لطلبة الصف الرابع، بهدف التعرف إلى أفضل طريقة للمعادلة عند استخدام اختبار ذي مفردات مختلطة، وأيضًا تأثير اختلاف النسبة بين عدد المفردات المتعددة إلى الثنائية على عملية المعادلة. بلغ حجم العينات المستخدمة في الدراسة (1000، 2000، 3000). أما النسبة بين عدد المفردات المتعددة إلى الثنائية، فقد بلغت (10/ 5، 15/ 5، 20/ 5) على التوالي، حيث إن المفردات المتعددة ذات خمس استجابات (5 four-category items)، وكانت نتيجة الدراسة أن استخدام طريقة المعايرة المتزامنة (Concurrent Calibration)، يعطي نتائج أكثر دقة وأقل تحيزًا، ويعطي تقديرات جيدة في عملية المعادلة عند استخدام اختبار ذي مفردات مختلطة (ثنائية ومتعددة). وظهرت نتيجة غير متوقعة، وهي أن زيادة عدد المفردات الثنائية (Dichotomous) في الاختبار أدى إلى ارتفاع متوسط خطأ القياس.

كما هدفت دراسة جوبس، وكليسيجلو (Gubes & Kelecioglu, 2016)، إلى معرفة تأثير أبعاد الاختبار، أو عامل البعدية (Dimensionality)، وعدد المفردات المشتركة، وطريقة الربط المستخدمة بين النماذج، على معادلة الاختبارات ذات المفردات المشتركة. طبقت الدراسة على بيانات اختبار (TIMSS 2011) للصف الثامن، والمكون من 194 مفردة، تم اختيار 50 فقرة اختبارية، منها 40 فقرة ثنائية الاستجابة، و10 مفردات متعددة الاستجابة، حيث تم تطبيقه على مجموعتين تكونت كل مجموعة من 3000 تلميذ. تم استخدام تصميم المجموعات غير المتكافئة ذات المفردات المشتركة، وتم استخدام برمجية (PARASCALE) للمعايرة، و برمجية (STUIRT) لعملية الربط بين النماذج، وبرمجية (POLYEQUATE) لإجراء عملية المعادلة. ثم تم استخدام تحليل التباين الثلاثي، لمعرفة أي العوامل له تأثير دال إحصائيًا على عملية المعادلة. أظهرت النتائج وجود دلالة إحصائية للعوامل الثلاثة: (أبعاد الاختبار، وعدد المفردات المشتركة، وطريقة الربط المستخدمة بين النماذج) في التأثير على نتائج عملية المعادلة، ولكن أكثر العوامل تأثيرًا، هو عامل أبعاد الاختبار (Dimensionality)؛ حيث إن لتعدد أبعاد الاختبار تأثيرًا عكسيًا على عملية المعادلة. وتشير هذه النتيجة على ضرورة التحقق من افتراض أحادية البعد قبل معادلة الاختبارات بالنظرية الحديثة.

مشكلة الدراسة

حصل التعليم في الوقت الحالي على اهتمام كبير من قبل مختلف دول العالم، ومن دول الخليج بصفة خاصة؛ ولعل من أحد الأسباب الرئيسة لهذا الاهتمام، يتمثل فيما يعرف باقتصاد المعرفة والمعلومات، الذي أدخل على المجتمعات جميعها في القرن الحادي والعشرين. ويحتاج الانتقال إلى تلك الاقتصادات نوعًا مختلفًا من التعليم، يهدف إلى تنمية المهارات اللازمة، للخوض في سوق عمل شديد التنافسية، مع القدرة على تحليل هذه المعارف وتوليفها وربطها بمعلومات أخرى متباينة ومختلفة. وما

يـزال التعليـم في دول مجلـس التعـاون الخليجـي، يحتـاج إلى تركيـز أكـبر عـلى مخرجـات عالميـة، تتضمّـن تبنّـي التعلّـم المبنـي عـلى مهـارات القـرن الحـادي والعشريـن، التـي تتطلّـب القـدرة عـلى التحليـل والنقـد والعمـل التعـاوني، والتفكـير خـارج الأطـر الاعتياديـة، لخلـق المعرفـة في عـالم صـار اقتصـاده أساسـا يعـرف باقتصـاد المعرفـة، (Taha, 2008).

فقـد كشـفت البيانـات أنّ نتائـج طـلاب جميـع الـدول الخليجيـة المشـاركة – بـما فيهـا سـلطنة عـمان وقطر – أتـت تحـت المعـدّل العالمـي عـلى رغـم مدخولاتهـا القوميـة المرتفعـة، ونسـبة صرفهـا عـلى قطـاع التعليـم. كـما أنهـا مـا زالـت تعـاني مـن ضعـف جـودة ومخرجـات هـذا التعليـم، والتـي تعكسـها نتائـج الطـلاب في الامتحانـات العالميـة المقنّنـة في العلـوم والرياضيـات والقرائيـة باللغـة العربيـة، ممـا يضعـف تنافسـية هـؤلاء الطـلاب مسـتقبلًا في سـوق العمـل. كـما أن مجتمعـات دول مجلـس التعـاون الخليجـي ليسـت مختلفـة عـن باقـي المجتمعـات الحديثـة في مجـال الاسـتثمار في التعليـم، وفي البنيـة البشريـة، فالتعليـم هـو المصـدّر الأهـم لتوفـير، وتسـليح القـوى العاملـة اللازمـة لسـوق العمـل، (United Nations, 2009).

وتلعـب الاختبـارات الدوليـة – بـما فيهـا اختبـار بيرلـز – دورًا كبـيرًا في الحكـم عـلى جـودة النظـام التعليمـي التـي تتبناهـا دول العـالم، حيـث إنهـا تقـدم مقارنـة دوليـة بـين مسـتوى مخرجـات الأنظمـة التعليمية، كانعـكاس لجـودة مُدخـات وعمليـات النظـام التعليمـي. وتسـتخدم هـذه الاختبـارات نُسـخًا مختلفـة لهـا، وذلـك حرصًـا عـلى تحقيـق السريـة، وتجنـب معوقـات التطبيـق في أوقـات مختلفـة وأماكـن مختلفـة؛ ونظـرًا لاختـلاف نُسـخ الاختبـار الواحـد، وللتمكـن مـن إصـدار أحـكام متشـابهة عـبر نُسـخ الاختبـار بشـكل تبـادلي، فإنـه يجـب إجـراء معادلـة لدرجـات النُّسـخ المختلفـة للاختبـار، حيـث تعتـبر معادلـة الاختبـارات مـن الإجـراءات الضروريـة، وتتطلـب اسـتخدام تصميـم مناسـب لجمـع البيانـات وتحليلهـا، واسـتخدام الطريقـة المناسـبة لتحويـل درجـات نُسـخ الاختبـار، بحيـث يمكـن المقارنـة بينهـا.

تتكـون بعـض الاختبـارات مـن مفـردات مختلطـة مـن المفـردات ثنائيـة الاسـتجابة، والمفـردات متعـددة الاسـتجابة؛ وذلـك لأسـباب متعـددة منهـا: تحقيـق الموضوعيـة، والتقليـل مـن ذاتيـة المقيـم، وكذلـك التخلـص مـن عـادة الحفـظ. ويُدخـل اسـتخدام الاختبـارات ذات المفـردات المختلطـة العديـد مـن المعوقـات في عمليـة معادلـة النُّسـخ المختلفـة لهـا، فقـد لا تقيـس هـذه المفـردات السـمة نفسـها، ممـا يسـبب تعـدد أبعـاد الاختبـار، لذلـك تسـعى العديـد مـن الاختبـارات إلى التحقـق مـن عامـل أحاديـة البعـد بغـض النظـر عـن البنـاء الهيـكلي للاختبـار، (Gubes & Kelecioglu, 2016). وتحتـاج الاختبـارات ذات المفـردات المختلطـة اسـتخدام نـماذج رياضيـة متقدمـة في نظريـة الاسـتجابة للمفـردة، ممـا يصعـب عمليـات تطبيـق المعادلـة.

ويعـد اختبـار "بيرلـز" مـن الاختبـارات الدوليـة التـي تسـتخدم المفـردات المختلطـة، والتـي تحتـوي عـلى مفـردات ثنائيـة الاسـتجابة (مفـردات اختيـار مـن متعـدد)، ومفـردات متعـددة الاسـتجابة (مفـردات

ذات إجابة إنشائية من الطالب). حيث يستخدم الاختبار هذين النوعين لقياس مهارة القراءة، نظرًا لإمكانية تصحيح مفردات الاختيار من متعدد، ثنائية الاستجابة آليًا، مما يحقق نسبة عالية من الدقة والموضوعية، مع قدرتها على قياس مهارات التفكير العليا، وأيضًا تساعد المفردات المتعددة الاستجابة على صياغة أسئلة تقيس مهارات تنظيم الأفكار وتكاملها وعرضها باتساق.

كما يستخدم اختبار "بيرلز" نُسَخًا مختلفة من الاختبار في التطبيق على كل دولة تجنبا للوقوع في معوقات اجراء الاختبار، كما ويسمح وجود نُسخ متعددة من الاختبار بتقديمه في أوقات مختلفة وأماكن مختلفة (Kolen, & Brennan, 2004; Andersson, Branberg & Wiberg, 2013). وعلى الرغم من الجهود التي يبذلها معدو الاختبارات في وضع نُسخ متكافئة من الاختبار في المحتوى، والأهداف السلوكية التي تقيسها، ونوع الأسئلة، تُظهر بيانات تطبيق الاختبار وجود اختلاف في مستوى صعوبتها، ويحرص واضعو الاختبارات عند استخدامهم درجات النُّسخ المتعددة من الاختبار ان تكون متعادلة، لتحقيق الغرض الأساسي من معادلة الاختبار، والمتمثل في استخدام درجات متعادلة فعالة للنُّسخ المتعددة من الاختبار، بحيث يمكن استخدام هذه الدرجات، لتمثل هذه النُّسح، وكأنها اختبار واحد.

وللقيمة التي يمثلها اختبار "بيرلز"، وأهمية القرارات المبنية على نتائجه، فإن من الأهمية بمكان القيام بإجراء معادلة بيانات طلبة سلطنة عمان لتكون أكثر دقة وملاءمة، للحصول على معايير وطنية لتفسير درجات الطلاب في هذا الاختبار. كما قد تختلف الدرجات المتحصل عليها من المعادلة باختلاف نوع الدرجات المستخدمة، سواء الدرجات الظاهرية أم الدرجات الحقيقية، وهذا يساعد في تحقيق أهداف الدراسة في الحصول على درجات للطلاب على نُسخ الاختبار، تكون انعكاسًا لقدراتهم من دون تدخل صياغة مفردات الاختبار، أو التخمين، أو التسرع في الاجابة. كما تساعد معادلة كُتيبات اختبار "بيرلز" – يطلق على نُسخ الاختبار في اختبار بيرلز بمسمى كتيبات – في تعديل الفروق المحتملة بين كتيبات الاختبار، بحيث يمكننا المقارنة بين تلك النماذج التي تقيس السمة ذاتها، وفقًا لمعيار مشترك تنسب إليه درجات الطلاب في النماذج المختلفة، مما يؤدي إلى تقديرات ثابتة عن مستوى تحصيل كل طالب. من هنا ظهرت الحاجة إلى إجراء معادلة نماذج اختبار "بيرلز" 2011 بسلطنة عمان، باستخدام طرق المعادلة بالدرجات الحقيقية، والمعادلة بالدرجات المشاهدة التابعة للنظرية الحديثة للقياس، بتصميم المجموعة الواحدة، وتصميم المجموعات غير المتكافئة ذات المفردات المشتركة، حيث يمكن تحديد مشكلة الدراسة من خلال الأسئلة التالية:

- ما الإحصاءات الوصفية لكتيبات اختبار (PIRLS, 2011)، قبل وبعد اجراء المعادلة؟
- ما قيم الدرجات المتعادلة بين كتيبات اختبار (PIRLS 2011)، باستخدام طريقتي المعادلة بالدرجات الحقيقية، والمعادلة بالدرجات المشاهدة؟
- هل تختلف قيم الدرجات المتعادلة لكتيبات اختبار (PIRLS 2011)، بين طريقتي المعادلة بالدرجات الحقيقية، والمعادلة بالدرجات المشاهدة؟

أهمية الدراسة

تعد هذه الدراسة – على حد علم الباحثين – أول دراسة استخدمت نظرية الاستجابة للمفردة في معادلة 13 كُتيب لاختبار "بيرلز"، باستخدام نموذج الاستجابة المتعددة بسلطنة عمان، على عكس الدراسات السابقة، التي اقتصرت على استخدام النماذج ثنائية الاستجابة في معادلة الاختبارات. كما تقدم هذه الدراسة معيارًا للحكم على درجات الطلاب في هذا الاختبار، الذي يمثل أهمية كبيرة، وتبنى على نتائجه قرارات تتعلق بتطوير المناهج القرائية، بحيث يمكن المقارنة، ليس فقط بين المجموعة التي أخذت نفس الاختبار، ولكن المقارنة بين جميع طلاب السلطنة. كما أن استخدام هذه الدراسة لتصميمي المجموعة الواحدة، وتصميم المجموعات غير المتكافئة ذات المفردات المشتركة في معادلة نماذج اختبار (PIRLS 2011)، يفتح الباب أمام الدراسات المستقبلية للتوسع في هذا المجال، والاستفادة من إمكانات النظرية الحديثة.

مصطلحات الدراسة

- معادلة الاختبارات (Tests Equating): "عملية إحصائية يتم فيها وضع درجات الاختبارات المختلفة، والتي تقيس السمة نفسها على مقياس مشترك، بحيث يحصل المفحوص على نفس التقدير للسمة التي يقيسها الاختبار، بغض النظر عن اختلاف النسخ الاختبارية"، (بركات، 2010، ص 10).
- نظرية الاستجابة للمفردة (Item Response Theory): هي "مجموعة من النماذج الرياضية التي تفترض أنه يمكن التنبؤ بسلوك الأفراد، أو يمكن تفسير أدائهم في اختبار نفسي أو تربوي معين، في ضوء خصائص مميزة لهذا الأداء تسمى سمات "Traits"، أو بمعنى آخر وجود واحدة أو أكثر من السمات الأساسية التي تحدد استجابات الفرد على مفردات الاختبار"، (هيبة، وعمر، 2011، ص 9).
- نموذج الاستجابة المتدرجة (Graded Response Model): اقترحت هذا النموذج "Samejima"، ويعد تعميمًا للنموذج ثنائي البارامتر (2PL)، بحيث يمثل العلاقة غير الخطية بين مستوى قدرة الفرد، واحتمال استجابته في استجابة معينة، وذلك يتطلب عملية ذات خطوتين من أجل تحديد الاحتمال المشروط لاستجابة فرد في قسم معين، (علام، 2005).
- الدراسة الدولية لقياس مدى تقدم القراءة في العالم "بيرلز" (Progress In International Reading Literacy Study): اختبار عالمي يقوم على أساس المقارنة، لقياس قدرات طلبة الصف الرابع في مهارات القراءة بلغتهم الأم، وذلك لتحديد جوانب القوة والضعف لديهم، ومن ثم تطوير تلك المهارات والارتقاء بها، بما يحقق أهداف المجلس الأعلى للتعليم، ويلبي متطلبات تطوير التعليم في الدولة، ويسهم في تطوير قدرات وكفايات الطلبة، (وزارة التربية والتعليم، 2014).

منهجية الدراسة

(أ) مجتمع وعينة الدراسة:

تكوّن مجتمع الدراسة من جميع طلبة الصف الرابع بسلطنة عمان، في العام الدراسي 2011/ 2012، وتكوّنت عينة الدراسة من البيانات الأرشيفية لعينة الطلبة الذين خضعوا لدراسة (PIRLS 2011) من سلطنة عمان، والتي تم اختيارها بناء على مواصفات اختيار عينة الدراسة الدولية في جميع الدول المشاركة، وفقًا لمعايير التصنيف العالمية للتربية التابعة للتصنيف الدولي الموحد للتعليم (ISCED 2011)، وهو التصنيف المرجعي لتنظيم برامج التعليم، والمؤهلات ذات الصلة حسب مستويات ومجالات التعليم. وتكونت العينة من طلبة الصف الرابع الأساسي (الذكور والاناث) في جميع محافظات السلطنة من مواليد (2001–2000)، البالغ عددهم 10394، والذين تم إجراء اختبار "بيرلز" 2011 عليهم (IEA, 2011، والمديرية العامة للتقويم التربوي، 2011). ويوضح الجدول (1) توزيع عينة الدراسة على كتيبات اختبار الدراسة PIRLS 2011.

جدول (1)

توزيع عينة الدراسة من طلاب سلطنة عمان على كتيبات اختبار PIRLS 2011

رقم الكتيب	1	2	3	4	5	6	7	8	9	10	11	12	13	الإجمالي
عدد الطلبة	700	684	675	683	692	694	691	693	691	688	694	690	2118	10394

(ب) أداة الدراسة:

استخدمت الدراسة اختبار "بيرلز"، وهو اختبار دولي يتكون من 5 نصوص أدبية، تتضمن حكاية، أو قصة واقعية من التراث، أو من الحياة اليومية و5 نصوص علمية تتضمن معلومات، أو مواد تثقيفية في مختلف فروع المعرفة، تقترحها الدول المشاركة موزعة على 13 كُتيب مختلف، بمعدل نصين بكل كُتيب، يطلب من التلميذ قراءة النصين والإجابة عن الأسئلة، وتوزع الكُتيبات على التلاميذ بطريقة مسبقة التنظيم من قبل اللجنة المنظمة للدراسة، (Marian & Jay, 2001) (Mullis, Martin, Kennedy, Tong & Sainsbury, 2009).

تستعمل في اختبار (PIRLS 2011) نوعان من الأسئلة للتقييم، أسئلة الاختيار من متعدد، وأسئلة تحتاج إجابات إنشائية، توفر الأسئلة ذات الاختيار من متعدد أربعة خيارات للإجابة، منها إجابة واحدة صحيحة، أما الأسئلة ذات الإجابة الإنشائية تتطلب إنشاء إجابات مكتوبة، ويستعمل هذا النوع لتقييم جوانب الفهم الذي يستدعي الطلاب تقديم دعم لإجاباتهم، أو تستدعي بتقديم شرح

بالاعتماد على معلوماتهم وتجاربهم (Mullis et al., 2009)، ويختلف عدد الدرجات المخصصة لكل سؤال من أسئلة الإجابة القصيرة من كُتيب إلى آخر، كما يختلف توزيع الأسئلة وعددها أيضًا في الكُتيبات، لذلك تختلف الدرجة الكلية من كتيب إلى آخر، كما هو واضح في الجدول (2)؛ إذ أن النصوص المختلفة قد تتطلب أنواعًا مختلفة من الأسئلة (Mullis & et. al, 2009، والتقويم التربوي، 2011).

جدول (2)

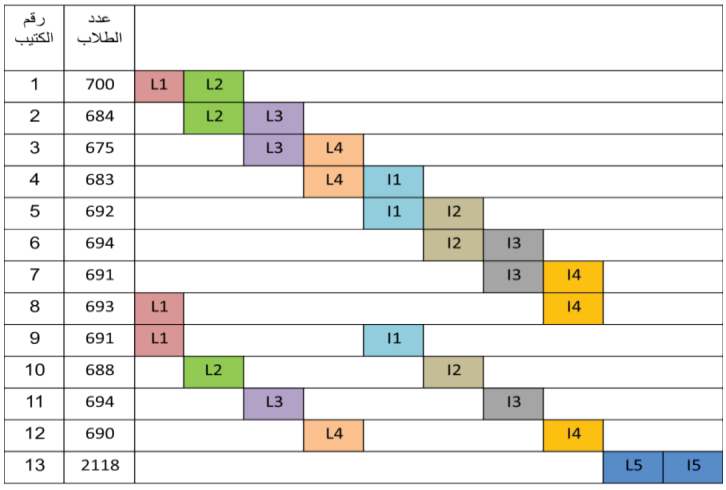
أدنى وأعلى درجة لمفردات كتيبات PIRLS 2011

رقم الكتيب	1	2	3	4	5	6	7	8	9	10	11	12	13
عدد المفردات	25	26	34	32	24	24	28	29	25	24	26	36	35
أدنى درجة	0	0	0	0	0	0	0	0	0	0	0	0	0
أعلى درجة	32	35	42	40	32	32	37	36	33	31	36	43	42

يتم توزيع النصين في كتيبات الاختبار باستخدام تصميم العينة المصفوفية، الذي يقسم النصوص والمفردات إلى نماذج أو كتل، بحيث يتم تكوين كُتيبات الطلاب، وفقًا لترتيب منهجي، بحيث يحتوي كل كُتيب على نصين. ويوضح الشكل (1) نظام الربط بين كُتيبات اختبار "بيرلز"، وتوزيع النصوص العلمية (L1 إلى L4)، والأدبية (I1 إلى I4) على 12 كتيب، بحيث تتكرر كل منها في 3 كُتيبات، أما النص الأدبي (L5) والنص العلمي (I5)، فتظهر في الكتيب (13)، فليست مرتبطة مباشرة مع أي كُتيب آخر.

شكل (1)

رسم توضيحي لتوزيع كتيبات اختبار PIRLS 2011 على عينة الدراسة من طلاب سلطنة عمان



تم التأكد من الخصائص السيكومترية لكتيبات اختبار "بيرلز" باستخراج معامل ثبات كل كُتيب باستخدام برنامج"SPSS" بطريقة" ألفا كرونباخ"، حيث تراوحت قيم معامل ثبات "ألفا كرونباخ"

لجميع الكُتيبات بين (0.83- 0.89)، وهي معاملات مرتفعة، ومناسبة لأغراض الدراسة مما يشير إلى دقة الأداة وثباتها. كما تم التحقق من صدق الاختبار عن طريق الصدق البنائي، وذلك من خلال التأكد من العلاقة الارتباطية بين النصين بكل كُتيب، ولتحقيق هذا الغرض تم حساب معامل ارتباط "بيرسون"، لمعرفة مدى الارتباط بين النصين في كل كُتيب، ويوضح الجدول (3) ذلك.

جدول (3)

معاملات الارتباط بين نماذج كل كتيب من كتيبات اختبار PIRLS 2011

رقم الكتيب	1	2	3	4	5	6	7	8	9	10	11	12	13
قيمة معامل الارتباط	0.73	0.71	0.73	0.68	0.64	0.67	0.64	0.68	0.63	0.69	0.69	0.67	0.69

ويتضح من خلال الجدول (3)، أن معاملات الارتباط بين النصين بكل كُتيب دالة إحصائيًا عند مستوى (0.01)، حيث تتراوح بين (0.63، 0.73)، وبلغ أعلى معامل ارتباط (0.73)، وهو بين النصين للكُتيبات (1) و(3)، بينما أقل معامل ارتباط (0.63)، وهو بين النصين للكُتيب (9)، مما يوفر درجة مقبولة من الصدق البنائي لكل كتيب.

وتم استخدام الكُتيب (1) ككُتيب مرجعي في عملية المعادلة، لارتفاع نسبة ثبات الكتيب، وتمثل المفردات المشتركة ما يزيد عن 20٪ من عدد المفردات الكلي فيه، وهي من ضمن شروط اختبار الفقرات المشتركة، كما أشار المدانات (2012).

(جـ) إجراءات الدراسة:

تم تطبيق كتيبات اختبار (PIRLS 2011) في يوم واحد، بحيث أدى الطلاب الاختبار في نفس اليوم مع فترة استراحة قصيرة بين النصين العلمي والأدبي، بحيث يكون الوقت المخصص لفترتي اختبار "بيرلز" ما يقارب 10 دقائق تقريبًا للتحضير للامتحان، بما في ذلك إعداد الطلاب، وقراءة التعليمات عليهم، وتوزيع كُتيبات الامتحان، ثم 40 دقيقة للإجابة عن أسئلة النص الأول من كتيب الأسئلة، وبعد ذلك استراحة قصيرة (15 دقيقة)، ثم 40 دقيقة للإجابة عن أسئلة النص الثاني من كتيب الأسئلة.

تم جمع البيانات وترتيبها، ثم تطبيق التصميم المناسب لمعادلة الاختبار، مع الأخذ بعين الاعتبار خصائص مجموعة المفحوصين، وطبيعة الاختبارات المراد معادلتها، باستخدام تصميم المجموعة الواحدة، وتصميم المجموعات غير المتكافئة ذات المفردات المشتركة. تم بعد ذلك بناء تدريج مشترك يربط بين السمة المراد قياسها، ومعلمات المفردات، باستخدام تصميم المجموعة الواحدة،

وتصميم المجموعات غير المتكافئة ذات المفردات المشتركة في تنظيم البيانات من خلال المعايرة المتزامنة (Concurrent Calibration)، ثم تنفيذ الإجراء الفعلي لعملية معادلة الاختبارات، وتجهيز بيانات الاختبار كملفات يحتاجها استخدام برنامج (Poly Equate (V0.5))، للحصول على الدرجات المعادلة بطريقتي الدرجات المشاهدة، وطريقة الدرجات الحقيقية.

(د) المعالجة الإحصائية:

للإجابة عن السؤال الأول في الدراسةُ تم استخدام البرنامج الإحصائي (Poly Equate(V0.5))، لإيجاد قيم الدرجات المعادلة الناتجة من طريقتي المعادلة، باستخدام نموذج الاستجابة المتدرجة (Graded Response Model) بافتراض أن قيمة ثابت المعادلة (D=1)، وهي القيمة النظرية التي تقلل من الاختلاف بين توزيع الدرجات الحقيقية بين الاختبار الجديد والاختبار المرجعي، ووضعها على تدريج مشترك (Baker, 1992)، حيث تم إيجاد قيم الدرجات المعادلة باستخدام طريقتي معادلة الدرجات الحقيقية (True Score Equating)، ومعادلة الدرجات المشاهدة (Observed Score Equating) على حد سواء.

كما تم حساب الدرجة المعيارية للقيم المتعادلة بطريقتي المعادلة، ليسهل مقارنة درجة الطالب بين جميع الكُتيبات، وكذلك مقارنة درجته مع أفراد مجموعته في جميع الكُتيبات، وتحديد موقعه النسبي مقارنة بالمتوسط الحسابي لمجموعته، على اعتبار أن المتوسط الحسابي في الدراسة الدولية (PIRLS 2011) يساوي (500)، وبانحراف معياري يساوي (100)، عن طريق المعادلة التالية:

$$\text{الدرجة المعيارية} = \left(\frac{\text{الدرجة} - \text{المتوسط}}{\text{الانحراف المعياري}} \times 100\right) + 500$$

كما تمت الإجابة عن السؤال الثاني عن طريق إيجاد الإحصاءات الوصفية للكُتيبات قبل المعادلة باستخدام برنامج SPSS. ونظرًا لاختلاف تدريج الكُتيبات، تم تحويل تدريج جميع الكُتيبات إلى تدريج الكُتيب المرجعي (الكُتيب 1)، ليسهل المقارنة بين المتوسطات والانحرافات المعيارية من خلال المعادلة الآتية:

$$\text{الدرجة المعدلة} = \frac{\text{درجة الطالب في الكتيب المراد معادلته}}{\text{الدرجة الكلية للكتيب المراد معادلته}} \times \text{الدرجة الكلية للكتيب المرجعي}$$

أيضا تم استخدام البرنامج الاحصائي SPSS لاستخراج الإحصاءات الوصفية للدرجات الخام بعد معادلتها بطريقة الدرجات الحقيقية، وطريقة الدرجات المشاهدة.

للإجابة عن السؤال الثالث تم استخدام اختبار "ت" للعينات المرتبطة (t-test paired samples)، لمعرفة فيما إذا كان هناك اختلاف في نتائج القيم المعادلة، باختلاف طريقة المعادلة في كل كُتيب من الكُتيبات المكونة لاختبار مهارات القراءة (PIRLS 2011).

نتائج الدراسة

أولًا- إجابة السؤال الأول: ما الإحصاءات الوصفية لكتيبات اختبار (PIRLS 2011) قبل وبعد اجراء المعادلة؟

(أ) الاحصاءات الوصفية قبل المعادلة:

أظهرت النتائج في الجدول (4) أن المتوسطات الحسابية للدرجات الخام قبل المعادلة، تختلف عبر جميع الكُتيبات المراد معادلتها، حيث تتراوح بين (7.8 - 10.81)، حيث كانت أعلى قيمة للمتوسط للكُتيب (10)، والتي بلغت (10.81)، بينما أدنى قيمة للمتوسط كانت للكتيب (8)، والتي بلغت (7.8). أما قيم الانحراف المعياري للدرجات الخام للكُتيبات قبل المعادلة، فتراوحت بين (4.67-7.88)، حيث كانت أعلى قيمة للانحراف المعياري للكتيب (12) بقيمة (7.88)، وأقلها كانت في الكتيب (3).

وعند مقارنة المتوسطات الحسابية، والانحرافات المعيارية للدرجات الخام لكُتيبات اختبار "بيرلز" قبل إجراء المعادلة بالنسبة للكُتيب المرجعي (الكتيب 1)، نجد أنها تتفاوت في درجة اختلافها عن الإحصاءات الوصفية للكتيب المرجعي، وكان ثلاثة من الكُتيبات (الكتيب 6، 7، 10) أقل صعوبة من الكُتيب المرجعي، فكانت متوسطاتها الحسابية أكبر من الكتيب المرجعي، بينما كانت بقية الكُتيبات أكثر صعوبة، حيث كانت متوسطاتها الحسابية أقل من الكتيب المرجعي. وهذا من الأسباب التي دعت إلى ضرورة معادلة هذه الكتيبات، والمتمثل في وجود فرق بين الإحصاءات الوصفية لها قبل المعادلة، حيث إن تطبيق هذه الكُتيبات على عينات مختلفة من الأفراد، يؤدي في غالب الأحيان إلى الاختلاف في المتوسطات الحسابية؛ بسبب اختلاف صعوبتها، أو بسبب اختلاف قدرات الأفراد (المحرزي، 2014).

جدول (4)

الاحصاءات الوصفية للدرجات الخام قبل المعادلة في كتيبات اختبار PIRLS 2011

النموذج	المتوسط	الانحراف المعياري	الالتواء	التفلطح
الكتيب المرجعي	9.62	6.42	0.795	0.077-
الكتيب (2)	9.51	5.74	1.005	0.376
الكتيب (3)	9.19	4.76	1.076	0.352
الكتيب (4)	9.50	4.77	0.884	0.183
الكتيب (5)	9.02	5.73	0.887	0.196
الكتيب (6)	10.11	6.26	0.617	0.458-
الكتيب (7)	9.70	6.39	1.194	1.421
الكتيب (8)	7.80	6.67	1.236	1.072
الكتيب (9)	8.97	5.87	0.734	0.100-
الكتيب (10)	10.81	6.94	0.646	0.516-
الكتيب (11)	9.12	6.81	0.890	0.280
الكتيب (12)	8.20	7.88	1.280	1.333
الكتيب (13)	7.34	7.16	1.337	1.318

(ب) الإحصاءات الوصفية بعد المعادلة:

يوضح الجدول (5) والجدول (6) الإحصاءات الوصفية للدرجات الخام بعد معادلتها بطريقة الدرجات الحقيقية، وطريقة الدرجات المشاهدة على التوالي.

فنجد من خلال الجدول (5) أن قيم المتوسطات للدرجات الخام بعد معادلتها بطريقة الدرجات الحقيقية في جميع الكتيبات، تقترب من القيمة (9.62)، والتي تمثل المتوسط الحسابي للدرجات الخام للكتيب المرجعي، وكذلك بالنسبة للانحراف المعياري الذي تتقارب قيمه من القيمة (6.42)، أما قيم الالتواء والتفلطح، فتختلف فيما بينها بفوارق بسيطة، وفي مجملها قريبة من قيم الالتواء والتفلطح لدرجات الكتيب المرجعي.

جدول (5)

الإحصاءات الوصفية للدرجات الخام بطريقة معادلة الدرجات الحقيقية في كتيبات اختبار PIRLS 2011

الكتيب	المتوسط	الانحراف المعياري	الالتواء	التفلطح
2	9.75	6.29	0.923	0.149
3	9.60	6.05	0.893	0.014
4	9.78	6.08	0.763	0.081-
5	9.15	6.02	0.911	0.154
6	9.87	6.39	0.660	0.400-
7	9.29	5.95	0.742	0.240
8	9.43	6.12	0.782	0.065
9	9.63	6.14	0.685	0.304-
10	10.29	6.58	0.713	0.390-
11	9.59	6.32	0.772	0.056-
12	9.37	5.86	0.859	0.358
13	9.89	7.69	0.937	0.172-

وعند النظر إلى الجدول 6 الخاص بالإحصاءات الوصفية للدرجات الخام للكتيب بعد معادلتها بطريقة الدرجات المشاهدة، فيلاحظ تكرار نمط النتائج نفسها التي ظهرت في الجدول (5) عند استخدام طريقة المعادلة بالدرجات الحقيقية.

وتحقق كل من طريقتي المعادلة تقارب الإحصاءات الوصفية للدرجات الخام لكُتيبات الاختبار بعد معادلتها مع الكتيب المرجعي، وهذا ما يؤكده (2004) Kolen & Brennan، أن توزيع الدرجات في الكتيب الجديد بعد تحويله تتقارب إلى حد كبير مع توزيع الدرجات في الكتيب المرجعي. وبالتالي تتضح أهمية معادلة الكُتيبات في أنها تساعد على تعديل تدريج الكتيبات المراد معادلتها، وفقًا لتدريج الكتيب المرجعي.

جدول (6)

الإحصاءات الوصفية للدرجات الخام بطريقة معادلة الدرجات المشاهدة في كتيبات PIRLS 2011

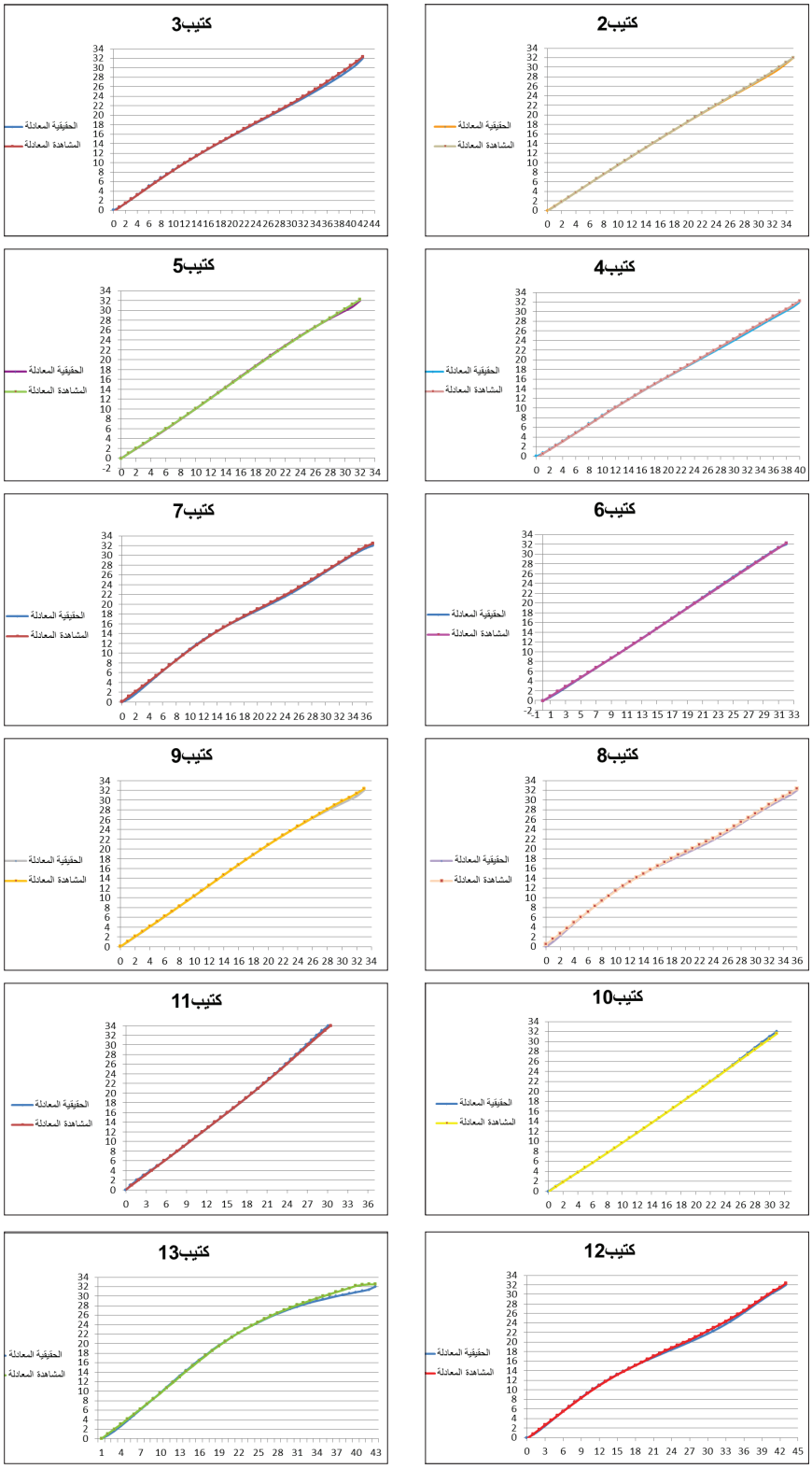
الكتيب	المتوسط	الانحراف المعياري	الالتواء	التفلطح
2	9.76	6.32	0.934	0.190
3	9.61	6.16	0.918	0.062
4	9.78	6.16	0.779	0.043-
5	9.14	5.10	0.913	0.187
6	9.88	6.32	0.665	0.395-
7	9.38	5.89	0.841	0.416
8	9.54	6.10	0.881	0.197
9	9.65	6.11	0.700	0.261-
10	10.27	6.55	0.697	0.420-
11	9.63	6.30	0.805	0.025
12	9.42	5.94	0.927	0.461
13	10.0	7.62	1.006	0.009

ثانيًا- إجابة السؤال الثاني: ما قيم الدرجات المتعادلة بين كتيبات اختبار (2011) PIRLS باستخدام طريقتي المعادلة بالدرجات الحقيقية والمعادلة بالدرجات المشاهدة؟

يعرض الشكل (2) العلاقة بين الدرجات الخام لجميع الكتيبات مع القيم المعادلة لها في تدريج الكتيب المرجعي (الكُتيب 1)، بطريقتي: معادلة الدرجات الحقيقية والدرجات المشاهدة. وتظهر جميع الأشكال تشابه طريقتي معادلة الدرجات الحقيقية، والدرجات المشاهدة إلى حد كبير في شكل العلاقة بين الدرجات الخام للكُتيبات مع القيم المعادلة لها في الكتيب المرجعي. كما يظهر الشكل أن هذا العلاقة تميل أن تكون غير خطية مع اختلاف درجة وشكل هذه العلاقة بين الكُتيبات. وتشير هذه النتيجة إلى اختلاف عملية المعادلة في كُتيبات اختبار "بيرلز"، ويعتمد شكل هذه المعادلة على مقدار الاختلاف في صعوبة كل كُتيب بالنسبة إلى الكتيب المرجعي، ويظهر التقارب الأكبر بين طريقتي المعادلة عند الدرجات الخام الأقل من (6)، في جميع الكُتيبات ماعدا الكُتيب (3، 4، 12)، بينما تختلف القيم المعادلة في الطريقتين عند الدرجات الأعلى من الدرجة 6 في جميع الكتيبات، وخاصة عند الدرجات الوسطية من (24-8) في الكُتيبات (3، 4، 8، 12). وفي المقابل عند الدرجات المرتفعة، تظهر الكتيبات (10، 12، 13) اختلافًا أكبر بين طريقتي المعادلة.

شكل (2)

قيم الدرجات المعادلة بطريقتي المعادلة (الحقيقية والمشاهدة) لكتيبات اختبار PIRLS 2011



ثالثًا- إجابة السؤال الثالث: هل تختلف قيم الدرجات المتعادلة لكتيبات اختبار (PIRLS 2011)، بين طريقتي المعادلة بالدرجات الحقيقية والمعادلة بالدرجات المشاهدة؟

يوضح الجدول (7) المتوسطات الحسابية، والانحرافات المعيارية، وقيمة اختبار "ت" للعينات المرتبطة، وحجم الأثر للفروق في المتوسطات الحسابية للدرجات الخام المعادلة في كتيبات اختبار (PIRLS 2011)، باختلاف طريقتي المعادلة الدرجات الحقيقية والدرجات المشاهدة.

جدول (7)

الاحصاءات الوصفية ونتائج اختبار "ت" وحجم الأثر للفروق في نتائج القيم المعادلة في كتيبات PIRLS 2011 بطريقتي المعادلة

الكتيب	طريقة المعادلة	المتوسط الحسابي	الانحراف المعياري	قيمة "ت"	درجات الحرية	الدلالة الاحصائية	حجم الأثر
2	الدرجات الحقيقية	9.75	6.29	5.282	683	0.000	0.20
	الدرجات المشاهدة	9.76	6.32				
3	الدرجات الحقيقية	9.60	6.05	2.655	675	0.008	0.10
	الدرجات المشاهدة	9.61	6.16				
4	الدرجات الحقيقية	9.78	6.08	0.100	682	0.920	0.003
	الدرجات المشاهدة	9.78	6.16				
5	الدرجات الحقيقية	9.14	6.02	0.596	691	0.552	0.02
	الدرجات المشاهدة	9.14	5.096				
6	الدرجات الحقيقية	9.87	6.39	2.629	692	0.009	0.09
	الدرجات المشاهدة	9.88	6.32				
7	الدرجات الحقيقية	9.29	5.95	13.07	690	0.00	0.49
	الدرجات المشاهدة	9.38	5.89				
8	الدرجات الحقيقية	9.43	6.12	14.26	692	0.00	0.54
	الدرجات المشاهدة	9.54	6.10				
9	الدرجات الحقيقية	9.63	6.14	10.93	690	0.00	0.41
	الدرجات المشاهدة	9.65	6.11				
10	الدرجات الحقيقية	10.29	6.58	12.66	687	0.00	0.48
	الدرجات المشاهدة	10.27	6.55				
11	الدرجات الحقيقية	9.59	6.32	11.88	693	0.00	0.45
	الدرجات المشاهدة	9.63	6.30				
12	الدرجات الحقيقية	9.37	5.86	7.26	689	0.00	0.28
	الدرجات المشاهدة	9.42	5.94				
13	الدرجات الحقيقية	9.89	7.69	23.04	2117	0.000	0.50
	الدرجات المشاهدة	10.0	7.62				

أظهرت نتائج اختبار "ت" للعينات المرتبطة، وجود فروق ذات دلالة إحصائية عند مستوى دلالة (0.05)، بين متوسطات القيم المعادلة في جميع كُتيبات الاختبار ما عدا الكتيب (4) و (5)، تُعزى إلى

الاختلاف في طريقة المعادلة بالدرجات الحقيقية والدرجات المشاهدة، وكانت الفروق لصالح طريقة المعادلة بالدرجات المشاهدة، بينما كانت الفروق في الكتيب (10) لصالح طريقة المعادلة بالدرجات الحقيقية. في حين لم تكن الفروق في متوسطات الدرجات الخام في الكتيب (4) والكُتيب (5) دالة إحصائيًا بين طريقتي المعادلة؛ أي أن الطريقتين تعطيان درجات خام معادلة متشابهة في متوسطاتها الحسابية.

المناقشة والاستنتاجات

هدفت الدراسة إلى المقارنة بين طريقة معادلة الدرجات الحقيقية، وطريقة معادلة الدرجات المشاهدة في إجراء معادلة كتيبات الاختبارات المكونة من فقرات مختلطة: (ثنائية الاستجابة ومتعددة الاستجابة)، باستخدام تصميم المجموعة الواحدة، وتصميم المجموعات غير المتكافئة ذات المفردات المشتركة. وأشارت النتائج إلى وجود اختلاف بين إحصاءات الدرجات الخام قبل المعادلة للكتيبات، وخصوصًا في الكتيبات (6، 8، 10) مقارنة بالكُتيب المرجعي؛ ويرجع ذلك إلى اختلاف مستوى الصعوبة بين الكتيب المرجعي وهذه الكتيبات. وقد يفسر ذلك ما ذهب إليه بعض الباحثين أمثال Baker (2001) وBranberg (2010)، في أنه من الصعب الحصول على نُسخ متكافئة في خصائصها السيكومترية، على الرغم من حرص معدي الاختبارات على ذلك، مما يسبب اختلاف في إحصاءاتها الوصفية، بسبب اختلاف مستوى قدرة الأفراد المختبرين، والظروف المحيطة بهم، وتفاعل الطلبة مع المفردات الاختبارية. أما بعد المعادلة فيتضح وجود تقارب كبير بين الإحصاءات الوصفية للدرجات المعادلة للكتيبات في طريقتي المعادلة مع الإحصاءات الوصفية للكتيب المرجعي.

وعند دراسة الدرجات الخام الناتجة من المعادلة المقابلة لكل درجة من درجات اختبار (PIRLS 2011)، أسفرت نتائج الدراسة عن وجود اختلاف الدرجات الخام المعادلة بين كل من الكُتيبات الاثني عشر في اختبار (PIRLS 2011)، والكتيب المرجعي، وظهرت الفروق الأكبر عند الدرجات الخام الأكبر من (6)، وقد يرجع ذلك إلى عدم تكافؤ مجموعات الطلاب المختبرين، حيث إن "Spence" المشار إليه في بركات (2010)، توصل إلى أن نتائج عملية المعادلة تتأثر في الحالات التي تكون فيها مجموعات المفحوصين غير متكافئة.

كما أظهرت النتائج وجود فروق ذات دلالة إحصائية بين طريقتي المعادلة، لصالح طريقة المعادلة بالدرجات المشاهدة في جميع الكُتيبات، ما عدا الكُتيب (10)، حيث كانت الفروق لصالح المعادلة بالدرجات الحقيقية، والكُتيبات (4) و(5)، حيث كانت الفروق غير دالة إحصائيًا بين طريقتي المعادلة. ويرجع اختلاف القيم المعادلة باختلاف طريقتي المعادلة إلى اختلاف الصعوبة بين النصوص المكونة لكل كُتيب، فيشير المحرزي (2015) إلى أن هناك عوامل تؤثر على نتائج كل طريقة من طرق المعادلة،

والـذي يـؤدي إلى الاختـلاف بينهـا، ومنهـا مقـدار الاختـلاف بـين نُسـخ الاختبـار في الإحصـاءات الوصفيـة والتوزيعـات التكراريـة، وعـدد أسـئلة الاختبـار، وعـدد المفـردات المشـتركة، وخصائصهـا السـيكومترية بصفـة خاصـة.

ويسـبب الاختـلاف بـين طريقتـي المعادلـة أيضًـا نمـوذج نظريـة الاسـتجابة للمفـردة المسـتخدم في التدريـج، وتصميـم جمـع البيانـات. كـما يضيـف Kolen & Brennan (2004)، أن اختـلاف النتائـج باسـتخدام نـماذج نظريـة الاسـتجابة للمفـردة، يعتمـد عـلى نوعيـة البرنامـج الحاسـوبي المسـتخدم في تقديـر بارامـترات المفـردات، ونوعيـة المعايـرة، سـواء كانـت متزامنـة أم منفصلـة، والإجـراءات المسـتخدمة لربـط النتائـج المختلفـة التـي نحصـل عليهـا مـن البرنامـج الحاسـوبي، مثـل طريقـة منحنـى الاختبـار، أو طريقـة المتوسـط/ الانحـراف.

وتـدل نتائـج الدراسـة عـلى أهميـة معادلـة نـماذج الاختبـارات في تقليـص الاختـلاف بـين مسـتويات الصعوبـة، بحيـث يتـم تعديـل تدريـج الكُتيـب الجديـد ليوافـق تدريـج الكتيـب المرجعـي، وبالتـالي يصبـح مـن السـهل اسـتخدام الدرجـات المتعادلـة بشـكل متبـادل للحكـم عـلى أداء المفحـوص عـلى هـذه النـماذج.

عـلى الرغـم النتائـج الإيجابيـة لهـذه الدراسـة، فإنهـا اقتـصرت عـلى اسـتخدام كُتيبـات اختبـار قيـاس مهـارات مـدى تقـدم القـراءة "بيرلـز" للعـام الـدراسي 2011، وعددهـا 13 كُتيبًـا لإجـراء المعادلـة. كـما تـم اسـتخدام نمـوذج الاسـتجابة المتدرجـة مـن نـماذج نظريـة الاسـتجابة للمفـردة في تدريـج نـماذج اختبـار "بيرلـز" 2011، كـما اسـتخدمت الدراسـة تصميـم المجموعـة الواحـدة، وتصميـم المجموعـات غـير المتكافئـة ذات المفـردات المشـتركة في نظريـة الاسـتجابة للمفـردة لإجـراء المعادلـة بـين النـماذج، باسـتخدام معادلـة الدرجـات الحقيقيـة، ومعادلـة الدرجـات المشـاهدة، وتعتمـد دقـة نتائـج الدراسـة عـلى دقـة البرمجيـات المسـتخدمة في عمليـة المعادلـة.

التوصيات

تقدم الدراسة مجموعة من التوصيات كالآتي:

- اسـتخدام نتائـج معادلـة اختبـار (PIRLS 2011)، المتحصـل عليهـا مـن الدراسـة للمقارنـة بـين الطلبـة باختـلاف النـماذج المطبقـة.
- تدريـب العاملـين في التقويـم التربـوي عـلى كيفيـة معادلـة الاختبـارات، باسـتخدام النظريـة الحديثـة، للتقليـل مـن مشـكلة هـدر الوقـت في بنـاء اختبـارات جديـدة كل مـرة.
- اسـتخدام نمـوذج الاسـتجابة المتدرجـة كأحـد نـماذج النظريـة الحديثـة في بنـاء الاختبـارات ومعادلتهـا؛ لأنـه يجمـع بـين النوعـين: مـن المفـردات ثنائيـة الاسـتجابة، ومتعـددة الاسـتجابة.

المقترحات

تقدم الدراسة مجموعة من المقترحات البحثية كالآتي:

- إجراء دراسات لمعادلة كُتيبات (PIRLS 2011)، باستخدام التصاميم، والطرق الأخرى التابعة للنظرية الحديثة في القياس.
- إجراء مقارنة بين معادلة كُتيبات اختبار (PIRLS 2011)، بالنظرية الكلاسيكية، ومقارنتها مع نتائج المعادلة التابعة للنظرية الحديثة.
- إجراء معادلة كُتيبات اختبار (PIRLS 2016)، لمعرفة القيم المتعادلة مع كُتيبات اختبار (PIRLS 2011)، للوقوف على مدى التقدم في مستوى التحصيل لدى طلاب الصف الرابع الأساسي؛ بسبب القيمة التي يمثلها هذا الاختبار، وأهمية القرارات المبنية على نتائجه.

المراجع

المراجع العربية:

-أيوب، حسين محمد (1994). **المقارنة بين أربع طرق للمعادلة عندما يكون التصميم من مجموعات متكافئة وغير متكافئة**. الأردن: الجامعة الأردنية.

-بركات، مايا إبراهيم (2010). **أثر تصميمات المعادلة ومتوسط صعوبة الاختبارات وتوزيع القدرة على معادلة درجات الاختبارات متعددة الأبعاد باستخدام نظرية الاستجابة للمفردة**، (رسالة دكتوراه غير منشورة)، جامعة القاهرة، جمهورية مصر العربية.

-الدوسري، راشد حماد (2001). معادلة الاختبارات، مفهومها، وطرقها، ومشكلات تطبيقها. **مجلة العلوم النفسية والتربوية**، البحرين، 2(4)، 106-141.

-الشمري، مها مطلق، والشريفين، نضال (2015). **معادلة درجات نُسخ مختلفة من اختبار القدرات المعرفية لدى طلبة الثانوية العامة بالمملكة العربية السعودية**، (دراسة ماجستير غير منشورة)، جامعة اليرموك، اربد.

-طيفور، مصطفى أحمد (2007). **دراسة مقارنة لنماذج نظرية الاستجابة للمفردة في معادلة درجات الاختبارات**. معهد القاهرة: معهد الدراسات التربوية.

-علام، صلاح الدين محمود (2005). **نماذج الاستجابة للمفردة الاختبارية أحادية البعد ومتعددة الأبعاد وتطبيقاتها في القياس النفسي والتربوي**. القاهرة: دار الفكر العربي.

-عودة، أحمد سليمان، وعبيدات، عمر سليمان (2013). فاعلية الاختبار التكيفي المحوسب في تقدير القدرة العقلية باستخدام مصفوفات رافن. **دراسات العلوم التربوية**، 2(40)، 1602-1621.

-المحرزي، راشد بن سيف (2014). المقارنة بين طرق المعادلة الكلاسيكية لدرجات نماذج اختبار القدرات العامة باستخدام تصميم الجماعات المتكافئة. **رسالة الخليج**، 134، 15-42.

-المحرزي، راشد بن سيف (2015). المفاضلة بين الدرجات المكافئة لنماذج اختبار القدرات العامة باستخدام طرق المعادلة الكلاسيكية في تصميم المفردات المشتركة بجماعات غير متكافئة. **مجلة العلوم التربوية والنفسية**، 16(3)، 394-429.

-محمد، محمد حبشي حسين (2006). تكافؤ القياس بين النسختين العربية والإنجليزية لاستبيان مؤشر أساليب التعلم في ضوء نظرية الاستجابة للمفردة. **مجلة دراسات نفسية**، مصر، 16(4)، 537-591.

-المدانات، رائد فايز (2012). مقارنة فاعلية طريقتي معادلة الدرجات الحقيقية والمشاهدة في معادلة الاختبارات باستخدام جذع مشترك ومجموعات غير متكافئة. **مجلة العلوم النفسية والتربوية**، البحرين، 13(2)، 365-394.

-المديرية العامة للتقويم التربوي (2011). **التقرير الوطني للدراسة الدولية لقياس مهارات القراءة PIRLS 2011**. سلطنة عمان.

-هيبة، محمد أحمد علي، وعمر، محمود أحمد (2011). تكافؤ قياس القائمة المختصرة للعوامل الخمسة للشخصية بين الجنسين في ضوء نظرية الاستجابة للمفردة ونمذجة المعادلة البنائية. **مجلة القراءة والمعرفة**، مصر، 115، 91-131.

-وزارة التربية والتعليم (2014). **دليل الدراسة الدولية لقياس مهارات القراءة (PIRLS)**. سلطنة عمان.

-الوليلي، اسماعيل حسن فهيم (2005). تكافؤ درجات الاختبارات في ضوء نظريتي القياس الكلاسيكية والحديثة، دراسة سيكومترية مقارنة. **مجلة كلية التربية، جامعة بنها**، مصر، 15(63)، 98-149.

المراجع الإنجليزية:

-Andersson, B., Branberg, K., & Wiberg, M. (2013). Performing the kernel method of test equating with the package equate. *Journal of Statistical Software*, *55*(6), 1-25.

-Baker, F. B. (2001). *The basics of item response theory*. USA, Eric.

-Baker, F. B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement*, 16(1), 87-96.

-Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement*, *28*(2), 147-162.

-Battauz, M. (2015). EquateIRT, An R package for IRT test equating. *Journal of Statistical Software*, *68*(7), 1-22.

-Branberg, K. (2010). *Observed score equating with covariates* (Doctoral dissertation, department of statistics, Umea University).

-Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications Inc.

-IEA's Progress in international reading literacy study. (2011). *Methods and procedures- sampling implementation*. Chestnut Hill, Boston College. Retrieved on December 18, 2016 from: http:timssandpirls.bc.edu.

-IEA's Progress in international reading literacy study. (2011). *Population coverage and sample participation rates-Appendix-C*. Chestnut Hill: Boston College.

-Ju, L. C. (2008). Comparisons between classical test theory and item response theory in automated assembly of parallel test forms. *The Journal of Technology, Learning and Assessment*, *6*(8), 4-42.

-Kabasakal, K. A., & Kelecioglu, H. (2015). Effect of differential item functioning on test equating. *Educational Sciences: Theory and Practice*, *15*(5), 1229-1246.

-Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: methods and practice*. New York: Springer.

-Li, Y. H., Lissitz, R. W., & Yang, Y. N. (1999). *Estimating IRT equating coefficients with polytomously and dichotomously scored items*. Paper presented at the annual meeting of the National Council on Measurement in Education, (Montreal, April 19-23, 1999).

-Livingston, S. A. (2004). *Equating test scores (without IRT)*, Educational Testing Service. Princeton.

-Marian, S., & Jay, C. (2001). Developing the PIRLS reading assessment. In Campbell, J. R., Kelly, D. L., Mullis, I. V. S., Martin, M. O., & Sainsbury, M. (2001), *Framework and* specifications for PIRLS assessment (2001), (2nd ed.). Chestnut Hill, MA: Boston College.

-Mullis, V. S., Martin, O., Kennedy, A. M., Tong, L., & Sainsbury, M. (2009). *PIRLS (2011), assessment framework*. Chestnut Hill: Boston College.

-Ozturk-Gubes, N., & Kelecioglu, H. (2016). The impact of test dimensionality, common-item set format, and scale linking methods on mixed-format test equating. *Educational Sciences: Theory and practice*, *16*(3), 715-734.

-Taha, H. (2008). The status of Arabic language today. *Journal of Education, Business and Society, Contemporary Middle Eastern Issues*, *1*(3), 186-192.

-United Nations Development Program (2009). *Human Development Report, Overcoming Barriers, Human Mobility and Development*. NY: United Nations.