أثر اختلاف معامل ثبات صورتي اختبار في دقة معادلة درجاتهما باستخدام الطرق القائمة على نظرية الاستجابة للمفردة

د. محمد محمود محمد عبد الوهاب*

المقدمة

زاد في الآونة الأخيرة اهتمام التربويين بضرورة تمكين الطلاب من إعادة الاختبار أكثر من مرة، مع اتخاذ الإجراءات المناسبة للتغلب على مشكلة سرية الاختبارات، كما نادى المتخصصون في القياس النفسي بأهمية تجنب أثر القياسات المتكررة للسمة أو القدرة على الطالب عند الرغبة في متابعة تقدمه بعد عمليات التعليم والتعلم والتدريب والمران؛ وهذا كله أدى إلى ضرورة وجود عدة صور اختبارية تقيس نفس السمة أو القدرة، كما ظهرت الحاجة إلى المقارنة بين طلاب المناطق التعليمية المختلفة الذين تعرضوا لاختبارات مختلفة لنفس السمة أو القدرة المقيسة؛ كل هذا أدى إلى البحث عن سبل تحقيق العدالة بين المفحوصين الذين يتعرضون إلى صور اختبارية مختلفة، فظهر تبعًا لذلك ما يعرف بمعادلة درجات الصور الاختبارية المختلفة.

وتعرف معادلة درجات الاختبارات Test Scores Equating بتلك العملية الإحصائية التي يمكن من خلالها تحويل نظام وحدات القياس التي تم الحصول عليها من صورة اختبارية إلى نظام وحدات قياس صورة اختبارية أخرى تقيس نفس السمة (González, Barrientos & Quintana, 2015; Kilmen & Demirtasli, 2012)، بحيث تصبح القياسات المستمدة من الصورتين متكافئة بعد إجراء التحويل (علام، ٢٠٠٥؛ Lord, 1980). وبذلك فإن عملية معادلة درجات الاختبارات تهدف إلى التغلب على التحيز الناتج عن اختلاف صعوبة الصور الاختبارية المختلفة لنفس السمة، في محاولة لتحقيق العدالة

* مدرس بقسم علم النفس التربوي – كلية التربية – جامعة المنيا

بين المفحوصين، والإفادة منها في بنوك الأسئلة، والاختبارات التكيفية الحاسوبية (Rui, Shu-Liang & Deng-Wen, 2010)، أي أن عملية المعادلة تهدف إلى إزالة الفروق بين مستويات صعوبة مفردات نموذجين مختلفين من الاختبار ومستويات تمييزها بهدف تحقيق التكافؤ في القياسات الناتجة عنهما (Kolen & Brennan, 1995).

وقد ظهر نوعان من معادلة درجات الاختبارات، أولهما: المعادلة الأفقية Horizontal equation التي يتم فيها معادلة درجات صورتي اختبار تقيسان نفس السمة على عينتين من المفحوصين لهما نفس توزيع القدرة تقريبًا ومن نفس المجتمع، ويكون الهدف منها في الغالب ضمان سرية الاختبارات مع تحقيق العدالة بين المفحوصين، وثانيهما: المعادلة العمودية Vertical equation وتتم للاختبارات التي تتضمن مستويات مختلفة من الصعوبة على مفحوصين توزيع قدراتهم مختلف من مستوى إلى آخر (Moghadamzadeh, Salehi & Khodaie, 2011)، ويكون الهدف منها الحصول على ميزان مشترك أو مقياس موحد Common scale يمتد عبر عدد من الصفوف الدراسية أو المراحل العمرية المختلفة (الرحيل، ٢٠١٣؛ علام، ٢٠٠٥)، وتنطبق المعادلة العمودية على بطاريات الاختبارات، كما تستخدم عندما تكون القدرة أو السمة المقيسة على متصل واسع المدى أو متعدد المستويات كما في السمات التطورية (الشريفين، ٢٠٠٩).

وقد أورد كثير من المتخصصين في القياس النفسي عددًا من طرق معادلة درجات الاختبارات، والتي صنفت إلى طرق تعتمد على النظرية الكلاسيكية Classical Test Theory، مثل: طريقة المتوسط الحسابي Mean Equating، والطريقة الخطية Linear Equating، ومنها طريقة تكر Tucker، وطريقة ليفين Levine سواء في حالة تساوي الثبات أو في حالة

عدم تساوي الثبات، وطريقة براون وهولند Braun & Holland، ومنها أيضًا طريقة المعادلة المئينية Equipercentile Equating، سواء بطريقة التكرارات الممهدة أو غير الممهدة، وطرق أخرى تعتمد على نظرية الاستجابة للمفردة Item Response Theory، هذا إضافة إلى طرق أخرى، مثل: طريقة اللب أو النواة Kernel (Cook, Eignor & Schmitt, 1990; Hambleton, Swaminathan & Rogers, 1991; Kilmen & Demirtasli, 2012; Kolen, 1988).

ومن طرق معادلة درجات الاختبارات باستخدام نظرية الاستجابة للمفردة طريقة المتوسط/ المتوسط التي قدمها لويد وهووفر (Loyd & Hoover, 1980) وتعتمد على تحويل بارامترات الصعوبة والتمييز لمفردات الاختبار، ويستخدم متوسط بارامترات التمييز في تحديد منحنى المعادلة، ومتوسط بارامترات الصعوبة في تحديد ثابت المعادلة، وطريقة المتوسط/ سيجما التي قدمها ماركو (Marco, 1977) وفيها يستخدم الإنحراف المعياري في تحديد منحنى المعادلة، ويستخدم متوسط بارامترات الصعوبة في تحديد ثابت المعادلة، وطريقة هايبارا (Haebara, 1980) التي يتم الحصول على منحنى المعادلة وثابت المعادلة من خلال الفرق بين منحنيات خصائص المفردات، حيث يكون الفرق بين منحنيات خصائص المفردات عبارة عن مجموع مربعات الفروق بين منحنيات خصائص المفردات الخاصة بكل مفردة، ثم يتم حساب ثابت المعادلة ومنحنى المعادلة المصغر لهذا الفرق، وطريقة ستوكنج – لورد (Stocking-Lord, 1983) وفيها أيضًا يتم الحصول على منحنى المعادلة وثابت المعادلة من خلال الفرق بين منحنيات خصائص المفردات، ولكن الفرق بين منحنيات خصائص المفردات في هذه الحالة عبارة عن مربع مجموع الفروق بين منحنيات خصائص المفردات الخاصة بكل مفردة، ثم يتم حساب ثابت المعادلة ومنحنى المعادلة لأقل فرق (in: Kilmen & Demirtasli, 2012).

وقد أكدت نتائج الدراسات وجود اختلاف في دقة معادلة درجات الاختبارات يعود إلى الطريقة المستخدمة في المعادلة نفسها أو إلى نموذج الاستجابة للمفردة المستخدم في تقدير بارامترات الأفراد والمفردات، ومن هذه الدراسات دراسة سكاجس وليسيتز (Skaggs & Lissitz, 1986) التي هدفت إلى مقارنة أربع طرق مختلفة لمعادلة درجات اختبارين لهما خصائص سيكومترية مختلفة، وهي: الطريقة الخطية، والطريقة المئينية، وباستخدام نموذج راش، وباستخدام النموذج ثلاثي البارامتر؛ وأسفرت نتائج المعادلة العمودية عن تحقيق النموذج ثلاثي البارامتر أعلى دقة في معادلة الدرجات، بينما كانت المعادلة الخطية غير دقيقة على الإطلاق، وكانت نتائج الطريقة المئينية جيدة بشكل عام عندما تختلف الاختبارات بمقدار واحد لوجيت فقط في صعوبتها، أما عند وجود اختلاف مقداره (٢) لوجيت، فإنه ستظهر أخطاء في المعادلة، ولم يحقق نموذج راش أية دقة في معادلة الدرجات إلا عند مطابقة البيانات في نموذجي الاختبار لافتراضات نموذج راش، أي عندما تتساوى بارامترات تمييز جميع المفردات وعدم وجود أية فرصة للتخمين، كما أكدت النتائج أفضلية نموذج راش عن النموذج ثلاثي البارامتر في حالة تحقق افتراضات نموذج راش، أما في باقي الحالات، فإن النموذج ثلاثي البارامتر يحقق أفضلية كبيرة عن نموذج راش.

وعلى نقيض ما توصل إليه سكاجس وليسيتز (Skaggs & Lissitz, 1986) فإن دراسة طيفور (٢٠٠٧) قد توصلت إلى أن النموذج الأحادي البارامتر أكثر النماذج دقة في معادلة درجات الاختبارات، يليه النموذج الثنائي البارامتر، وكان النموذج الثلاثي البارامتر أقل النماذج الثلاثة دقة، وذلك عند استخدام تصميم المفردات المشتركة، وكانت النماذج الثلاثة متكافئة في معادلة درجات الاختبارات

باستخدام تصميم الأفراد المشتركين، أما عند استخدام تصميم المجموعات المتكافئة، فإن النموذج الأحادي البارامتر أكثر النماذج الثلاثة دقة في معادلة درجات الاختبارات، أما النموذجان الثنائي والثلاثي البارامتر فكانا متكافئين.

كما توصلت بالخيور (٢٠٠٩) إلى أن استخدام النموذج أحادي البارامتر أدى إلى نتائج أفضل من النموذجين الثنائي والثلاثي البارامتر عند الأحجام المختلفة للعينة (١٠٠، ٥٠٠، ١٠٠٠ مفحوص)، وذلك في دقة واستقرار نتائج معادلة درجات النماذج المختلفة لاختبار القدرات العامة بالمركز الوطني للقياس والتقويم المستخدم في استكمال متطلبات القبول بجامعة أم القرى، كما أن طريقة المتوسط / المتوسط أفضل من طريقة المتوسط / سيجما، وأن الطرق المعتمدة على نظرية الاستجابة للمفردة كانت أفضل من الطرق الكلاسيكية بشكل عام عند الأحجام المختلفة للعينة.

وقد قارن سونج (Song, 2009) بين نماذج الاستجابة للمفردة ثنائية الاستجابة dichotomous IRT models وخليط من النماذج ثنائية الاستجابة والنماذج متعددة الاستجابات Polytomous IRT models في معادلة درجات صورتين من امتحان شهادة الكفاءة في اللغة الإنجليزية، وقد أظهرت النتائج أن الخليط من النموذج ثلاثي البارامتر ونموذج التقدير الجزئي المعمم أعطى نتائج مشابهة للطرق التقليدية في جزء الاستماع، وكانت هذه النتائج أفضل من نتائج استخدام النموذج ثلاثي البارامتر خاصة عند المستويات المنخفضة والمتوسطة، بينما كان النموذج ثلاثي البارامتر أفضل في جزء القواعد وملء الفراغات وحصيلة المفردات اللغوية والقراءة؛ وقد أرجع هذا الاختلاف بين جزءي الاختبار إلى درجة انتهاك افتراض الاستقلال الموضعي Local Independence للمفردات في كل جزء.

وتوصلت دراسة أبو مسلم (٢٠١٠) إلى أن طريقة المعادلة المتوسط / سيجما التي تقوم على نظرية الاستجابة للمفردة كانت أكثر فاعلية من طريقة المتوسط / المتوسط عندما تمت معادلة درجات صيغتي اختبار توني Tony للذكاء، كما تفوقت في الوقت نفسه على طرق المعادلة القائمة على النظرية الكلاسيكية سواء الطريقة الخطية أو طريقة المئينيات.

وقد قام بانج ومادیرا ورضوان وزانج (Pang, Madera, Radwan & Zhang, 2010) بالكشف عن فاعلية أربع طرق لمعادلة درجات الاختبارات، وهي: المعادلة بالتدريج المتداخل Concurrent Calibration Equating، والمعادلة عن طريق بارامترات المفردات المشتركة المثبتة Fixed Common Item Parameter Equating، والمعادلة عن طريق منحنى خصائص الاختبار للورد وستوكنج، والمعادلة بطريقة المتوسط / سيجما؛ وقد أظهرت النتائج تفوق طرق المعادلة التي تعتمد على التدريج المنفصل للمفردات، مثل: طريقة بارامترات المفردات المشتركة وطريقة لورد وستوكنج وطريقة المتوسط / سيجما على المعادلة بالتدريج المتداخل.

كما سعى بعض الباحثين إلى تعرف العوامل التي قد تؤثر في دقة معادلة درجات الاختبارات، فقد قام المحروق (٢٠١١) بتقصي أثر طول الاختبار (٣٠، ٤٠، ٦٠) مفردة، وحجم العينة (٢٠٠، ٦٠٠، ١٠٠٠) مفحوص، ومستويات الصعوبة (متشابهة – مختلفة) في دقة معادلة درجات الاختبارات باستخدام طريقة كيرنيل وطرق نظرية الاستجابة للمفردة، وذلك باستخدام تصميم المفردات المشتركة، عن طريق توليد بيانات تجريبية ببرنامج Wingen2؛ وقد أكدت النتائج أن طريقة المعادلة باستخدام نظرية الاستجابة للمفردة كانت الأكثر دقة عند مختلف أطوال الاختبار وأحجام العينة

ومستويات الصعوبة، وذلك بحساب جذر متوسط مربع الخطأ The root mean squared error (RMSE).

وقد سعى كيلمين وديميرتاسلي (Kilmen & Demirtasli, 2012) إلى دراسة أثر حجم العينة (٥٠٠، ١٠٠٠) مفحوص، وتوزيع قدرات المفحوصين (متشابه، مختلف) في دقة أربع طرق معادلة للدرجات قائمة على نظرية الاستجابة للمفردة، وهي: طريقة المتوسط / المتوسط، وطريقة المتوسط / سيجما، وطريقة هايبارا، وطريقة ستوكنج لورد؛ وقد أظهرت النتائج أن طريقة ستوكنج لورد تعطي أقل خطأ للمعادلة في حالة اختلاف حجم العينة واختلاف توزيع قدرات المفحوصين.

كما قام وولف (Wolf, 2013) بتقصي أثر البنية الاختبارية (أحادية البعد أو متعددة الأبعاد)، وتوزيع قدرات المفحوصين في المجموعات (متكافئة أو غير متكافئة) في قدرة طرق المعادلة المختلفة على تحقيق خصائص المساواة Equity Properties، وقد توصل إلى أن طرق المعادلة القائمة على نظرية الاستجابة للمفردة من خلال الدرجات الحقيقية أفضل الطرق في تحقيق خصائص المساواة بغض النظر عن البنية الاختبارية، وقد كانت طرق المعادلة القائمة على نظرية الاستجابة للمفردة من خلال الدرجات الملاحظة أفضل الطرق في حالة زيادة الفروق بين متوسطات قدرات المفحوصين أي في حالة عدم تكافؤ توزيع القدرات في المجموعتين.

ويتضح مما سبق تفوق الطرق القائمة على نظرية الاستجابة للمفردة بشكل عام في معادلة درجات الاختبارات على الطرق القائمة على النظرية الكلاسيكية، وقد أشار الدوسري (٢٠٠١)، وعلام (٢٠٠٥) ولامبريانو (Lamprianou, 2007)وكذلك موجادمزاده(Moghadamzadeh, et al., 2011) إلى أن ذلك التفوق يرجع إلى عدم قدرة الطرق الكلاسيكية على تحقيق بعض شروط المعادلة، مثل: شرط استقلالية التحويلات Transformation عن

مجتمعات الأفراد المختبرين عند استخدام الطرق المعتمدة على النظرية الكلاسيكية في إجراء المعادلة، كما أكد هامبلتون وجونز (Hambleton & Jones, 1993) ذلك عندما أشارا إلى أن أهم مزايا استخدام طرق نظرية الاستجابة للمفردة هو تحقيق استقلال إحصائيات المفردات عن خصائص المفحوصين، واستقلال قدرات المفحوصين عن بارامترات المفردات والاختبار فيما يعرف باللاتغاير أو اللاتباين Invariance.

كما يتضح مما سبق وجود اختلاف بين طرق المعادلة القائمة على نظرية الاستجابة للمفردة سواء كانت طريقة المتوسط / المتوسط أو المتوسط / سيجما أو طريقة هاييارا، أو طريقة ستوكنج لورد أو غيرها في دقة معادلة درجات الاختبارات، وذلك بسبب وجود أخطاء تؤثر في دقة المعادلة، والتي منها: الأخطاء في تقدير بارامترات المفردات، والأخطاء الناتجة عن عدم ثبات الاختبار (Ryan & Brockmann, 2011)، والأخطاء الناتجة عن عدم ملاءمة نموذج الاستجابة للمفردة المستخدم في تحليل المفردات، والخطأ في تقدير الخطأ العشوائي للمفحوصين أو الخطأ العشوائي للمفردات، والخطأ الناتج عن تغيير في المفردات وموضعها أو ترتيبها، والخطأ الناتج عن اختيار الطريقة المناسبة للمعادلة (Xu, 2009).

وقد أشار الدوسري (٢٠٠١) إلى نوعين من الخطأ يؤثران في معادلة الاختبارات، أولهما: الخطأ العشوائي Random error، وهو خطأ يحدث دائمًا طالما تم استخدام عينات من المجتمع، ويمكن خفض هذا الخطأ باستخدام عينات كبيرة من المفحوصين المطبق عليهم معادلة الاختبارات، واختيار التصميم الأنسب للمعادلة، وثانيهما الخطأ المنتظم Systematic error، وهو خطأ ينتج عن عدم مراعاة العوامل التي قد تؤثر في أداء المفحوصين على مفردات الاختبار، مثل: التعب والخبرة

والممارسة، كما ينتج عن اختلاف صورتي الاختبار في محتوى مفرداتهما وفي صعوبتهما وفي ثبات درجاتهما.

ولعل أحد مصادر الخطأ في معادلة درجات الاختبارات هو التصميم المستخدم في المعادلة، ومن هذه التصميمات تصميم المجموعة الواحدة الذي يتم فيه تطبيق صورتي الاختبار المراد معادلة درجاتهما على نفس المفحوصين بالتناوب، ويفضل أن يكون عدد المفحوصين كبيرًا حتى يتم الحد من خطأ المعادلة، وكذلك تصميم المجموعات المتكافئة الذي يتم فيه تكوين مجموعتين عشوائيتين من المفحوصين، والتحقق من تكافؤهما، ثم تعطى كل مجموعة منهما إحدى صورتي الاختبار المراد معادلة درجاتهما، والتصميم القائم على الجذع المشترك ومجموعات غير متكافئة الذي يتم فيه تطبيق الصور الاختبارية على مجموعتين مختلفتين من المفحوصين في توزيع قدراتهم، مع إدراج مجموعة من المفردات في كل من صورتي الاختبار كجذع مشترك Anchor أو مفردات مشتركة Common items، وتصميم الأفراد المشتركين الذي يتم فيه تطبيق صورتي الاختبار على مجموعتين من المفحوصين، مع وجود مجموعة من المفحوصين يطبق عليهم كلتا الصورتين (Moghadamzadeh, Salehi & Khodaie, 2011).

هذا، ويتم الحكم على فاعلية معادلة الاختبارات من خلال استقرار النتائج ودقتها، ويستخدم في معرفة مدى استقرار النتائج معامل الصدق التقاطعي Cross Validation Coefficient لمعرفة مدى استقرار النتائج (الحواري، ٢٠٠٧؛ الرحيل، ٢٠١٣)، كما يستخدم جذر متوسط مربعات الخطأ RMSE كمؤشر لدقة عملية المعادلة (بالخيور، ٢٠٠٩؛ Kim, 2013).

مشكلة البحث

وضع كثير من المتخصصين عددًا من الشروط اللازم تحققها في عملية معادلة الاختبارات، وهي: ضرورة قياس الاختبارات لنفس السمة أو القدرة، وتحقق خاصية المساواة Equity التي تعني أن يكون التوزيع التكراري المشروط للدرجات عند مستوى معين من مستويات القدرة بعد معادلة الدرجات متماثلًا في صورتي الاختبار، وبذلك تكون الدرجات قابلة للتبادل بين الصورتين، وكذلك اللاتباين أو اللاتغاير في المجتمع Population invariance والذي يعني أن تحويل الدرجات يجب أن يبقى كما هو بصرف النظر عن مجموعة المفحوصين، وكذلك تحقق خاصية التماثل Symmetry الذي يقصد به قابلية عملية تحويل الدرجات من صورة إلى أخرى للانعكاس Invertible بمعنى أن الدالة التي يتم تحويل الدرجات بها من الصورة الاختبارية الأولى إلى الثانية تؤدي النتائج نفسها إذا تم تحويل الدرجات من الصورة الثانية إلى الأولى (Holland & Dorans, 2006; Kolen & Brennan, 1995; Lord, 1980; Petersen, Kolen & Hoover, 1989; von Davier, Holland & Thayer, 2004).

وقد أضاف متخصصون آخرون شرطًا إضافيًا يتعلق بضرورة تساوي ثبات صورتي الاختبار المراد معادلة درجاتهما للحصول على معادلة دقيقة للدرجات، مثل: لورد (Lord, 1980) وهامبلتون وسواميناثان وروجرز (Hambleton, Swaminathan & Rogers, 1991) ودورانز وهولاند (Dorans & Holland, 2000) والشريفين (٢٠٠٩) وريان وبروكمان (Ryan & Brockmann, 2011) وفون دافير (von Davier, 2011) ودورانز ومزيس وايجنور (Dorans, Moses & Eignor, 2011) ومينج (Meng, 2012)، وقد فسر علام (٢٠٠٥) ذلك بأن محك التبادل Interchangeability بين درجات الاختبارات سيتحقق فقط إذا كانت هذه الاختبارات على درجة واحدة من الثبات. كما أكد لوي وزو وكيرلي وكاري

(Lui, Zu, Curley & Carey, 2014) أن انخفاض ثبات الاختبار قد يسبب خطأ التحيز في المعادلة، وقد شدد الدوسري (٢٠٠١) على ضرورة أن تتمتع الاختبارات المراد معادلة درجاتها بثبات كامل، بينما خفف كروكر وألجينا (Crocker & Algina, 1986) والمدانات (٢٠١٢) من هذا الشرط عندما أشاروا إلى إمكانية معادلة درجات الاختبارات المتقاربة في معامل ثباتها.

بينما أشار لامبريانو (Lamprianou, 2007) إلى أن هناك طرقًا تصلح لمعادلة درجات الاختبارات غير متساوية الثبات، فقد قدم ليفين طريقتين لمعادلة درجات الاختبارات إحداهما تتطلب تساوي الثبات والأخرى تطبق في حالة عدم تساوي الثبات. وقد أكد كينجستون وهولاند (Kingston & Holland, 1986) أن طريقة ليفين للاختبارات مختلفة الثبات تؤدي إلى معادلة أدق وأفضل إذا ما قورن بين طريقتي ليفين سواء أكان ذلك في حالة تساوي طول الاختبار أو عدم تساويه.

ولبحث أثر اختلاف ثبات صورتي الاختبار في معادلة درجاتهما، قام ماركس وليندساي (Marks & Lindsay, 1972) بتقصي أثر عدد المفردات في الصورة الاختبارية وحجم العينة وثبات صورتي الاختبار في دقة معادلة درجات الاختبارين، وكان من بين النتائج التي توصلا إليها عدم وجود فروق دالة إحصائيًا في دقة معادلة الدرجات نتيجة اختلاف ثباتهما، على ألا يقل ثبات أي منهما عن ٠,٧٥ حتى لا تكون نتائج المعادلة مشكوكاً في صحتها.

كما أكد كوك وايجنور (Cook & Eignor, 1991) أنه في حالة اختلاف صعوبة الصور الاختبارية ومحتواها وثباتها فإن عددًا كثيرًا من طرق معادلة الاختبارات لا تصلح لهذا الغرض، وقد أشار دورانز وهولاند (Dorans & Holland, 2000) إلى أن متطلب تساوي الاختبارين المراد معادلة درجاتهما في الثبات غالبًا ما ينتهك، ورغم ذلك فإن عملية المعادلة في هذه الحالة تكون مرضية، وأنه من الناحية العملية لا يوجد دليل على وجود مشكلات في معادلة درجات الاختبارات التي لها

معاملات ثبات مرتفعة ومختلفة اختلافًا واضحًا في الثبات، وبالتالي فإن متطلب تساوي الثبات يعد متطلبًا ثانويًا وليس أساسيًا. وقد قدما شرحًا للعلاقة بين تساوي الثبات ومتطلب المساواة Equity؛ من أجل توضيح أنهما غير مرتبطين ارتباطًا وثيقًا، وقد دعوا إلى ضرورة الاهتمام بارتفاع ثبات الاختبارات المراد معادلة درجاتها أكثر من التحقق من تساويها.

ويلاحظ من الدراسات التي سبق عرضها وجود اختلاف في دقة معادلة درجات الاختبارات يعود إلى الطريقة المستخدمة في عملية المعادلة، أو إلى نموذج الاستجابة للمفردة المستخدم في تدريج مفردات الاختبارات، أو إلى بعض المتغيرات السيكومترية، مثل: طول الاختبار أو حجم العينة أو شكل توزيع قدرات المفحوصين ... ، كما يتضح أنه غالبًا ما تتفوق الطرق القائمة على نظرية الاستجابة للمفردة على الطرق الكلاسيكية، ونظرًا لكون افتراض تساوي ثبات صورتي الاختبار غالبًا ما ينتهك، وأن طريقة ليفين التي تصلح لعملية معادلة الاختبارات غير متساوية الثبات تقوم على النظرية الكلاسيكية في القياس، فإنه لكي يتم الاستفادة من المزايا التي تحققها نظرية الاستجابة للمفردة في معادلة الدرجات يجب تحديد ما إذا كانت دقة طرق المعادلة القائمة على نظرية الاستجابة للمفردة تتأثر بوجود اختلاف بين ثبات صورتي الاختبار أم لا؛ لذا فإن هذا البحث يسعى إلى الإجابة عن الأسئلة الآتية:

١- ما درجة دقة معادلة درجات صورتي الاختبار باستخدام الطرق القائمة على نظرية الاستجابة للمفردة مع اختلاف معاملي ثباتهما؟

٢- هل يؤثر اختلاف معاملي ثبات صورتي الاختبار في دقة معادلة درجاتهما باستخدام الطرق القائمة على نظرية الاستجابة للمفردة؟

أهداف البحث:

يهدف هذا البحث إلى

١- تعرف درجة دقة معادلة درجات صورتي الاختبار باستخدام الطرق القائمة على نظرية الاستجابة للمفردة مع اختلاف معاملي ثباتهما.

٢- الكشف عن أثر اختلاف معاملي ثبات صورتي الاختبار المراد معادلة درجاتهما في دقة هذه المعادلة عند استخدام طرق المعادلة القائمة على نظرية الاستجابة للمفردة.

أهمية البحث

تكمن أهمية هذا البحث فيما يلي:

١- يعد هذا البحث من البحوث القليلة التي تناولت المقارنة بين الطرق القائمة على نظرية الاستجابة للمفردة في دقة معادلتها لدرجات الاختبارات في حالة اختلاف ثبات هذه الاختبارات، مستخدمًا بيانات امبريقية وواقعية، وليست بيانات مولدة باستخدام البرامج الكمبيوترية مثل Wingen2.

٢- قد ينهي هذا البحث الخلاف بين المتخصصين في القياس النفسي في وضع شرط تساوي ثبات الاختبارات المراد معادلة درجاتها ضمن الشروط اللازمة لمعادلة الدرجات.

٣- يقدم هذا البحث الخطوات العلمية المتبعة في معادلة درجات الاختبارات باستخدام الطرق القائمة على نظرية الاستجابة للمفردة والكشف عن دقة هذه المعادلة.

٤- قد يساعد هذا البحث المؤسسات والهيئات المختلفة التي تعقد اختبارات القبول أو الكفاءة في اختيار أكثر الطرق ملاءمة لمعادلة درجات الاختبارات التي تقيس نفس السمة في حالة اختلاف ثباتها.

حدود البحث

يقتصر هذا البحث على

١- صورتي اختبار في مقرر سيكلوجية التعلم المقرر على طلاب الفرقة الثانية بكلية التربية في الفصل الأول من العام الجامعي ٢٠١٤ – ٢٠١٥، وتم اشتقاق عدة صور اختبارية من كلتا الصورتين.

٢- عينة مكونة من (٩٥١) طالبًا وطالبة من طلاب الفرقة الثانية بكلية التربية جامعة المنيا في العام الجامعي ٢٠١٤- ٢٠١٥م.

٣- تصميم معادلة درجات الاختبارات باستخدام المجموعة الواحدة.

مصطلحات البحث

- **معادلة درجات الاختبارات:** هي تلك العملية الإحصائية التي يمكن من خلالها تحويل نظام وحدات القياس التي تم الحصول عليها من صورة اختبارية أو أكثر إلى نظام وحدات قياس صورة اختبارية أخرى تقيس نفس السمة (Kilmen & Demirtasli, 2012).
- **دقة معادلة درجات الاختبارات:** هي الحصول على أقل قدر ممكن من مقدار الخطأ في معادلة درجات الاختبارات سواء أكان خطًا منتظمًا، أم عشوائيًا، وتقاس دقة المعادلة في هذا البحث من خلال قيمة جذر متوسط مربعات الخطأ.
- **ثبات الاختبار:** ويقصد به ثبات النتائج التي يقدمها الاختبار عند إعادة تطبيقه مرة أخرى على نفس العينة بعد فترة، وهذا الثبات يعني ثبات الاستقرار، أما ثبات الاتساق فيعني أن أجزاء الاختبار متسقة أي أنها تقيس نفس الشيء، ويحسب ثبات الاختبار في هذا البحث باستخدام معادلة ألفا كرونباك باستخدام برنامج SPSS.
- **تصميم المجموعة الواحدة:** وهو أحد تصميمات معادلة الاختبار يعتمد على تطبيق صورتي الاختبار المراد معادلة درجاتهما على المجموعة

نفسها من المفحوصين الواحدة تلو الأخرى وفي اليوم نفسه؛ لتفادي تأثير عوامل التعب والملل والدافعية والتعلم السابق على أداء المفحوصين.

أداة البحث

تتمثل أداة هذا البحث في صورتي اختبار لمقرر سيكلوجية التعلم على طلاب الفرقة الثانية بكلية التربية في الفصل الأول من العام الجامعي ٢٠١٤ – ٢٠١٥، وقد تم بناؤهما وفق الخطوات الآتية:

١- تحديد الغرض من صورتي الاختبار: حيث تهدف كلتا الصورتين إلى قياس تحصيل الطلاب في مقرر سيكلوجية التعلم، وذلك في أربعة مستويات معرفية، هي: المعرفة، والفهم، والتطبيق، والتحليل.

٢- تحديد محتوى صورتي الاختبار: تم تحديد موضوعين رئيسين لتمثيلهما في كل صورة من صورتي الاختبار، وهما: مفهوم التعلم وشروطه، ونظريات التعلم.

٣- تحديد الأهداف المراد قياسها: تم تحديد أهداف صورتي الاختبار بحيث تمثل أربعة مستويات معرفية، هي: المعرفة، والفهم، والتطبيق، والتحليل، وقد بلغ عدد الأهداف التي تسعى كل صورة إلى قياسها ٤٨ هدفًا بواقع (١٢) هدفًا في كل مستوى معرفي، أي بنسبة (٢٥٪) في كل مستوى.

٤- تحديد الأوزان النسبية لموضوعات محتوى صورتي الاختبار: تم ذلك بناء على زمن تدريس كل موضوع من هذه الموضوعات، وذلك بواقع محاضرتين لموضوع مفهوم التعلم وشروطه، وست محاضرات في موضوع نظريات التعلم، وبالتالي فإن الوزن النسبي لموضوع مفهوم التعلم وشروطه بلغ (٢٥٪)، والوزن النسبي لموضوع نظريات التعلم بلغ (٧٥٪).

٥- وضع جدول مواصفات صورتي الاختبار: وذلك لضمان توزيع مفردات الاختبار بما يتناسب مع الأوزان النسبية لموضوعات المحتوى ومستويات الأهداف المختلفة.

٦- تحديد عدد مفردات صورتي الاختبار: تمت صياغة (٧٥) مفردة في كل صورة من صورتي الاختبار، وكانت جميعها على هيئة اختيار من متعدد يتم فيها اختيار بديل واحد صحيح من بين أربعة بدائل، وتم توحيد جميع التعليمات في صورتي الاختبار، وكذلك ترتيب مفردات كل صورة اختبارية بنفس طريقة ترتيبها في الصورة الأخرى في ضوء الهدف المراد قياسه.

٧- التأكد من صدق صورتي الاختبار: وذلك باستخدام عدة طرق، كما يلي:

أ) عرض مفردات صورتي الاختبار على السادة المحكمين من أعضاء هيئة التدريس بقسم علم النفس التربوي ممن يقومون بتدريس المقرر، وكان عددهم ٩ محكمين؛ وذلك للوقوف على قياس المفردة للهدف المراد قياسه، وصحة الصياغة اللغوية للمفردة، ومناسبتها للطلاب، وقد أفاد جميع المحكمين بملاءمة المفردات لقياس الأهداف التي وضعت لقياسها وكذلك مناسبتها للطلاب، وقد اقترح بعضهم إعادة الصياغة اللغوية لبعض المفردات وتغيير بعض المشتتات، وقد التزم الباحث بإجراء هذه التعديلات.

ب) صدق المحك: من خلال حساب معامل الارتباط بين درجات الطلاب على كل صورة من صورتي الاختبار ودرجاتهم على اختبار سيكلوجية التعلم في نهاية الفصل الدراسي الأول (إعداد قسم علم النفس التربوي)، وكان معامل صدق المحك للصورة الأولى ٠.٨٧٣، وللثانية ٠.٩١١.

ج) كما تم التأكد من صدق صورتي الاختبار من خلال مطابقة البيانات المستمدة من كل منهما مع افتراضات النموذج ثلاثي البارامتر (كما سيلي توضيح ذلك).

٨- تم تطبيق صورتي الاختبار على طلاب الفرقة الثانية بكلية التربية في العام الجامعي ٢٠١٤-٢٠١٥م، وقد بلغ عددهم (٩٥١) طالبًا وطالبة، بحيث يقوم نصف عدد الطلاب بالإجابة عن مفردات الصورة الاختبارية الأولى، في الوقت الذي يجيب فيه النصف الباقي عن مفردات الصورة الاختبارية الثانية، ويعطى جميع الطلاب فترة راحة، ثم يعطى لكل الطلاب الصورة الاختبارية التي لم يجيبوا عنها؛ وذلك لضمان تعرض الإجابة عن مفردات الصورتين إلى العوامل نفسها التي قد تؤثر في الإجابة مثل التعب والملل والدافعية والتعلم السابق؛ وذلك للتقليل من الخطأ المنتظم لمعادلة صورتي الاختبار، وهذا ما يعرف بتصميم المجموعة الواحدة.

٩- تم التأكد من مطابقة بيانات كل من صورتي الاختبار لافتراضات نظرية الاستجابة للمفردة، وذلك وفق الخطوات الآتية:

أ) التحقق من أحادية البعد: وذلك بالاعتماد على التحليل العاملي وحساب النسبة بين الجذر الكامن للعامل الأول والجذر الكامن للعامل الثاني في كل صورة من صورتي الاختبار، وكانت النتائج كما في جدول (١):

جدول (١) قيمتا الجذر الكامن للعاملين الأول والثاني في كل صورة اختبارية والنسبة بينهما

صورة الاختبار	الجذر الكامن للعامل الأول	الجذر الكامن للعامل الثاني	النسبة بين الجذرين الكامنين
الأولى	٦.٠٦٨	٢.٧٢٣	٢.٢٢٨
الثانية	٦.٨٥٨	٢.٧٥٧	٢.٤٨٧

يتضح من هذه النتائج زيادة النسبة بين الجذر الكامن للعامل الأول والجذر الكامن للعامل الثاني عن القيمة (٢) التي حددها ريكاس (Reckase, 1979) كشرط لتحقق أحادية البعد في كل صورة من صورتي الاختبار؛ مما يؤكد تحقق افتراض أحادية البعد في كل منهما.

ب) التحقق من الاستقلال الموضعي لمفردات كلتا الصورتين: وتم ذلك من خلال حساب مؤشر Z لفيشر، حيث كان عدد أزواج المفردات المرتبطة موضعيًا في الصورة الاختبارية الأولى والثانية (٣٤، ٢٢) زوجًا على الترتيب من بين (٢٧٧٥) زوجًا من أزواج المفردات في كل صورة أي بنسبة (٠.٠١٢ ٪، ٠.٠٠٨ ٪) على الترتيب، وهذا يدل على تحقق الاستقلال الموضعي بدرجة كبيرة بين مفردات كل صورة من صورتي الاختبار.

ج) التحقق من التحرر من السرعة: وذلك بإتاحة الوقت الكافي لجميع الطلاب للإجابة عن مفردات كل صورة من صورتي الاختبار، والتأكد من أن عامل السرعة في الإجابة لم يؤثر في استجابات الطلاب على مفردات صورتي الاختبار.

د) التحقق من العلاقة الوتيرية بين احتمالية الإجابة الصحيحة عن كل مفردة من مفردات صورتي الاختبار وقدرة المفحوصين في تحصيل المقرر، وذلك من خلال تحقق شكل منحنى خصائص المفردة ICC في كل مفردة، وتم ذلك بعد تدريج المفردات.

١٠- تم تحليل مفردات صورتي الاختبار باستخدام النموذج ثلاثي البارامتر، وذلك باستخدام برنامج (PARSCALE 4.1)؛ لكونه أكثر نماذج الاستجابة للمفردة مناسبة لمفردات الاختيار من متعدد ثنائية الاستجابة (Wagner & Harvey, 2003)، وقد نتج عن ذلك عدم مطابقة (٤) مفردات من الصورة الاختبارية الأولى، وكذلك عدم مطابقة (٥) مفردات من الصورة الثانية، وكانت من بين هذه المفردات غير المطابقة مفردتان في الصورة الأولى تتفقان مع مفردتين غير مطابقتين من مفردات الصورة الثانية في قياس نفس الهدفين، فتم حذفهما، كما تم حذف المفردتين غير المطابقتين

الباقيتين في الصورة الأولى مع المفردتين اللتين تقيسان نفس الهدفين في الصورة الثانية، ومثل ذلك مع المفردات الثلاث غير المطابقة المتبقية في الصورة الثانية وما يقابلهما في الصورة الأولى في قياس نفس الأهداف؛ ليصل عدد المفردات المستبعدة من كل صورة إلى سبع مفردات؛ وبذلك وصل عدد المفردات المقبولة (٦٨) مفردة في كلتا الصورتين وفقًا للنموذج ثلاثي البارامتر؛ لتتكون بذلك صورتا الاختبار (أ، ب).

١١- تم حساب معامل ثبات صورتي الاختبار باستخدام معادلة ألفا كرونباك باستخدام برنامج SPSS، وبلغ معامل الثبات (٠.٧٦٧) للصورة الاختبارية (أ)، وبلغ (٠.٨١٩) للصورة (ب)، وهي قيم مقبولة للدلالة على ثبات صورتي الاختبار.

عينة البحث

تكونت عينة البحث من (٩٥١) طالبًا وطالبة من بين (١٠٦٤) طالبًا وطالبة في الفرقة الثانية بكلية التربية جامعة المنيا في العام الجامعي ٢٠١٤- ٢٠١٥م، وقد كانت العينة تمثل جميع الطلاب في التخصصات المختلفة بالكلية عدا الغائبين يوم تطبيق صورتي الاختبار.

إجراءات البحث

١- تم تقدير بارامترات الصعوبة والتمييز والتخمين لمفردات كل من صورتي الاختبار (أ، ب) وتقدير قدرات المفحوصين باستخدام النموذج ثلاثي البارامتر، وإجراء معادلة للدرجات باستخدام طرق: (المتوسط / المتوسط، والمتوسط / سيجما، وطريقة هاييارا، وطريقة ستوكنج لورد).

٢- تم حذف مفردتين تقيسان نفس الهدف من صورتي الاختبار (واحدة من كل صورة بما لا يخل بجدول المواصفات) بحيث يتم الحصول على معامل ثبات متساوي تقريبًا في الحالتين (٠.٨٠٤، ٠.٨٠٦) على الترتيب، وذلك باستخدام تقنية

(Alpha - Scale if item deleted) في برنامج SPSS (v.16)، وهاتان المفردتان تحملان رقم (٥٢) في صورتي الاختبار، وبذلك تكونت الصورتين (ج، د).

٣- تم إجراء معادلة لدرجات الطلاب على صورتي الاختبار (ج، د) باستخدام الطرق القائمة على نظرية الاستجابة للمفردة التي سبقت الإشارة إليها، والحكم على دقة المعادلة في كل حالة.

٤- تم إجراء نفس الإجراءين الثاني والثالث عدة مرات لتكوين صور اختبارية مختلفة في معامل الثبات بحيث يتم الحصول على فروق مختلفة بين معاملي الثبات، وبذلك تم تكوين صورتي الاختبار (هـ ، و) بحذف المفردة التي تحمل رقم (٦٧) في صورتي الاختبار، وصورتي الاختبار (ز ، ح) من خلال حذف المفردة التي تحمل رقم (٤٥) في صورتي الاختبار، وصورتي الاختبار (ط ، ي) من خلال حذف المفردتين اللتين تحملان رقمي (٢٣، ٤٥) في صورتي الاختبار.

نتائج البحث وتفسيرها

الإجابة عن السؤال الأول

ينص السؤال الأول على: "**ما درجة دقة معادلة درجات صورتي الاختبار باستخدام الطرق القائمة على نظرية الاستجابة للمفردة مع اختلاف معاملي ثباتهما؟**".

وللإجابة عن هذا السؤال تمت معادلة درجات صورتي الاختبار (أ، ب)، (ج، د)، (هـ، و)، (ز، ح)، (ط، ي)، حيث تم حساب معاملي المعادلة Equating coefficients (A ، B) باستخدام برنامج IRTEQ في كل حالة، حيث يمثل معامل المعادلة A ميل خط التحويل الخطي ، ويمثل معامل المعادلة B المقطع الصادي للتحويل الخطي، كما تم حساب قيمة

جذر متوسط مربعات الخطأ RMSE للدلالة على دقة المعادلة، وكانت النتائج كما في جدول (٢)

جدول (٢) معاملي المعادلة وقيمة جذر متوسط مربعات الخطأ في معادلة الاختبارات

طريقة المعادلة	الإحصائي	صورتا الاختبار مرتبة تصاعديًا في ضوء الفرق بين معاملي الثبات.				
		ج، د	هـ، و	أ، ب	ز، ح	ط، ي
	قيمتا الثبات	٠.٨٠٤، ٠.٨٠٦	٠.٧٨٨، ٠.٨١٤	٠.٧٦٧ ، ٠.٨١٩	٠.٧٣٩، ٠.٨١١	٠.٦٩٤، ٠.٨٠٦
	فرق الثبات	٠.٠٠٢	٠.٠٢٦	٠.٠٥٢	٠.٠٧٢	٠.١١٢
المتوسط / المتوسط	معامل المعادلة A	٠.٩٢٠	٠.٩٧٠	٠.٩٨٠	٠.٩٨٠	٠.٩٦٠
	معامل المعادلة B	- ٠.١٤٠	- ٠.٠٨٠	- ٠.٠٥٠	٠.٠٣٠	٠.١٢٠
	جذر متوسط مربعات الخطأ	٠.٠٢٤٥	٠.٠١٢٣	٠.٠٠٨٣	٠.٠٠٦٣	٠.٠١٨١
المتوسط / سيجما	معامل المعادلة A	٠.٨٧٠	٠.٩٣٠	١.٠٥٠	١.٠٥٠	١.٠٦٠
	معامل المعادلة B	- ٠.١٤٠	- ٠.٠٨٠	- ٠.٠٦٠	٠.٠٢٠	٠.١١٠
	جذر متوسط مربعات الخطأ	٠.٠٣٣٤	٠.٠٢٠١	٠.٠١٥٢	٠.٠١٣٩	٠.٠٢٢٨
هاييارا	معامل المعادلة A	٠.٩٦٠	١.٠٧٠	١.٠٨٠	١.٠٦٠	١.٠٧٠
	معامل المعادلة B	٠.٠٧٠	٠.٠٩٠	٠.٢٥٠	٠.٣١٠	٠.٣٥٠
	جذر متوسط مربعات الخطأ	٠.٠١٣٣	٠.٠٢٣٥	٠.٠٣٨١	٠.٠٤١٨	٠.٠٤٩٦
ستوكنج – لورد	معامل المعادلة A	١.٠٢٠	١.٠٤٠	١.٠٩٠	١.٠٩٠	١.١٣٠
	معامل المعادلة B	٠.٠٦٠	٠.٠٥٠	٠.٠٤٠	٠.٠٩٠	٠.١٢٠
	جذر متوسط مربعات الخطأ	٠.٠٠٨٦	٠.٠١١٨	٠.٠٢٥٠	٠.٠٢٧٢	٠.٠٤١١

يتضح من جدول (٢) أنه في جميع الحالات – سواء باختلاف طرق المعادلة أو باختلاف الفرق بين ثبات صورتي الاختبار – فإن قيمة جذر متوسط مربعات الخطأ RMSE كانت صغيرة جدا مما يدل على دقة معادلة درجات صورتي الاختبار المختلفتين في الثبات باستخدام أية طريقة من طرق نظرية الاستجابة للمفردة، وهذا يتفق مع ما أشار إليه دورانز وهولاند (Dorans & Holland, 2000) بأن متطلب تساوي الاختبارين المراد معادلة درجاتهما في الثبات غالبًا ما ينتهك، ورغم ذلك فإن عملية

المعادلة في هذه الحالة تكون مرضية، وأنه من الناحية العملية لا يوجد دليل على وجود مشكلات في معادلة درجات الاختبارات التي لها معاملات ثبات مرتفعة ومختلفة اختلافًا واضحًا في الثبات، وبالتالي فإن متطلب تساوي الثبات يعد متطلبًا ثانويًا وليس أساسيًا، ويمكن تفسير ذلك بأنه ما دامت الاختبارات المراد معادلة درجاتها تتمتع بدرجة ثبات مقبولة فإن معادلة هذه الاختبارات ستكون مرضية بغض النظر عن الاختلاف في قيمة ثبات هذه الاختبارات، وقد يرجع ذلك إلى تحقق الافتراضات الأساسية لنظرية الاستجابة للمفردة بشكل عام، وهي: أحادية البعد والاستقلال الموضعي والتحرر من السرعة والعلاقة الوتيرية، كما قد يكون السبب وراء دقة طرق معادلة درجات صورتي الاختبار في جميع الحالات مع الاختلاف بين معاملي ثبات صورتي الاختبار إلى اختيار النموذج ثلاثي البارامتر في تدريج مفردات صورتي الاختبار، وهو أكثر نماذج الاستجابة للمفردة ملاءمة لمفردات الاختيار من متعدد ثنائية الاستجابة؛ وذلك لمعالجته أثر التخمين عند تقدير قدرات المفحوصين، وقد تعود دقة المعادلة أيضًا إلى اعتماد تصميم المعادلة القائم على مجموعة واحدة، والذي يعطي أقل خطأ ممكن في المعادلة.

الإجابة عن السؤال الثاني

ينص السؤال الثاني على: " **هل يؤثر اختلاف معاملي ثبات صورتي الاختبار في دقة معادلة درجاتهما باستخدام الطرق القائمة على نظرية الاستجابة للمفردة؟**"

وبالنظر إلى جدول (٢) يتضح أن جذر متوسط مربعات الخطأ تأثر بشكل غير منتظم مع زيادة الفرق بين ثبات صورتي الاختبار في حالة استخدام طريقة (المتوسط / المتوسط) وطريقة (المتوسط / سيجما)، أي أن دقة معادلة صورتي الاختبار تتأثر بشكل غير منتظم مع زيادة الفرق بين

ثبات صورتي الاختبار، وهذا يعني أنه لا توجد قاعدة يمكن الاستناد إليها في الحكم على العلاقة بين دقة معادلة الاختبارات ودرجة الفرق بين ثبات صورتي الاختبار في حالة استخدام طريقة (المتوسط / المتوسط) وطريقة (المتوسط / سيجما).

أما في حالة استخدام طريقتي هايبارا وستوكنج لورد في معادلة درجات صورتي الاختبار فإن قيمة جذر متوسط مربعات الخطأ ازدادت تدريجيًا مع زيادة الفرق بين ثبات صورتي الاختبار؛ مما يدل على انخفاض دقة المعادلة مع زيادة هذا الفرق، وهذا يدل على أهمية تقارب قيمتي ثبات صورتي الاختبار في حالة استخدام طريقتي هايبارا وستوكنج لورد للحصول على درجة عالية من الدقة في معادلة الاختبارات، وهذا يعني تأثر دقة معادلة درجات الاختبارات باختلاف ثبات صورتي الاختبار عند استخدام هاتين الطريقتين.

وبهذا، فإن النتائج السابقة تعني أن معادلة الاختبارات تتم بشكل مقبول في حالة اختلاف صورتي الاختبار المراد معادلة درجاتهما في درجة الثبات، إلا أن دقة هذه المعادلة تقل بزيادة الفرق بين معاملي الثبات في حالة استخدام طريقتي هايبارا وستوكنج لورد، وبالتالي يمكن استنتاج أنه ما دامت الاختبارات المراد معادلة درجاتها تتمتع بدرجة مقبولة من الثبات وتتحقق فيها افتراضات نظرية الاستجابة للمفردة فإن هذه المعادلة تتم بدقة مقبولة، إلا أن هذه الدقة ترتفع كلما تقاربت قيمتا ثبات صورتي الاختبار في حالة استخدام طريقتي هايبارا وستوكنج لورد؛ ولعل هذا يرجع إلى أن هاتين الطريقتين تعتمدان بالدرجة الأولى على تقدير الفرق في منحنيات خصائص المفردات ICCs سواء عن طريق حساب مجموع مربعات الفروق بين منحنيات خصائص المفردات لكل مفردة، أو عن طريق حساب مربع مجموع الفروق بينها، وهاتان الطريقتان تختلفان في ذلك عن طريقتي المتوسط/ المتوسط، والمتوسط/ سيجما.

أوجه الاستفادة من البحث

يمكن الاستفادة من نتائج هذا البحث فيما يلي:

١- ضرورة التحقق من افتراضات نظرية الاستجابة للمفردة في البيانات المستمدة من تطبيق كل صورة من الصور الاختبارية عند استخدام طرق معادلة درجات الاختبارات القائمة عليها.

٢- ضرورة التحقق من تقارب معاملي ثبات صورتي الاختبار المراد معادلة درجاتهما عند استخدام طريقتي هاييارا وستوكنج لورد، والتحقق من تمتع صورتي الاختبار بدرجة مقبولة من الثبات عند استخدام أية طريقة من الطرق القائمة على نظرية الاستجابة للمفردة.

٣- توجيه أنظار القائمين على وضع اختبارات القبول أو اختبارات الكفاءة في أي تخصص من التخصصات إلى ضرورة التحقق من دقة معادلة درجات الاختبارات عند استخدام عدة صور اختبارية.

البحوث المقترحة

بناء على البحث الحالي يقترح الباحث إجراء البحوث الآتية:

١- أثر اختلاف ثبات صورتي الاختبار على دقة معادلة درجات الاختبارات باستخدام الطرق القائمة على نظرية الاستجابة للمفردة عند استخدام تصميم الجذع المشترك أو الأفراد المشتركين أوو المجموعات المتكافئة.

٢- أثر انتهاك افتراض الاستقلال الموضعي لمفردات صورتي الاختبار على دقة معادلة درجاتهما باستخدام الطرق القائمة على نظرية الاستجابة للمفردة.

٣- أثر اختلاف ثبات صورتي الاختبار على دقة معادلة درجات الاختبارات متعددة الأبعاد باستخدام الطرق القائمة على نظرية الاستجابة للمفردة.

المراجع والمصادر

١. أبو مسلم، مايسة فاضل (٢٠١٠). معادلة صيغتى اختبار توني للذكاء غير اللفظي باستخدام طرق مختلفة للمعادلة في ضوء بعض المتغيرات المؤثرة على نتائجها. **مجلة الجمعية المصرية للدراسات النفسية**، ٦٦، ٣٧١-٤١١.

٢. الحواري، أروى عيسى (٢٠٠٧). الخصائص السيكومترية لصور مختارة من اختبارات الرخصة الدولية لقيادة الحاسوب في الأردن ومعادلة درجاتها. **رسالة ماجستير غير منشورة**، كلية التربية بجامعة اليرموك.

٣. الدوسري، راشد حماد (٢٠٠١). معادلة الاختبارات: مفهومها، وطرقها، ومشكلات تطبيقها. **مجلة العلوم التربوية والنفسية بالبحرين**، ٤(٢)، ١٠٧-١٤١.

٤. الرحيل، راتب صايل الخضر (٢٠١٣). أثر وجود أداء تفاضلي في الفقرات المرساوية على دقة المعادلة العمودية لاختبار اوتيس لينون للقدرة العقلية. **المجلة الدولية التربوية المتخصصة**، ٢(٨)، ٧٥٤ – ٧٧١.

٥. الشريفين، نضال كمال (٢٠٠٧). معادلة درجات نماذج مختلفة من اختبار الكفاءة اللغوية في اللغة الإنجليزية لدى طلبة جامعة اليرموك. **مجلة جامعة أم القرى للعلوم التربوية والنفسية**، ١(٢)، ١١-٦٢.

٦. المحروق، يوسف عبد العاطي (٢٠١١). مقارنة طرق كيرنيل والمئينيات وطرق نظرية استجابة الفقرة عند استخدام تصميم الفقرات المشتركة في دقة معادلة درجات الاختبارات متعددة الحدود. **رسالة دكتوراه** غير منشورة، كلية الدراسات العليا، الجامعة الأردنية.

٧. المدانات، رائد فايز (٢٠١٢). مقارنة فاعلية طريقتي معادلة العلامات الحقيقية والمشاهدة في معادلة الاختبارات باستخدام جذع مشترك ومجموعات غير متكافئة. **مجلة العلوم التربوية والنفسية بالبحرين**، ١٣(٢)، ٣٦٥- ٣٩٤.

٨. بالخيور، شفاء عبد الله عبد القادر (٢٠٠٩). فاعلية طرق معادلة نماذج اختبار القدرات العامة بالمركز الوطني للقياس والتقويم وفق نظريتي القياس التقليدية والحديثة في ضوء بعض المتغيرات. **رسالة دكتوراه** غير منشورة، كلية التربية، جامعة أم القرى.

٩. طيفور، مصطفى أحمد (٢٠٠٧). دراسة مقارنة لنماذج نظرية الاستجابة للمفردة في معادلة درجات الاختبارات. **رسالة دكتوراه** غير منشورة، معهد الدراسات التربوية، جامعة القاهرة.

١٠. علام، صلاح الدين محمود (٢٠٠٥). نماذج الاستجابة للمفردة الاختبارية أحادية البعد ومتعددة الأبعاد وتطبيقاتها في القياس النفسي والتربوي. القاهرة، **دار الفكر العربي**.

11.Cook, L. L.; Eignor, D. R. & Schmitt, A. P. (1990). Equating Achievement Tests Using Samples Matched on Ability. New York: College Entrance Examination Board.

12.Crocker, L. & Algina, J. (1986).Introduction to Classical and Modern Test Theory. New York: Holt, Rinehart & Winston.

13.MacCann, R. (1989). A Comparison of Two Observed-Score Equating Methods That Assume Equally Reliable, Congeneric Tests. Applied Psychological Measurement, 13(3), 263-276.

14. Dorans, N. J. & Holland, P. W. (2000). Population Invariance and the Equatability of Tests: Basic Theory and the Linear Case. Princeton: Educational Testing Service.
15. Dorans, N. J.; Moses, T. P. & Eignor, D. R. (2011). Equating Test Scores: Toward Best Practices. In A.A. von Davier (ed.), Statistical Models for Test Equating, Scaling, and Linking: Statistics for Social and Behavioral Sciences, (pp.23-42). New York: Springer.
16. González, J., Barrientos, A. F., & Quintana, F. A. (2015). Bayesian nonparametric estimation of test equating functions with covariates. Computational Statistics & Data Analysis, 89(0), 222-244.
17. Hambleton, R. K. & Jones, R. W. (1993). Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. Educational Measurement: Issues and Practice, 12(3), 38- 47.
18. Hambleton, R.; Swaminathan, H. &Rogers, H.(1991). Fundamentals of Item Response Theory, International Educational and Professional. Publisher Newbury Park.
19. Holland, P. W., & Dorans, N. J. (2006). Linking and Equating. In R. L. Brennan (Ed.),Educational measurement (4th ed., pp. 187–220). Westport, CT: American Council on Education and Praeger.
20. Kilmen, S. & Demirtasli, N. (2012). Comparison of Test Equating Methods Based on Item Response Theory According to the Sample Size and Ability Distribution. Procedia - Social and Behavioral Sciences, 46, 130 – 134.

21. Cook, L.L.,& Eignor ,D.R.(1991). An NCMF Instructional Module on IRT Equating Methods. Educational Measurement: Issues and Practice,10,37-45.
22. Kim, J. Y. (2013). Factors Affecting Accuracy of Comparable Scores for Augmented Tests under Common Core State Standards. Unpublished doctoral dissertation. University of Lowa.
23. Kolen, M. J. (1988). Traditional Equating Methodology. Educational Measurement: Issues and Practice, 7(4), 29-36.
24. Kolen, M. J. & Brennan, R. L. (1995). Test equating: Methods and practices. (1st ed.). New York: Springer.
25. Lamprianou, I. (2007). An Investigation into the Test Equating Methods Used During 2006, and the Potential for Strengthening Their Validity and Reliability. University of Manchester and Cyprus Testing Service: the Qualifications and Curriculum Authority.
26. Lord, F.M. (1980). Application of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Lawrence Erlbaum Associates.
27. Lui, J.; Zu, J.; Curley, E. & Carey, J. (2014). Test Score Equating Using Discrete Anchor Items Versus Passage-Based Anchor Items: A Case Study Using SAT® Data. ETS Research Report No. RR-14-14. © 2014 Educational Testing Service.
28. Marks, E. & Lindsay, C. A. (1972), Some Results Relating to Test Equating under Relaxed Test Form Equivalence. **Journal** of Educational Measurement, 9, 45–56.

29. Meng, Y. (2012). Comparison of Kernel Equating and Item Response Theory Equating Methods. Unpublished doctoral dissertation. University of Massachusetts.
30. Moghadamzadeh, A., Salehi, K. & Khodaie, E. (2011). A comparison Method of Equating Classic and Item Response Theory (IRT): A Case of Iranian Study in the University Entrance Exam. Procedia - Social and Behavioral Sciences, 29, 1368-1372.
31. Pang, X., Madera, E., Radwan, N. & Zhang, S. (2010). A Comparison of Four Test Equating Methods. Education Quality and Accountability Office. Ontario: Queen's Printer for Ontario. www.eqao.com.
32. Petersen, N. S.; Kolen, M. J. & Hoover, H. D. (1989). Scaling, Norming, and Equating. In R. L. Linn (Ed.), Educational Measurement (3rd ed.) (pp. 221-262). New York: Macmillan.
33. Reckase, M. D. (1979). Unifactor Latent Trait Models Applied to Multifactor Tests: Results and Implications. **Journal** of Educational Statistics, 4, 207-230.
34. Rui, W.; Shu-Liang, D. & Deng-Wen, G. (2010). Test Equating with Testlets. Acta Psychologica Sinica, 42 (3), 434-442.
35. Ryan, J. & Brockmann, F. (2011). A Practitioner's Introduction to Equating with Primers on Classical Test Theory and Item Response Theory. Technical Issues in Large Scale Assessment, the Council of Chief State School Officers.
36. Skages, G & Lissitz, R. (1986). An Exploration of The Robustness of Four Test Equating Models. Applied Psychological Measurement, 10(3), 303-317

37. Song, T. (2009). Investigating Different Item Response Models in Equating the Examination for the Certificate of Proficiency in English (ECPE). Spaan Fellow Working Papers in Second or Foreign Language Assessment, 7, 85–98, www.lsa.umich.edu/eli/research/spaan.
38. von Davier, A. A.(2011). A Statistical Perspective on Equating Test Scores. In A.A. von Davier (ed.), Statistical Models for Test Equating, Scaling, and Linking: Statistics for Social and Behavioral Sciences, (pp.1-20). New York: Springer.
39. von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). The kernel Method of Test Equating. New York: Springer-Verlag.
40. Wagner, T. A. & Harvey, R. J. (2003). Developing a New Critical Thinking Test Using Item Response Theory. Paper presented at the 2003 annual conference of the Society for Industrial and Organizational Psychology. Retrieved June 11, 2008, from http://harvey.psyc.vt.edu/Documents/WagnerHarveySIOP2003.pdf
41. Wolf, R. (2013). Assessing the Impact of Characteristics of the Test, Common-items, and Examinees on the Preservation of Equity Properties in Mixed-format Test Equating. Unpublished doctoral dissertation. University of Pittsburgh.
42. Xu, Y.(2009). Measuring Change in Jurisdiction Achievement over Time: Equating Issues in Current International Assessment Programs. Unpublished doctoral dissertation. University of Toronto.

الملخص

هدف هذا البحث إلى تعرف أثر اختلاف معامل الثبات بين صورتي اختبار في مقرر سيكلوجية التعلم في دقة معادلة درجاتهما باستخدام الطرق القائمة على نظرية الاستجابة للمفردة، وهي: طريقة المتوسط / المتوسط، وطريقة المتوسط / سيجما، وطريقة هايبارا، وطريقة ستوكنج لورد، وفي سبيل ذلك تم بناء صورتي الاختبار والتحقق من صلاحيتهما وتحقيقهما لافتراضات نظرية الاستجابة للمفردة، ثم تم اشتقاق عدة صور اختبارية من كل منهما بهدف الحصول في كل مرة على فرق مختلف بين ثبات صورتي الاختبار، مع إجراء المعادلة في كل مرة باستخدام الطرق القائمة على نظرية الاستجابة للمفردة، وكذلك حساب قيمة جذر متوسط مربعات الخطأ في كل مرة للحكم على دقة عملية المعادلة، وقد أسفرت النتائج عن تمتع معادلة الاختبارات بدرجة مقبولة من الدقة في جميع الحالات رغم زيادة الفرق في الثبات، ولكن تأثرت درجة دقة هذه المعادلة بشكل غير منتظم عند استخدام طريقة المتوسط / المتوسط وطريقة المتوسط / سيجما مع زيادة الفرق بين ثبات صورتي الاختبار، بينما أكدت النتائج انخفاض دقة معادلة الاختبارات عند استخدام طريقتي ستوكنج لورد وهايبارا مع زيادة الفرق بين ثبات صورتي الاختبار.

الكلمات المفتاحية: معادلة درجات الاختبارات، ثبات الصور الاختبارية، افتراضات نظرية الاستجابة للمفردة، طريقة المتوسط / المتوسط، طريقة المتوسط / سيجما، طريقة هايبارا، طريقة ستوكنج لورد.

The Effect of the Difference between Test Forms Reliability in the Accuracy of Test Scores Equating Using Item Response Theory Methods

Abstract

The present research was conducted to identify the effect of the difference between reliability coefficients of two test forms in Learning Psychology course upon test scores equating accuracy using Item Response Theory (IRT) methods such as Mean/Mean, Mean/Sigma, Haebara & Stocking-Lord methods. To achieve this, two test forms were prepared and validated. In addition, IRT assumptions were verified. Also, multiple test forms were built from the two forms to obtain different varied differences in test reliability each time and to equate their scores using IRT methods. Root Mean Squares Error (RMSE) was calculated to measure the accuracy of equating. The findings indicated that in all cases the accuracy of equating was obtained regardless of the difference between test forms reliability. The accuracy of test scores equating was affected indiscriminately with the increased difference in test forms reliability when Mean/Mean and Mean/Sigma methods were used, while equating accuracy was decreased with the increased difference in test forms reliability when Haebara and Stocking-Lord methods were used.

Key words: Test Scores Equating, Test Forms Reliability, Item Response Theory Assumptions, Mean/Mean, Mean/Sigma, Haebara & Stocking-Lord methods.