معادلة الاختبارات
مفهومها، وطرقها، ومشكلات تطبيقها

د. راشـــد حمّـــاد الدوســري
رئيس قسم الأصول والإدارة التربوية
كليــة التربيــة – جامعــة البحــرين

المجلد 2 العدد 4 ديسمبر 2001

معادلة الاختبارات مفهومها، وطرقها، ومشكلات تطبيقها

د. راشد حمّاد الدوسريّ
رئيس قسم الأصول والإدارة التربوية
كلية التربية–جامعة البحرين

الملخص

هذه ورقة بحثية نظرية تتعلق بالقضايا الفنية والعملية في معادلة الاختبار. يعرض الباحث في هذه الورقة موضوع معادلة الاختبار بالتفصيل. تنقسم الورقة إلى ستة أجزاء كبيرة. يتعلق الجزء الأول بتعريف معادلة الاختبار وأغراضها، والشروط الضرورية لتطبيقها. ويتناول الجزء الثاني أهم تصميمات جمع البيانات التي تستخدم في معادلة الاختبار، مثل تصميم المجموعة الواحدة، وتصميم المجموعات المتكافئة، وتصميم المجموعات العشوائية المرتبطة باختبار مشترك، وتصميم المجموعات غير المتكافئة المرتبطة باختبار مشترك. أما الجزء الثالث فهو مخصص للطرق المتعددة المستخدمة في معادلة الاختبار، والتي يعتمد بعضها على النظرية التقليدية في الاختبار ويعتمد البعض الآخر على النظرية الحديثة في الاختبار (نظرية الاستجابة لمفردة الاختبار). الطرق التقليدية في معادلة الاختبار تشمل طريقة المئينيات المتساوية، والطريقة الخطية. أما طرق معادلة الاختبار المعتمدة على النظرية الحديثة في الاختبار فتشمل طريقة درجة القدرة، وطريقة الدرجة الحقيقية، وطريقة الدرجة المشاهدة (الملاحظة). وتعتمد طرق معادلة الاختبار في النظرية الحديثة في الاختبار على النماذج اللوجستية المرتبطة بهذه النظرية. يلخص الجزء الرابع عدداً من الدراسات التطبيقية حول معادلة الاختبار للتعرف على أفضل طريقة منها، حيث إن معيار الأفضلية للطريقة التي تنتج أقل خطأ في معادلة الاختبار. أما الجزء الخامس فيناقش بعض القضايا الفنية والعملية المرتبطة بمعادلة الاختبار، كمعادلة الاختبارات غير المتوازنة، والحصول على درجات متكافئة من الاختبار، وأثر العينات الصغير على معادلة الاختبار، وأهمية التواصل بين المتخصصين في القياس النفسي، لتجنب الفهم الخاطئ لنتائج معادلة الاختبار. أما الجزء الأخير فيستخلص بعض النتائج والمضامين للبحث التربوي حول موضوع معادلة الاختبار، مثل حجم العينة، والطرق المستخدمة لمعادلة الاختبار، وأنواع الاختبارات المستخدمة لغرض المعادلة، ومشكلة وجود أكثر من بعد (سِمَة) في معادلة الاختبار، ومعادلة الاختبارات المعتمدة على الوسائط المتعددة.

* تاريخ تسلم البحث ٢٠٠١/٢/٢٠م
* تاريخ قبوله للنشر ٢٠٠١/١١/٥م

Technical and Practical Issues on Test Equating

Dr. Rashid Hammad Al-Dosary
College of Education
University of Bahrain

Abstract

This is a theoretical research paper about the technical and practical issues on test equating. In this paper the researcher addresses the issue of test equating in detail. The paper is divided into six major parts. The first part deals with the definition of test equating, purposes of test equating, and the necessary conditions for equating. The second part tackles the main data collection designs that are commonly used for test equating, such as single group design, counterbalanced random groups design, equivalent groups design, anchor test random groups design, and anchor test nonequivalent groups design. The third part is devoted to various methods of test equating. Some methods are based on classical test theory (CTT), whereas others are based on item response theory (IRT). The classical methods of test equating include equipercentile equating and linear equating. The IRT methods include ability- score equating, true-score equating, and observed-score equating; utilizing the three IRT logistics models. The fourth part summarizes a number of research studies on test equating to detect the dominant equating method(s), and what method might be preferred, where the main criterion is the least equating error to judge the quality of any method. The fifth part is intended to discuss some technical and practical issues on test equating, such as equating non-parallel test, obtaining equivalent scores on tests, effects of small sample size on test equating, and the importance of communication among psychometricians to avoid misinterpretations of equating results. The last part is concerned with some conclusions and implications for educational research regarding the issue of test equating, such as sample size, methods of equating used, types of tests used for equating, the problem of multidimensionality, and equating multimedia-based tests.

المجلد 2 العدد 4 ديسمبر 2001

معادلة الاختبارات
مفهومها، وطرقها، ومشكلات تطبيقيها

د. راشد حمّاد الدوسريّ
رئيس قسم الأصول والإدارة التربوية
كلية التربية-جامعة البحرين

مقدمــة

على الرغم من أن موضوع معادلة الاختبارات قد بحثه المختصصون في القياس والتقويم التربوي والنفسي بشكلٍ مستفيض بوصفه جزءاً من مهماتهم البحثية والمهنية؛ إلا أن بعض المتخصصين في القياس التربوي بشكلٍ عام قد تجاهلوا هذا الموضوع المهم. فخلال العشرين سنة المنصرمة بدأ الاعتراف بأهمية معادلة الاختبارات من قبل قطاع كبير من المتخصصين المسئولين عن بناء الاختبارات في الكثير من المؤسسات التربوية، مثل: المجلس القومي للقياس في التربية، والرابطة الأمريكية للبحث التربوي، وجمعية علم النفس الأمريكية (Kolen & Brennan ، ١٩٩٥. ويمكن أن يعزى هذا الاهتمام المتزايد بموضوع معادلة الاختبارات إلى ثلاثة تطورات مهمة خلال الخمس عشرة سنة الماضية. التطور الأول: يتعلق بزيادة عدد برامج الاختبارات، وتنوعها التي تستخدم صوراً متعددة من الاختبارات. بالإضافة إلى أن المتخصصين في بناء الاختبارات والمسئولين عن هذه البرامج قد أدركوا ضرورة معادلة درجات الاختبارات المستقاة من استمارات متعددة للاختبار الواحد، أو لاختبارات متعددة تقيس السَّمة نفسها. التطور الثاني: هو أن مطوري الاختبارات، ودور النشر المتخصصة بنشر الاختبارات التربوية، والنفسية قد أكدوا أهمية معادلة الاختبارات ودورها عند تحليل درجاتها، وذلك لمعالجة الكثير من القضايا الفنية التي يثيرها النقاد في هذا المجال. أما التطور الثالث: فيتعلق بقضية المساءلة والمحاسبة في التربية، وقضية قدرة الاختبارات على العدل في قياس السَّمة المراد قياسها لمجموعات من الطلبة متباينة في الجنس، والأصل العرقي. كل هذه التطورات مجتمعة قد أعطت عملية معادلة الاختبارات أهمية كبيرة في أوساط المتخصصين في القياس التربوي بشكلٍ عام، وكذلك في أوساط مستخدمي نتائج الاختبارات (Kolen & Brennan, 1995).

مشكلة الدراسة

هناك الكثير من المشكلات التي تعانيها نظم التعليم في الكثير من دول العالم، ومنها دولة البحرين، ومن أهم هذه المشكلات عملية بناء الامتحانات الفصلية، والامتحانات الملحقة للشهادات العامة بالمرحلتين الإعدادية، والثانوية؛ إذ يقوم المتخصصون في إدارة المناهج ببناء الكثير من الامتحانات لنهاية كل فصل دراسي، ويتطلب ذلك وقتاً كبيراً، وجهداً أكبر، ويضطرون إلى بناء اختبارات ملحقة للطلبة الذين لم يتمكنوا من اجتياز الامتحانات الفصلية بنجاح. ولا يوجد دليل علمي على قدرة هذه الامتحانات على تمثيل المحتوى المطلوب لتأكيد جانب الصدق فيها. كما لا يمكن اعتبارها متكافئة فيما بينها بسبب عدم تحليلها للحصول على الخصائص السيكومترية الضرورية لعملية التكافؤ، والمعادلة؛ كالوسط الحسابي، والانحراف المعياري، ومعامل الصعوبة، ومعامل التمييز. ومن الصعوبة بمكان السماح للطلبة، الذين لم يتجاوزوا بعض المقررات بنجاح، بتقديم الامتحانات في أي وقت يشاؤون، بسبب عدم وجود مصرف للامتحانات يضم جميع الامتحانات العامة بعد معادلتها علمياً وفنياً. وكثيراً ما يشتكي الطلبة في كل عام دراسي من تفاوت الامتحانات من امتحان إلى آخر، وعدم العدل في قياس سمة التحصيل قياساً علمياً سليماً. ومما يزيد من خطورة المشكلة أن نتائج الطلبة في الامتحانات العامة تترتب عليها قرارات متعلقة بالانتقال إلى مراحل دراسية أعلى، أو الالتحاق بالجامعات، أو بسوق العمل. ومن الجدير بالذكر أن هناك بعض القضايا التي تعرضت لها الأدبيات، والدراسات التي عالجت موضوع معادلة الاختبارات ومشكلاتها، ولكنها لم تعالجها بعمق، ومن مختلف الزوايا والظروف المحيطة بتطبيق الاختبارات، كتلك التي أشرنا إليها أعلاه، ولها خصوصية من بلد إلى آخر، على الرغم من أن المشكلات العامة هي مشكلات مشتركة.

ومن هنا، فإنه من الضروري أن تتصدى وزارة التربية والتعليم في كل دولة عربية لهذه المشكلة عن طريق تحليل الامتحانات العامة، ومعادلتها، ووضعها في مصارف الامتحانات. وتأتي هذه الدراسة لتسليط الضوء على أهمية موضوع معادلة الاختبارات علمياً وفنيا، وأثره على النظام التربوي والسياسات التربوية المترتبة على استخدام نتائج الامتحانات العامة، وارتباط ذلك بمفهوم صدق العواقب Consequential Validity.

الهدف من الدراسة

تهدف هذه الدراسة إلى التعرف على مفهوم معادلة الاختبارات علمياً وفنياً، والطرق المستخدمة في معادلة الاختبارات ضمن النظرية التقليدية، والنظرية الحديثة في القياس والتقويم التربوي والنفسي. وتسليط الضوء على بعض المشكلات الفنية، والتطبيقية في عملية معادلة الاختبارات. كما توضح الدراسة أهمية موضوع معادلة الاختبارات في برامج الاختبارات في القطاع التربوي، وقدرتها على حل الكثير من المشكلات التي يعانيها المتخصصون في بناء الاختبارات في المدارس، وغيرها.

أسئلة الدراسة

تحاول هذه الدراسة الإجابة عن الأسئلة الآتية.

١. ما المفهوم العلمي لمعادلة الاختبار؟

٢. ما الطرق المستخدمة في معادلة الاختبار في النظرية التقليدية في القياس والتقويم التربوي والنفسي؟

٣. ما الطرق المستخدمة في معادلة الاختبار في النظرية الحديثة في القياس والتقويم التربوي والنفسي؟

٤. ما المشكلات التطبيقية، والفنية التي تواجه المتخصصين في عملية معادلة الاختبار؟

ستتم الإجابة عن هذه الأسئلة ضمن العرض، والشرح، والتحليل في متن الدراسة، بوصفها أسئلة مترابطة فيما بينها.

أهمية الدراسة

تستمد هذه الدراسة أهميتها من الآتي:

١. أنها أول دراسة (حسب علم الباحث) في الأدبيات التربوية المكتوبة بالعربية، تتطرق إلى موضوع مهم جداً، ومعقد من الناحيتين الفنية، والتطبيقية.

٢. أنها توجه نظر المسئولين في وزارات التربية والتعليم، والجامعات في العالم العربي إلى الاهتمام بهذا الموضوع المهم الذي يحل الكثير من المشكلات التي تعانيها نظم الامتحانات في الدول العربية، وضرورة تبني عملية معادلة الاختبارات، واستثمار

تكنولوجيا المعلومات في عملية بناء الامتحانات الرسمية، والعادية، وكذلك إعادة النظر في نظم الامتحانات الحالية، وعلاقتها بتكنولوجيا المعلومات.

٣. أنها تسلط الضوء على الكثير من المشكلات التي تعانيها برامج ونظم الامتحانات، مثل: مشكلة هدر الوقت في بناء اختبارات جديدة كل مرة لنفس المقرر، وما يرتبط بذلك (خاصة في العالم العربي) من بناء إجابات نموذجية لفقرات الاختبارات التي لا حصر لها، والجهد والوقت الذي يبذله المتخصصون في بناء الامتحانات لكل فصل دراسي، وعدم توافر البرامج التي تقوم ببناء مصارف الامتحانات

مفهوم معادلة الاختبارات، وشروطها

يقول سون (Suen, 1990): إنه عندما يكون لدينا أكثر من صورة واحدة من الاختبار نفسه تكون معادلة الاختبار قضية مهمة للحصول على تفسير ملائم، وصحيح لدرجات الاختبار. أما دورانز(Dorans, 1990, P.3) فيُعرِّف معادلة الاختبار؛ بأنها عملية إجراء تعديل إحصائي على درجات صورة واحدة من الاختبار لجعل تلك الدرجات مكافئة بطريقةٍ ما لدرجات صورة أخرى من الاختبار نفسه. كما يمكن تعريف معادلة الاختبارات على أنها إجراء لإزالة الآثار التي يتركها الفرق بين متوسط مستويات الصعوبة لفقرات الاختبار، وكذلك الفروق في متوسط مستويات التمييز لفقرات الاختبار، على درجات الاختبار؛ وذلك بين صورتين من الاختبار نفسه لجعلهما متكافئتين (Brennan, 1995 & Kolen) أما كروكر، وألجاينا (Crocker & Algina, 1986) ، وكذلك هلز، وصبحيه، وهيرش (Hills, Subhyiah, & Hirsch, 1988) فيعرِّفون معادلة الاختبار بأنها عملية الحصول على درجات متكافئة لأداتين تقيسان السِّمة نفسها. ويُعرِف هلز، وزملاؤه مفهوم الدرجات المتكافئة بأنها تلك الدرجات التي تقيس السِّمة نفسها وبمقدارٍ متساوٍ من الثبات، مع تساوي الرُّتب المئينية المناظرة للدرجات. ومن جانبٍ آخر، يجب أن تقيس جميع صور الاختبار السِّمة نفسها لضرورة تحويل الدرجات رياضياً من صورة إلى أخرى . (Hambleton & Swaminathan, 1985)

يواجه المعنيون بعملية معادلة الاختبارات موقفين مختلفين في معادلة الاختبارات. الموقف الأول: هو عندما يكون لدينا اختباران متكافئان في مستوى الصعوبة، وتوزيعات السمة المراد قياسها متشابهة بين المفحوصين الذين طُبق عليهم الاختباران. والموقف الثاني:

هو عندما يكون لدينا اختباران متفاوتان في مستوى الصعوبة، وتوزيعات السِّمة المراد قياسها متباينة بين المفحوصين الذين طُبق عليهم الاختباران. معادلة الاختبار مرتبطة في الواقع بهذين الموقفين. الموقف الأول مرتبط بما يُعرف بالمعادلة الأفقية للاختبار equating Horizontal، حيث يتم إعداد صور متكافئة من الاختبار نفسه ذات فقرات مختلفة، ومتساوية في مستوى الصعوبة. أما الموقف الثاني فيرتبط بما يُعرف بالمعادلة الرأسية للاختبار Vertical equating، حيث يكون لدينا اختباران غير متساويين في مستويات الصعوبة، مع اختلاف المفحوصين في توزيعات السِّمة المراد قياسها. وينطبق ذلك على بطاريات الاختبارات، وكذلك الصفوف المختلفة في مستواها التعليمي (الرابع، الخامس، ... إلخ)، لذلك فإن معادلة الاختبار أفقياً تكون ملائمة عندما تكون لدينا صور متعددة من الاختبار نفسه، ومتوازية (رياضياً وإحصائياً) Parallel، ولكنها ليست متطابقة في الفقرات. أما معادلة الاختبار رأسياً فتستخدم عندما نريد بناء تدريج Scale واحد يسمح بالمقارنة بين مستويات المفحوصين في السِّمة المراد قياسها في صفوف دراسية مختلفة؛ حيث تكون الفقرات مختلفة في مستويات صعوبتها من صف إلى آخر، على الرغم من أنها تقيس السِّمة نفسها. بالإضافة إلى ذلك، يجب الاحتفاظ هنا بدرجة عالية من الثبات في عملية معادلة الاختبار رأسياً؛ وخاصة عندما نجد أن بعض الطلبة في الصفوف العليا، والذين لديهم قدرات منخفضة قد يحصلون على درجات غير ثابتة عندما تطبق عليهم صورة من الاختبار نفسه مخصصة لصف أدنى من صفهم (Suen, 1990; Crocker & Algina, 1986) إن أحد الحلول لمشكلة معادلة الاختبارات رأسياً هو وضع الفقرات على تدريج مشترك Common scale دون الإشارة إلى مجموعة الطلبة الذين يطبق عليهم الاختبار؛ وذلك عند بناء مصرف للأسئلة (Hambleton & Swaminathan, 1985).

شروط معادلة الاختبارات

اتفق المتخصصون على مجموعة من الشروط الضرورية لإجراء عملية معادلة الاختبارات، بشرط استيفاء جميع هذه الشروط، وعدم الإخلال بأي منها؛ لتكون عملية معادلة الاختبار صحيحة علمياً وفنياً (Hambleton & Swaminathan, 1985; Kolen & Brennan, 1995; Peterson, Kolen, & Hoover, 1989) ويمكن إيجاز هذه الشروط دون التعرض للمعادلات الرياضية المعقدة المرتبطة بها، والتي يمكن الرجوع إليها في تلك المصادر؛ وذلك على النحو التالي:

١. **قياس السِّمة نفسها (القدرة)**: يجب أن تقيس الاختبارات المراد معادلتها الخاصية نفسها ، أو السِّمة الكامنة نفسها ، أو المهارات نفسها.

٢. **العدالة (المساواة) Equity**: وهو أن يكون التوزيع التكراري المشروط للدرجات عند مستوى معين من مستويات القدرة (Θ) للاختبار Y بعد تحويل الدرجات هو التوزيع التكراري نفسه المشروط للدرجات عند مستوى معين من مستويات القدرة (Θ) للاختبار X أي أن $f([Y|\theta]) = f([X|\theta])$.

٣. **اللاتباين في مجتمع الدراسة Population Invariance**: أي أن تحويل

الدرجات يجب أن يبقى كما هو بصرف النظر عن مجموعة المفحوصين التي تم اشتقاقه من نتائجها في الاختبار.

٤. **التساوق (التماثل) Symmetry**: ويعني أن تحويل الدرجات من صورة إلى أخرى في الاختبار يجب أن يكون قابلاً للانعكاس . Invertible أي أن الرسم البياني لتوزيع الدرجات Mapping من الصورة X للاختبار إلى الصورة Y يجب أن يكون هو الانعكاس نفسه من الصورة Y للاختبار إلى الصورة X من الاختبار نفسه.

٥. **الثبات**: الاختبارات التي تتمتع بثبات كامل هي التي يمكن معادلتها.

٦. **الاختبارات المتوازية Parallel tests**: الاختبارات المراد معادلتها يجب أن تكون متوازية (متطابقة / متماثلة). ويعني التطابق هنا أن الوسط الحسابي، والانحراف المعياري، والتباين، ومستوى الصعوبة، ومستوى التمييز لفقرات الصورة الأولى من الاختبار تساوي نظيراتها في الصورة الثانية من الاختبار نفسه.

تصميمات جمع البيانات في معادلة الاختبارات

تتطلب عملية معادلة الاختبار جمع بيانات ميدانية تجريبية. كما أن طرق وإجراءات جمع البيانات الخاصة بمعادلة درجات الاختبارات يجب أن توفر نوعاً من الشيوع Communality بين بيانات فقرات الاختبارات المراد معادلتها لدى المفحوصين. وبصرف النظر عن نوع التصميم المستخدم في جمع البيانات الخاصة بمعادلة الاختبارات، فإنه لكي يكون التصميم مناسباً لمعادلة الاختبار يجب أن تكون لدينا مجموعة مشتركة من المفحوصين تُطبق عليها صورتا الاختبار؛ أو أن تكون لدينا فقرات مشتركة بين صورتي الاختبار التي تُطبق على جميع المفحوصين.

أهم التصميمات الخاصة بجمع بيانات معادلة الاختبار تشمل التصميم المتوازن عكسياً، والتصميم ذا المجموعات المتكافئة، والتصميم ذا الاختبار المشترك. وفي الواقع هناك امتداد فرعي لكل من التصميمات الرئيسة الثلاثة هذه؛ مثل تصميم المجموعة المفردة، وتصميم المجموعات العشوائية المتوازنة عكسياً، وتصميم المجموعات المتكافئة، وتصميم المجموعات العشوائية المرتبطة بفقرات مشتركة (اختبار مشترك anchor test) (Peterson et. al., 1989; Kolen & Brennan, 1995) نستعرض في الفقرات التالية التصميمات الرئيسة تلك وامتداد كل منها.

(١). تصميم المجموعة المفردة.

ويعدّ أبسط تصميمات جمع البيانات في عملية معادلة الاختبارات. وهنا نطبق صورتي الاختبار المراد معادلته على المجموعة نفسها من المفحوصين. ويجب أن تطبق الصورتان الواحدة تلو الأخرى، وفي اليوم نفسه، لتفادي العوامل التي قد تؤثر سلباً على أداء المفحوصين، كالتعب، والتعلم السابق، والممارسة practice ، كما يجب ألا تكون هناك فروق في مستويات صعوبة الفقرات في صورتي الاختبار، وعدم وجود فروق في مستويات القدرة (السِّمة) لدى المفحوصين.

(٢). تصميم المجموعة العشوائية المتوازنة عكسياً.

يستخدم هذا التصميم بشكلٍ كبير في المواقف العملية أكثر من استخدام تصميم المجموعة المفردة. والفكرة وراء هذا التصميم هي ضمان أن العوامل المذكورة سابقاً (التعب، وأثر التعلم السابق، والممارسة والخبرة) لها التأثير نفسه في صورة الاختبار قيد المعادلة. وفي هذا التصميم يتم تقسيم مجموعة واحدة من المفحوصين إلى مجموعتين متساويتين تقسيماً عشوائياً، وتطبيق صورتي الاختبار عليهما بطريقة التوازن العكسي. أي أن إحدى المجموعتين تأخذ الصورة الأولى للاختبار، والمجموعة الثانية تأخذ الصورة الثانية (الجديدة). ثم تأخذ كل مجموعة الصورة التي لم تأخذها في المرة الأولى، مباشرة بعد أخذها الاختبار السابق، بشرط أن يكون الوقت المحدد لتقديم صورتي الاختبار متساوياً للمجموعتين، وأن تكون كتيبات الاختبار معّدة بشكلٍ متسلسل بحيث يكون كل كتيب صورة جديدة من الاختبار نفسه.

(٣). تصميم المجموعات المتكافئة

هذا النوع من التصميمات يتجنب تماماً العوامل التي تؤثر في أداء الطالب سلباً، أو إيجاباً. والمسلمة، أو الفرضية التي ينطلق منها هذا التصميم هي أنه لا يمكن ترتيب وقت اختبار لكل مفحوص على حدة في أكثر من صورة واحدة من الاختبار؛ لذلك لا يتطلب الأمر هنا أن يتقدم المفحوص لكل صورة من صور الاختبار؛ لذلك يتم تطبيق صورة واحدة من صور الاختبار المتكافئة على المفحوصين، أو صورة واحدة فقط تُطبق على كل مجموعة. ويجب أن تكون المجموعتان متماثلتين في السِّمة المراد قياسها بغرض إزالة أي تحيز غير معروف في الفقرات خلال عملية معادلة الاختبارات. وللتخلص من الفروق العشوائية في قدرات المفحوصين، يجب تطبيق الاختبار على عينات كبيرة.

(٤). تصميم المجموعات العشوائية ذات الاختبار المشترك.

يمتاز هذا التصميم من تصميم المجموعات المتكافئة المشار إليه سابقاً بأنه يسمح بالتخلص من الفروق العشوائية في السِّمة المراد قياسها بين مجموعتي المفحوصين. وفي هذا التصميم تُطبق الصورة (X) من الاختبار على المجموعة الأولى، وتطبق الصورة (Y) من الاختبار نفسه على المجموعة الثانية؛ ثم يتم تطبيق اختبار مشترك anchor test على المجموعتين معاً. ويجب أن يُطبق هذا الاختبار المشترك على المجموعتين بالترتيب نفسه، بحيث تكون درجات أفراد المجموعتين على صورتي الاختبار وعلى الاختبار المشترك متأثرة بشكلٍ متساوٍ بعوامل التعب، وأثر التعلم، والخبرة السابقة.

يمكن أن يكون الاختبار المشترك داخلياً،أو خارجياً. فالاختبار المشترك الداخليّ يحتوي على مجموعات جزئية من الفقرات يتم تضمينها في صورتي الاختبارين المراد معادلتهما. أما الاختبار المشترك الخارجيّ فهو اختبار ذو فقرات مشتركة تُطبق في وقتين مختلفين على المجموعتين. تُستخدم الدرجات الناتجة عن تطبيق الاختبار المشترك الداخليّ في حساب درجات الطلبة في الاختبارين. ومن فوائد الاختبار المشترك بشكلٍ عام، استخدام درجاته لتقدير مستوى المجموعة المشتركة من المفحوصين في صورتي الاختبار، وذلك لإجراء عملية محاكاة للواقع عندما تُطبق صورتا الاختبار على المجموعة نفسها من المفحوصين. وتعتمد فائدة الاختبار المشترك على درجة العلاقة بين درجات المفحوصين في هذا الاختبار، ودرجاتهم في كلٍ من صورتي الاختبار (القديمة، والجديدة).

(٥). تصميم المجموعات غير المتكافئة ذات الاختبار المشترك.

نظراً لوجود اعتبارات تتعلق بقضية أمن الاختبارات، وسريّتها في الكثير من برامج الاختبارات، فإنه من غير المنطقيّ، ومن غير العمليَ تطبيق أكثر من صورة واحدة من الاختبار نفسه. وفي هذه الحالة يكون تصميم المجموعات غير المتكافئة ذات الاختبار المشترك هو المناسب. يتم في هذا التصميم جمع البيانات المتعلقة بمعادلة الاختبار بالطريقة نفسها التي تم بها في تصميم المجموعات العشوائية ذات الاختبار المشترك. والفرق بين التصميمين أنه يكون لدينا في هذا التصميم مجموعات غير متكافئة، حيث يتم تطبيق صورة واحدة من الاختبار مع الاختبار المشترك على مجموعة واحدة من المفحوصين في فصل الخريف مثلاً، ونطبق الصورة الأخرى من الاختبار نفسه مع الاختبار المشترك على المجموعة الأخرى في الربيع، ويجب أن يكون الاختبار المشترك مساوياً للاختبار المراد معادلته في مستوى الصعوبة والمحتوى.

وتجدر الإشارة هنا إلى أن جميع تصميمات جمع البيانات التي أشرنا إليها تتطلب تطبيق الاختبار المراد معادلته على مجموعات المفحوصين المسحوبة كعينة من مجتمع الدراسة نفسه.

طرق معادلة الاختبارات

يُصَّنفُ المتخصصون في القياس والتقويم التربوي والنفسي طرق معادلة الاختبارات إلى نوعين. يندرج النوع الأول ضمن الطرق التي تعتمد على النظرية التقليدية في الاختبارات Classical Test Theory (CTT). أما النوع الثاني فيُصنَّفُ ضمن الطرق التي تعتمد على النظرية الحديثة في القياس التربوي (نظرية الاستجابة لمفردة الاختبار). Theory (IRT) Item Responseوسنعرض للطرق الأساسية لمعادلة الاختبارات ضمن هذين النوعين، دون التعمق في بعض القضايا الفنية البحتة التي ربما تكون خارج نطاق هذا البحث، مثل: المعادلة الدائرية للاختبارات Circular Equating ، وطريقة اللب (النواة) . Kernel Method

أولاً: طرق معادلة الاختبارات بواسطة النظرية التقليدية للاختبارات

الطرق الرئيسة المستخدمة في معادلة الاختبارات ضمن هذه النظرية هي معادلة الاختبارات بواسطة الوسط الحسابي، ومعادلة الاختبارات خطياً، ومعادلة الاختبارات بواسطة الرُّتب المئينية المتساوية . (Keeves, 1988; Kolen, 1988; Kolen & Brennan, 1995)

(Suen, 1990; Dorans, 1990; Crocker & Algina, 1986; Peterson et. al., 1989;

وسنتناول كل طريقة على حدة، مدعمين شرحنا بمثال تطبيقي كلما أمكن ذلك.

(١) معادلة الاختبار بطريقة الوسط الحسابي Mean Equating

في هذه الطريقة نجعل الوسط الحسابي للدرجات متساوياً في صورتي الاختبار لمجموعة المفحوصين التي تُستخدم بياناتها في إجراء عملية معادلة الاختبار. أي يتم إجراء تحويل رياضي يؤدي إلى تساوي متوسطي الدرجات في صورتي الاختبارين. مثلاً، إذا كان متوسط درجات المفحوصين في الصورة الأولى للاختبار هو ٧٧ درجة، ومتوسط درجاتهم في الصورة الثانية من الاختبار هو ٧٢ درجة. فذلك يعني أن الدرجة ٧٧ في الصورة الأولى يتم تحويلها رياضياً لتصبح مكافئة للدرجة ٧٢ في الصورة الثانية؛ ويكون الفرق بين الدرجتين (٧٧−٧٢=٥) ثابتاً خلال عملية تدريج الدرجات. Calibration فإذا حصل طالب على درجة ٧٥ في الصورة الثانية للاختبار، تكون درجته في الصورة الأولى للاختبار ٨٠ درجة.

ويمكن التعبير رياضياً عن معادلة الاختبار بواسطة الوسط الحسابي على النحو التالي:

$$X_1 - \bar{X}_1 = X_2 - \bar{X}_2$$

حيث X_1 و X_2 هما درجة الطالب في الصورة الأولى للاختبار، ودرجته في الصورة الثانية للاختبار، على التوالي. وأن X_1 و X_2 هما متوسط الدرجات في الصورة الأولى للاختبار والصورة الثانية له على التوالي. ويمكن وضع المعادلة السابقة (بالنسبة إلى X_1) في الصورة الآتية:

$$X_1 = X_2 - \bar{X}_2 + \bar{X}_1$$

وتعني هذه المعادلة حساب درجة طالب في الصورة الأولى للاختبار من خلال درجته في الصورة الثانية للاختبار نفسه، أي أن:

$$X_1 = X_2 - 77 + 72 = X_2 - 5$$

فمثلاً، إذا حصل طالب على درجة ٧٥ في الصورة الثانية للاختبار، تكون درجته في الصورة الأولى، للاختبار مكافئة للدرجة ٧٠ في الصورة الأولى. أي أن:

$$X_1 = X_2 - 5 = 75 - 5 = 70 \quad .$$

(٢) معادلة الاختبار خطـــياً Linear Equating .

في هذه الطريقة يتم تحويل الدرجات رياضياً بحيث تتساوى المتوسطات الحسابية، والانحرافات المعيارية للفقرات في صورتي الاختبار المراد معادلتهما للمجموعة المستهدفة من المفحوصين من خلال بياناتهما. وتنطلق طريقة معادلة الاختبار خطياً من فرضية مؤداها أن الفروق في مستويات الصعوبة بين الفقرات في صورتي الاختبار المراد معادلته تكون ثابتة على امتداد سلم تدريج الدرجات. وفي كثير من الأحيان تُعَدُّ الفروق النسبية بين مستويات صعوبة الفقرات على أنها متغير variable على امتداد تدريج (سُلَّمْ) الدرجات. فقد تكون الصورة الأولى للاختبار أكثر صعوبة نسبياً للطلبة منخفضي التحصيل منها لدى الطلبة مرتفعي التحصيل؛ إذ إن التحويلات الخطية linear transformations هنا تسمح للصعوبة النسبية للفقرات بأن تتباين على امتداد تدريج الدرجات.

ويمكن تحويل الدرجات خطياً بتحويلها من درجات خام إلى درجات معيارية زائية، ومتساوية في صورتي الاختبار ($Z_1 = Z_2$) ، وذلك على النحو التالي:

$$\frac{X_1 - \overline{X}_1}{S_1} = \frac{X_2 - \overline{X}_2}{S_2}$$

حيث S_1 و S_2 هما الانحرافان المعياريان للفقرات في كل من صورتي الاختبار على التوالي. فللحصول مثلاً على درجة طالب في الصورة الأولى للاختبار من مكافئها في الصورة الثانية للاختبار، تتحول المعادلة السابقة إلى الآتي:

$$X_1 = \frac{S_1}{S_2} X_2 + \left[\overline{X}_1 - \frac{S_1}{S_2} \overline{X}_2 \right] = AX_2 + B \quad ، \text{ حيث:}$$

$$A = \frac{S_1}{S_2} \quad ، \quad B = \overline{X}_1 - \frac{S_1}{S_2} \overline{X}_2$$

وأن هو ميل التحويل الخطي slope ، B هو القاطع . intercept فعلى سبيل المثال، لو افترضنا أن متوسط درجات المجموعة في الصورة الأولى للاختبار هو ٧٢، ومتوسط درجاتهم في الصورة الثانية له هو ٧٧. وأن الانحراف المعياري لفقرات كل من صورتي الاختبار هو ٩، ١٠ على التوالي. فإن التحويل الخطي المطلوب للحصول على درجات معادلة لصورة من الاختبار إلى صورة أخرى هي على النحو التالي:

$$X_1 = \frac{9}{10} X_2 \left[72 - \frac{9}{10}(77) \right] = 0.9X_2 + 2.7$$

فإذا حصل طالب مثلاً على الدرجة ٨٠ في الصورة الثانية للاختبار، فإن الدرجة التي تعادلها في الصورة الأولى للاختبار هي:

$$X_1 = 0.9(80) + 2.7 = 74.7$$

(٣) طريقة الرُّتب المئينية المتساوية Equipercentile Equating

في هذه الطريقة نجعل التوزيع الإحصائي للدرجات في الصورة الثانية للاختبار مساوياً لنظيره في الصورة الأولى، وذلك لمجموعة المفحوصين؛ لذلك فإن تحويل درجات الصورة الثانية للاختبار رياضيا، باستخدام الرُّتب المئينية المتساوية، له الوسط الحسابي نفسه، والانحراف المعياري، والشكل التوزيعي الإحصائي نفسه تقريباً (الالتواء والتفلطح، .. إلخ) كنظيره في الصورة الأولى للاختبار. وهذه التحويلات الرياضية في معادلة الاختبار بهذه الطريقة يمكن استخدامها في أي من التصميمات المشروحة سابقاً. ومن هنا، فإن الرُّتب المئينية المتساوية لدرجات المفحوصين في صورتي الاختبار تدل على تساوي مستوى الأداء في كلتا الصورتين. وهذا هو لُبّ معادلة الاختبار بهذه الطريقة. أما الجانب الإجرائي في معادلة الاختبار بهذه الطريقة، فيتم بواسطة رسم بياني لدرجات الطلبة في صورتي الاختبار معاً يبين التوزيع التكراري النسبي المتجمع (الرُّتب المئينية مقسومة على ١٠٠). وبذلك تكون درجات الطلبة في أي من الصورتين متكافئة إذا تساوت رتبها المئينية.

مقارنة بين الطرق التقليدية لمعادلة الاختبار

يمكن تركيز الفروق بين طرق معادلة الاختبار التقليدية في الآتي .(Brennan, 1995 & Kolen)

- يمكننا القول بشكلٍ عام أنه إذا كانت صورتا الاختبار المراد معادلتهما لهما مقدار الانحراف المعياري نفسه، فإن معادلة الاختبار بطريقة الوسط الحسابي، وبطريقة المعادلة الخطية تؤديان إلى النتائج نفسها. أما إذا كان شكل التوزيع الإحصائي للدرجات في الصورتين هو نفسه (نفس الالتواء والتفلطح، إلخ ..)، فإن طريقة المعادلة الخطية للاختبار، وطريقة الرُّتب المئينية المتساوية تعطيان النتائج نفسها.

- تتطلب معادلة الاختبار بطريقة الرُّتب المئينية المتساوية عينة كبيرة من المفحوصين؛ لأنها أكثر تعقيداً من الناحية الرياضية من طريقة الوسط الحسابي، وطريقة المعادلة الخطية. وينطبق ذلك بشكلٍ خاص على تصميم المجموعات غير المتكافئة الذي يتضمن فقرات اختبارية مشتركة بين صورتي الاختبار.
- معادلة الاختبار خطياً هي عملية تحليلية بحتة، بخلاف طريقة الرُّتب المئينية المتساوية.
- لا يمكن للتحويل الرياضي بطريقتي المعادلة الخطية، والرُّتب المئينية المتساوية أن يُنْتِجَ معادلة اختبار صحيحة، وتامة، ما لم يتم استيفاء شرط العدالة (المساواة) .Equity

خطأ معادلة الاختبار Equating Error

من المشكلات المهمة التي يتصدى لها المتخصصون في معادلة الاختبارات مشكلة الخطأ الناجم عن عملية المعادلة نفسها، وكيفية تقليص هذا الخطأ إلى الحد الأدنى، لرفع مستوى دقة القياس إلى الحد الأقصى من جهة، وللحصول على معادلة اختبار صحيحة، يمكن تفسير نتائجها تفسيراً علمياً، وعملياً صحيحاً.

ولا شك أن الأنواع المختلفة لأخطاء معادلة الاختبار تؤثر في تفسيرنا للنتائج من خلال تطبيق طرق معادلة الاختبار؛ لذلك يتوجب انتقاء تصميمات معادلة الاختبار بعناية فائقة؛ لخفض خطأ القياس إلى حدِّه الأدنى؛ آخذين في الاعتبار الاعتبارات العملية أثناء التطبيق. وفي بعض الأحيان يكون مقدار الخطأ من الضخامة بحيث يُفضَّل عدم إجراء المعادلة.

هناك نوعان من الخطأ يؤثران في عملية معادلة الاختبار؛ هما الخطأ العشوائي Random error ، والخطأ المنتظم . Systematic error أما الخطأ العشوائي فهو موجود دائماً متى استخدمنا عينات من مجتمع الدراسة لتقدير مَعْلَم Parameter أو أكثر (كالوسط الحسابي، والانحراف المعياري، وغيرهما). ولكن يمكن خفض هذا النوع من الخطأ باستخدام عينات كبيرة من المفحوصين المطبقة عليهم معادلة الاختبار، ثم اختيار التصميم الأنسب لعملية المعادلة. أما إذا تطلب الوضع العملي لمعادلة الاختبار استخدام عينات صغيرة من المفحوصين، يصبح الخطأ العشوائي عندئذٍ مشكلة كبرى. أما الخطأ المنتظم لمعادلة الاختبار، فينجم عادة عن عدم التزام الباحث بالافتراضات، والشروط التي تتطلبها طريقة معينة في المعادلة. فمثلاً: في تصميم المجموعات المفردة، إذا أخفقنا في ضبط الآثار الناجمة عن عاملي التعب، والخبرة والممارسة، وهما مصدران أساسيان للخطأ المنتظم،

نكون قد قللنا من دقة وصحة معادلة الاختبار. كذلك في تصميم المجموعات العشوائية، ينتج الخطأ المنتظم عن عدم قدرتنا على جعل مجموعة المفحوصين متفاوتة، أو متكافئة في مستوى القدرة (السِّمة) المراد قياسها، وفي مستوى صعوبة الفقرات. وتكبر المشكلة إذا كانت لدينا مجموعات غير متكافئة. أما في تصميم المجموعات ذات الاختبار المشترك، فنجد أن الخطأ المنتظم يأخذ في الزيادة إذا كانت الفقرات المشتركة تلك لا تمثل الاختبار ككل في محتواه، وفي خصائصه الإحصائية. كذلك تزيد قيمة الخطأ المنتظم إذا كان تأثير الفقرات المشتركة مختلفاً في صورتي الاختبار. ويمكننا القول بشكلٍ عام، أن أي تصميم من تصميمات معادلة الاختبار يُنْتِجُ الخطأ المنتظم إذا كانت صورتا الاختبار مختلفتين في المحتوى، وفي مستوى صعوبة الفقرات، وفي الثبات.

معادلة الاختبارات بطريقة نظرية الاستجابة لمفردة الاختبار

إن معادلة الاختبارات المعتمدة على الدرجات الخام raw scores ضمن النظرية التقليدية للاختبارات قد لا تكون مرغوبة؛ وذلك بسبب إخفاقها في تحقيق جميع شروط معادلة الاختبارات التي أشرنا إليها سابقاً، وهي العدل، والمساواة، والتساوق، واللاتباين (Hambleton & Swaminathan,1985)؛ لذلك فإن معادلة الاختبار عن طريق نظرية الاستجابة لمفردة الاختبار (النظرية الحديثة في القياس التربويّ، والنفسيّ) Theory Item Response تحل الكثير من المشكلات التي عجزت عن حلها النظرية التقليدية في الاختبارات؛ بشرط أن يكون النموذج IRT model المستخدم في النظرية الحديثة هذه مطابقاً للبيانات المُعدة لعملية المعادلة.

ويشير بيكر، والقرني (Baker & Alkarni, 1991, p. 147) إلى أنه من الإسهامات الكبيرة للنظرية الحديثة في القياس التربويّ، والنفسيّ، في الممارسة التربوية، قدرة هذه النظرية على وضع عدة اختبارات، ومجموعات من المفحوصين على تدريج مشترك common scale في عملية القياس؛ وإمكانية استخدامها في المعادلة الأفقية والرأسية للاختبار.

وتقوم نظرية الاستجابة لمفردة الاختبار على مجموعة من الافتراضات، تؤدي إلى تفسير صحيح لنتائج الاختبار ومعادلته، بشرط يتم تطبيقها بشكلٍ صحيح ودقيق. تفترض هذه النظرية أن أداء المفحوصين في الاختبار يمكن تفسيره عن طريق السِّمة، أو السِّمات الكامنة latent traits المراد قياسها، والتي لا يمكن قياسها بصورة مباشرة؛ إذ يتم استخدام الدرجات

التي تم تقديرها للمفحوص في تلك السِّمة في التنبؤ بأدائه في اختبارٍ ما، أو في فقرة من الاختبار؛ لأن العلاقة الحقيقية بين الدرجات المشاهدة (الخام) للمفحوص، والسِّمة المراد قياسها لا يمكن الحصول عليها بطريق مباشر. ومن هنا، تقوم نظرية الاستجابة لمفردة الاختبار بوصف هذه العلاقة بواسطة دالة تعتمد على مجموعة من الافتراضات، هي: أحادية البُعْد (أحادية السِّمة) unidimensionality حيث يقيس الاختبار سَّمة واحدة فقط؛ والاستقلال الموضعي local independence وهو استقلال أداء المفحوص على فقرة اختبار عن أدائه على فقرة أخرى. وكذلك جعل السِّمات الأخرى التي تؤثر في أداء المفحوص ثابتة، ومتسقة . consistent أما الافتراض الأخير فهو افتراض اللاتباين invariance، والذي يعني أن معالم (معلمات) parameters الفقرة (الصعوبة، والتمييز، والتخمين) لا تعتمد على التوزيع الإحصائي للسِّمة المراد قياسها؛ وأن المعالم التي تصف أداء المفحوصين لا تعتمد على فقرات الاختبـــار (Lord, 1980; Hambleton, Swaminathan, & Rogers, 1991) ويمكن التعبير عن العلاقة الرياضية بين درجة المفحوص (الدرجة الخام) على الفقرة، ودرجته على سُلَّم السِّمة المراد قياسها بعدة نماذج رياضية في النظرية الحديثة في القياس التربويّ، والنفسيّ، وهي نموذج راش Rasch أحادي المَعْلَم (نسبةً إلى العالم الرياضي الدنمركي جورج راش)، والنموذج ثنائي المَعْلَمْ two-parameter model ، والنموذج ثلاثي المَعْلَم three-parameter model ؛ وهي على النحو التالي:

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

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

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

حيث $P(\theta)$ = احتمال السِّمة المراد قياسها.

a = معامل تمييز الفقرة (قيمته ثابتة في نموذج راش).

b = معامل صعوبة الفقرة.

c = مَعْلَم التخمين للفقرة.

D = مَعْلَم تدريج scale parameter يجعل الدالة اللوجستية logistic function قريبة إلى الحد الأقصى من دالة القوسية الطبيعية .normal ogive function وقيمة هذا المَعْلَم تساوي ١,٧

e = الثابت الرياضي الذي يحول الدالة التي تربط بين الدرجة الخام، والسِّمة المراد قياسها، من دالة ما لانهاية إلى دالة احتمالية تحصر العلاقة بين أداء الطالب على الفقرة، وصعوبة الفقرة بين الصفر والواحد الصحيح. وقيمة هذا الثابت الرياضي تساوي ٢,٧١٨.

طرق معادلة الاختبار بواسطة النظرية الحديثة في القياس التربويّ

هناك ثلاث طرق رئيسة لمعادلة الاختبار بهذه النظرية (Kolen & Brennan, 1995; 1991 Lord, 1980; Peterson et. al., 1989; Hambleton, Swaminathan, & Rogers,)، نعرضها بصورة موجزة ومركزة على النحو التالي:

(أ) معادلة الاختبار باستخدام درجات القدرة (السَّمة) Ability Score Equating.

في هذا النوع من معادلة الاختبارات نفترض أن كلا الاختبارين يقيسان القُدرة، أو السِّمة نفسها. فإذا قمنا بقياس معلمات القدرة هذه، ومعلمات الفقرات في الوقت نفسه لكل من الاختبارين، تكون هذه المعلمات على التدريج نفسه؛ لذلك فإننا نحتاج إلى معادلة درجات القدرة أثناء معايرة calibration الاختبارات. وتعني المعايرة هنا تقدير معلمات الفقرات، كالصعوبة، والتمييز.وينتج عن ذلك علاقة خطية بين تدريج القدرة بعد تقدير معلماتها، وتقدير معلمات الفقرات. بشرط أن يكون التقديران منفصلين.

(ب) معادلة الاختبار باستخدام الدرجة الحقيقية True-Score Equating.

في النظرية التقليدية للاختبارات، يُطْلَقُ على الدرجة المتوقعة للمفحوص (g) على الفقرة (i) مصطلح «الدرجة الحقيقية». بينما في النظرية الحديثة في القياس التربويّ، والنفسيّ، تصاغ رياضياً على الصورة: $\sum_{i=1}^{n} P_i(\theta_g)$ ، .

فلو افترضنا أن الاختبار بصورتيه يقيس السِّمة نفسها، وأن معلمات الفقرات لكلا الصورتين أو الاختبارين قد تم وضعهما على التدريج نفسه؛ فإن الدرجة الحقيقية الصحيحة للطالب هي ξ في الاختبار X، وتكون η في الاختبار Y ؛ وتكونان مرتبطتين بدرجات القدرة بواسطة دالة خاصية الاختبار Test characteristic curve ، بالصورة الرياضية الآتية (انظر Lord, 1980) :

$$\eta = \sum_{i=1}^{n} P_j(\theta) \qquad , \qquad \zeta = \sum_{i=1}^{n} P_i(\theta)$$

ويجب أن تكون تدريجات القياس للدرجات ξ ، η مستقلة عن مجموعة المفحوصين

n في الاختبار، وأن يكون تدريج قياس درجات القدرة مستقلاً عن عدد فقرات الاختبارين. كما يُشترط أن تكون مستويات صعوبة الفقرات متطابقة في الاختبارين.

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

(ج) معادلة الاختبار بطريقة الدرجات المشاهَدَة (الدرجات الخام) Observed-score Equating

من المشكلات التي يواجهها المتخصصون في معادلة الاختبار بطريقة الدرجات الحقيقية، أن هذه الطريقة لا تنتج عنها درجات معادلة للمفحوص الذي درجته الخام أدنى من مستوى الصدفة (C) ؛ وذلك بسبب أن العلاقة بين الدرجات الخام ليست مثل العلاقة بين الدرجات الحقيقية. ففي الدرجات الخام تكون أدنى درجة هي الصفر، وفي الدرجات الحقيقية تكون أدنى درجة هي $\sum_{i=1}^{n} C_i$ ، أي درجة التخمين.

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

$$f(X|\theta_g)=\binom{n}{x}P^xQ^{n-x}$$

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

٣- وضع معلمات القدرة، ومعلمات الفقرات على تدريج مشترك لكل المجموعات والاختبارات.

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

$$f(X)=\sum_{g=1}^{n} f(X|\theta_g)$$

٣- تكرار الخطوة رقم (٢) للاختبار الثاني.

٤- إجراء معادلة للاختبار بطريقة الرُّتب المئينية المتساوية بين الدرجات الخام في الاختبار الأول، ونظيراتها في الاختبار الثاني؛ وذلك باستخدام التوزيع التكراري الهامشي الذي تم إنشاؤه. ومن المهم جداً في معادلة الاختبار بهذه الطريقة تغطية مدى الدرجات الخام بالكامل.

خطوات معادلة الاختبارات بواسطة النظرية الحديثة في القياس التربوي

يوضح كل من هامبلتون، وسواميناثان (Hambleton & Swaminathan, 1985)، وكذلك كولن، وبرينان (Kolen & Brennan, 1995)، الخطوات الضرورية لمعادلة الاختبارات بواسطة النظرية الحديثة في القياس التربويّ، والنفسيّ؛ وهي على النحو التالي.

- اختيار التصميم المناسب لمعادلة الاختبار مع الأخذ بعين الاعتبار خصائص مجموعة المفحوصين، وطبيعة الاختبارات المراد معادلتها.
- اختيار النموذج المناسب الذي يطابق التصميم المناسب، والاختبار المناسب (نموذج راش، أو غيره من نماذج هذه النظرية).
- بناء تدريج مشترك يربط العلاقة بين السِّمة المراد قياسها ومَعْلَم الفقرة . item parameter
- اختيار التدريج المناسب لوضع درجات الاختبار، أي: هل تُكْتَبُ الدرجاتُ كدرجات خام؟ (درجات مُشاهَدَة)، أو على صورة درجات قُدرة؟ ability scores ، أو على صورة درجات حقيقية مُقدَّرة؟ . estimated true -score ويمكن أن تقوم بهذه المهمة المعقدة رياضياً بعض البرامج الحاسوبية، مثل برنامج BILOG ،وبرنامج MULTILOG، وبرنامج LOGIST ، وغيرها. ومن المتعذر عملياً إلى حدٍ كبير القيام بهذه العملية يدوياً، وخاصة في هذه النظرية.

ويشرح كوك، وآيجنور (Cook & Eignor, 1991) الآليات الأساسية لعملية معادلة الاختبارات في هذه النظرية، والتي يمكن إيجازها في الآتي:

- **اختيار التصميم المناسب**: هناك ثلاثة تصميمات أساسية تستخدمها هذه النظرية لمعادلة

الاختبارات، وهي: تصميم المجموعة المفردة، وتصميم المجموعات العشوائية، والتصميم ذو الاختبار المشترك. ويعتمد حجم العينة الملائم لإجراء معادلة الاختبار بشكلٍ صحيح على العدد الملائم للمفحوصين للحصول على تقديرات مستقرة للمعلمات المستخدمة في النموذج المختار لوضع الدرجات، ومعلمات القدرة (السِّمة) على تدريج واحد،، وفي عملية المعايرة هذه نحتاج إلى عينة تصل إلى ٣٠٠٠ مفحوص. وإذا كانت هناك فقرات مشتركة بين صورتي الاختبار على هيئة اختبار مشترك anchor test ، فيجب أن تعكس هذه الفقرات المحتوى، والخصائص الإحصائية لصورتي الاختبار، وألاَّ تقل نسبة فقرات الاختبار المشترك عن ٢٠٪ من الطول الاحتمالي للاختبار. كما أن اختيار أي من التصميمات المشار إليها أعلاه يرتبط بنوعية البرنامج الحاسوبي المستخدم لمعادلة الاختبار.

- وضع تقديرات المعالم (المعلمات) على تدريج مشترك: لو افترضنا أن مجموعتين من المفحوصين طُبقت عليهما المجموعة نفسها من فقرات الاختبار، وتم تقدير معلمات الفقرة (الصعوبة، والتمييز) لكل مجموعة على حده، وأن النموذج الرياضي المستخدم هو النموذج ثنائي المعلم، حيث إن منحنيات خاصية الفقرة characteristic curves (ICC) item مستقلة عن المجموعتين المستخدمتين لرسم هذه المنحنيات. ويمكن أن نستنتج من ذلك أن تقديرات معلمات الفقرات متطابقة في كلٍ من المجموعتين على حده. ويجب هنا الأخذ بعين الاعتبار أثر خطأ المعاينة .sampling error ولكن الواقع ليس كذلك! إذ يجب إجراء تحويل رياضي معين يوحد التدريج في عملية المعايرة. ويعتمد ذلك طبعاً على نوع البرنامج الحاسوبي المستخدم في إجراء عملية التحويل الرياضي للحصول على نقطة أصل. ووحدة قياس للسِّمة ولمستوى الصعوبة. ويكون متوسط درجات القدرة (السِّمة) هو الصفر، وانحرافها المعياري هو الواحد الصحيح (برنامج LOGIST يمكن أن يحقق ذلك). ويجب ملاحظة أن عملية التحويل الرياضي التي تضع الدرجات وتقديرات معلمات السِّمة على التدريج نفسه تستهدف الحصول على معلمَي المَيْل slope، والقاطع intercept، بحيث يكون الوسط الحسابي، والانحراف المعياري لتوزيع مستويات صعوبة الفقرات التي تم تقديرها في عملية المعايرة الثانية للدرجات، مساويين لنظيريهما اللذين تم تقديرهما في المعايرة الأولى للدرجات
- **معادلة درجات الاختبار**: تُعَدُّ عملية معادلة الاختبار منتهية إذا تم تقدير المعلمات لفقرات كل من صورتي الاختبار المستهدف، وتم وضعها على تدريج مشترك، وتم كذلك

الحصول على تقدير للسِّمة المراد قياسها لدى المفحوص؛ بحيث تكون هي نفسها في أي من صورتي الاختبار. ويؤخذ خطأ القياس بعين الاعتبار هنا. وبناء على ذلك، يكون التعبير عن الدرجات الخام بما يكافئها من درجات السِّمة. أما إذا أخفق البرنامج الحاسوبي المستخدم في استخراج نتيجة السِّمة، يلجأ المتخصصون إلى ترجمة، أو تحويل أي درجة من درجات القدرة إلى الدرجة الحقيقية المقدّرة المناظرة لها في صورتي الاختبار؛ وعدّها الدرجة التي تمت معادلتها في الاختبار. أما الصورة الرياضية للدوال التي تربط بين درجات القدرة، وتقديرات الدرجات الحقيقية فهي على النحو التالي:

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

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

حيث إن:

= الدرجة الحقيقية المقدرة للصورة الأولى من الاختبار.

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

$\hat{T}_Y$ = الدالة المقدِّرة لاستجابة الفقرة في الفقرات للصورة الأولى للاختبار.

$\hat{P}_i(\theta)$ = الدالة المقدِّرة لاستجابة الفقرة في الفقرات للصورة الثانية في الاختبار.

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

الفوائد العملية والتطبيقية لمعادلة الاختبار باستخدام النظرية الحديثة في القياس التربويّ، والنفسيّ.

تشير نتائج الأبحاث التي أجريت على معادلة الاختبار بواسطة هذه النظرية، إلى أن معادلة الاختبار بهذه النظرية لها فوائد جمة من الناحيتين العملية، والتطبيقية (Kolen & Brennan, 1995; Yang, 1997)، يمكن إيجازها في الآتي:

١. معادلة الاختبار بهذه النظرية هي الأفضل عندما تكون الاختبارات المختلفة فقراتها في مستويات الصعوبة مطبّقة على مفحوصين من مجموعات غير عشوائية، تختلف في مستويات السِّمة المراد قياسها.

٢. معادلة الاختبار بهذه النظرية، وبسبب خصائصها المشروحة سابقاً، تؤدي إلى تحويلات رياضية للدرجات تكون مستقلة عن مجموعة، أو مجموعات المفحوصين التي طُبّقت عليها الاختبارات.

٣. المعادلة بنظرية IRT هي أفضل من نظيرتها في النظرية التقليدية للقياس التربوي، وخاصة في النهايات العليا لتدريج الدرجات، إذ غالباً ما تتم عملية اتخاذ القرارات المهمة، وحيث يمكن معادلة الدرجات الخام مع كل قيم درجات السِّمة.

٤. المعادلة بهذه النظرية تسمح باستخدام الصورة القديمة للاختبار لمعادلة درجاتها بالصورة الجديدة للاختبار متى تم وضع الدرجات في كلا الاختبارين، وكذلك تقديرات معلمات القدرة، على تدريج واحد.

٥. معادلة الاختبار بهذه النظرية تسمح بمعادلة الاختبار على مستوى الفقرة الواحدة؛ وذلك قبل الشروع في تطبيق الاختبار في صوره المتعددة، وقبل تطبيق صورتيه أيضاً. ويتم ذلك إذا توافرت البيانات القبلية على مستوى الفقرة الواحدة، مع إمكانية معايرتها. ثم يتم وضع تقديرات المعلمات على تدريج مشترك. وهذه الخاصية لا يمكن الحصول عليها من خلال معادلة الاختبار بواسطة النظرية التقليدية في الاختبارات. ويوضح الجدول (١) طرق معادلة الاختبارات التقليدية، والحديثة، وبعض خصائصها (Kolen, 1988, P.149).

الجدول رقم (١)

طرق معادلة الاختبار وبعض خصائصها

طريقة المعادلة	الدالة	الإحصاءات المتشابه لصورتي الاختبار	طريقة عرض النتائج
الوسط الحسابيّ	Y=X+B	الوسط الحسابي	كثابت تحويل (B) ،وقاعدة تقريب
الخطية	Y=AX+B	الوسط الحسابيّ، والانحراف المعياريّ	كميل (A)، وقاطع (B)، وقاعدة تقريب
الرُّتب المئينية المتساوية	معقدة	الوسط الحسابيّ، والانحراف المعياريّ، وشكل توزيع الدرجات	تتطلب جدول تحويل للدرجات
نظرية الاستجابة لمفردة الاختبار(IRT)	معقدة		تتطلب برنامج حاسوبي لتحويل الدرجات

دراسات سابقة في معادلة الاختبارات

نعرض في هذا الجزء من الدراسة بعض الدراسات، والبحوث التي تناولت موضوع معادلة الاختبار في النظرية التقليدية للاختبارات، وفي النظرية الحديثة في القياس التربويّ، والنفسيّ على حدٍ سواء، والتي استخدمت طرقاً متعددة في معادلة الاختبارات في كلتا النظريتين.

ففي الدراسة التي أجراها رايت، ودورانز (Wright & Dorans, 1993) حول تحسين نتائج معادلة الاختبار عن طريق مطابقة متغير الانتقاء selection variable، استخدم الباحثان درجات اختبار من مجتمع دراسة حقيقي، وآخر مفترض، بواسطة اختبار مشترك؛ لغرض عملية المعادلة. وقد قارنا طريقة معادلة الاختبار الخطية، وطريقة الرُّتب المئينية المتساوية، فوجدا أن دقة معادلة الاختبار تتحسن باستخدام هاتين الطريقتين.

أما عن تأثير الأوزان التفاضلية في دقة وثبات معادلة الاختبار، فقد تم تطبيق إحدى طرق معادلة الاختبار متعددة الأبعاد multidimensional (MD) بواسطة النظرية الحديثة في القياس التربويّ على بيانات اختبار في الرياضيات لعينة من المفحوصين عن طريق مؤسسة

الاختبارات الجامعية الأمريكية، باستخدام تصميم الفقرات المشتركة غير المتكافئة، ودلت النتائج على تفوق طرق الأوزان التقليدية على الطرق الأخرى في تحسين دقة وجودة معادلة الاختبار (Kromrey, Parshall, & Yi, 1998).

أما لونز، وبرجستروم (Lunz & Bergestrom, 1995) فقد استخدما النظرية الحديثة في القياس التربويّ في مقارنة معادلة الاختبار لعينة مكونة من (٣٣٩٨)من الموظفين في الشئون الطبية، مرة باستخدام اختبار الورقة، والقلم،وأخرى باستخدام الاختبار التكيفي الحاسوبي Computer Adaptive Testing (CAT)، ودلت نتائج بحثهما على تقارب النتائج في الحصول على تقدير متشابه للسِّمة في الاختبارين.

أما بارشال، وزملاؤه (Parshall et. al. , 1991) فقد أجروا دراسة للتعرف على تأثير العينات الصغيرة في معادلة الاختبار، وقارنوا الأخطاء المعيارية، وكذلك التحيز الإحصائي، وطَبَّق الباحثون دراستهم على(١٠٠٠) عينة، يتراوح حجم كل منها بين ١٥، ٢٥، ٥٠، ١٠٠ من مختلف المواد المدرسية. واستخدم الباحثون طريقة معادلة الاختبار خطياً، وتصميم المجموعات غير المتكافئة. ونتج عن معادلة الاختبار تلك تحيزاً ضئيلاً جداً، مع وجود أخطاء معيارية كبيرة جداً.

وهناك بعض الدراسات التي استخدمت تصميم الفقرات المشتركة (الاختبار المشترك) لمعادلة الاختبار بواسطة نظرية الاستجابة لفقرة الاختبار (IRT)؛ لأغراض بحثية مختلفة في معادلة الاختبار. فعلى سبيل المثال، وجد ليستز، ويانغ (Lissitz, & Yang, 1999) أن معادلة الاختبار بطرق هذه النظرية أنتجت تحويلات رياضية دقيقة مع خليط من الفقرات المتنوعة (اختيار من متعدد، مقالية، إلخ ..)، وكذلك لكل نوع على حدة من أنواع الفقرات؛ وذلك للتعرف على أثر معادلة الاختبار بهذه النظرية في دقة التحويلات أثناء تقدير المعلمات.

أما سميث وجروس (Smith & Gross, 1997) فقد استخدما خمسة أنواع من الفقرات في اختبارات الترخيص licensure في برنامج اختبارات على مستوى الولايات المتحدة الأمريكية، بهدف الحصول على مستوى عال من الصدق، بطريقة ندلسكي المُعدلة لبناء محكات الأداء standard setting، وجد الباحثان أن هذه الطريقة ذات فاعلية كبيرة في معادلة الاختبار وخاصة أن نتائجها في حدود الخطأ المقبول في عملية المعادلة هذه.

أما فيما يتعلق بأثر تمثيل المحتوى content representativeness، والطريقة المستخدمة لمعادلة

الاختبار على دقة معادلة الاختبار، فقد دلت نتائج الدراسة التي أجراها يانغ (Yang, 1997) على نتائج صورتين من اختبار الكفاءة المهنية مجموع فقراتهما (١٩٧٢٠٣) فقرة، بأن معادلة الاختبار بواسطة نظرية الاستجابة لفقرة الاختبار أفضل من نظيرتها التي استخدمت فيها الطريقة الخطية. وفي دراسة أخرى قام بها يانغ، وهوانغ (Yang & Houang, 1996) استخدما فيها تصميم الفقرات المشتركة لمعادلة بيانات لصورتين من اختبار الحد الأدنى للكفاءة، وجدا أن طريقة نظرية IRT ، والطريقة الخطية تعطيان نتائج دقيقة في المتوسط.

أما سميث، وكريمر (Smith & Kramer, 1992) في مقارنتهما لنماذج راش المختلفة، وأثرها في دقة معادلة الاختبار؛ فقد استخدما بيانات المعادلة من ست صور متكافئة لاختبار التناظر الإدراكي، ووجدا فروقاً طفيفةً جداً بين تلك النماذج في دقة معادلة الاختبار.

أما فيما يتعلق بقضية معادلة اختبارات الكفاءة اللغوية، فقد قام كينيون، وستانزفيلد (Kenyon & Stansfield, 1993) باختبار قدرة أحد البرامج الحاسوبية المستخدمة في عملية معادلة الاختبار بنظرية IRT، وهو برنامج BIGSTEPS، من خلال إجراء عملية المعايرة المتزامنة، وكذلك في المعادلة الرأسية، وطبق الباحثان دراستهما على اختبارين في الكفاءة اللغوية في اللغة الصينية، وتوصلا إلى قدرة البرنامج المذكور على دقة عملية المعايرة تلك. كذلك موريسون، وفيتزباتريك (Morrison & Fitzpatrick, 1992) فقد توصلا إلى نتائج مشابهة لنتائج كينيون، وستانزفيلد، وخاصة فيما يتعلق بقدرة المعايرة المتزامنة على إعطاء المقدار الأقل من خطأ معادلة الاختبار.

أما جلواكي (Glowacki, 1991) فقد توصل إلى نتائج تختلف عن تلك التي توصل إليها الباحثون الذين أشرنا إليهم، فقد قام جلواكي باستخدام طريقة IRT، ومقارنتها بالطرق التقليدية في معادلة الاختبار، للتعرف على أكثرها ملاءمة لاختبار ولاية ألباما لتخريج طلبة المرحلة الثانوية، وقد توصل إلى أن الطرق التقليدية التي استخدمها في معادلة الاختبار أكثر ملاءمة إذا كانت صور الاختبار المراد معادلته قد تم تطويرها بواسطة الطرق التقليدية في بناء الاختبار. وفي دراسة لمقارنة معادلة الاختبار بطريقة اختبار الورقة، والقلم، والاختبار المعتمد على الوسائط المتعددة، وجد ستيبلز، ولوزو(Staples & Luzzo, 1999) أن نتائج التطبيقين متقاربة في معادلة الاختبار؛ وخاصة في فروق متوسط درجات التدريج، والارتباط بين التدريج المناظر، وأنماط الارتباطات البينية للتدريج.

نلاحظ من الدراسات التي عرضناها حول معادلة الاختبارات، سواء في النظرية

التقليدية للاختبار، أو في النظرية الحديثة، أن النظرية الحديثة في القياس التربويّ، والنفسيّ تعطي نتائج أفضل في معادلة الاختبار، وخاصة إذا استوفت شروط تطبيقها، وكانت عينة الدراسة كبيرة جداً، وتم استخدام البرنامج الحاسوبي المناسب من قبل متخصصين محترفين.

مشكلات فنية، وتطبيقية في معادلة الاختبار

يقول أنجوف (Angoff, 1980) أن أدبيات القياس والتقويم التربويّ، والنفسيّ قد اهتمت بتطوير نماذج لمعادلة الاختبار، ودوالها الخاصة بخطأ المعادلة، وكذلك مدى قوة وإحكام robustness هذه النماذج في مواجهة عدم استيفاء متطلبات الفروض الأساسية لتطبيق، وتطوير هذه النماذج.

نعرض في هذا الجزء من الدراسة بعض المشكلات الفنية، والتطبيقية التي يواجهها المتخصصون الذين يتولون مهمة معادلة الاختبارات، وتأثيرها في دقة عملية المعادلة، وبالتالي رفع مقدار الخطأ الناتج عن عملية المعادلة. ويمكن إيجاز هذه المشكلات في النقاط الآتية.

- **معادلة درجات في اختبارات غير متوازية للحصول على درجات متكافئة**: يشير لندكويست (Lindquist, 1967) إلى أن معادلة اختبارين تتطلب الحصول على درجات كلية متكافئة، بشرط الحصول على رتب مئينية متساوية للمجموعات التي طُبقت عليها الاختبارات. ولكن في بعض الأحيان، نجد أن بعض الطلبة قد أخذوا أحد الاختبارين، أو كليهما، مما يفسح المجال لتأثير عامل الزمن في درجات الطلبة تلك. وخاصة إذا تزامن مع ذلك اختلاف الاختبارين المراد معادلتهما في المحتوى إلى حدٍ ما. مما يعرقل عملية إحلال اختبار محل آخر ضمن عملية تكافؤ الاختبارين.

وفي مواقف أخرى، يود الباحثون أحياناً استخدام نتائج تقويم اختبارات المواد في المرحلة الثانوية؛ لغرض التعرف على جوانب القوة، والضعف البينشخصية individual-intra ؛وكذلك لمعرفة قيمتها الإرشادية، ثم جعلها متكافئة مع بعضها عن طريق معادلة فقراته (Lennon, 1967) والمشكلة في هذه الاختبارات أنه لا يمكن تطبيقها على مجموعة تقنين مفردة، أو مشتركة، وجذور هذه المشكلة تعود إلى تقدم الطالب في التحصيل من صف إلى آخر؛ لذلك فإن أيّ رُتب مئينية يتم الحصول عليها من نتائج هذا النوع من الاختبارات لا يمكن عدّها متكافئة، على الرغم من إمكانية التعامل مع هذه المشكلة من خلال وجود اختبار مشترك anchor test ، ويرتبط بهذه المشكلة عدة عوامل تجعل عملية

معادلة الاختبارات أكثر صعوبة. وأبرز هذه العوامل هو إعداد جدول درجات متكافئة، وحجم العينة المسحوبة، وانخفاض ثبات درجات هذه الاختبارات، والطريقة المستخدمة في عملية المعادلة.

- **أخطاء معادلة الاختبار**: في الكثير من المواضع، من المهم مقارنة الطرق المستخدمة في عملية معادلة الاختبار، للتعرف على أيها يعطي المقدار الأقل من خطأ المعادلة؛ لغرض الحصول على معادلة اختبارات عالية الثبات. فعلى سبيل المثال: أجرى جفني وميلامد (Gafni & Melamed, 1991) دراسة في معادلة الاختبارات، واستخدما الطرق الخطية في عملية المعادلة؛ لتقدير مقدار خطأ المعادلة لاختبار واحد، والمقاييس المعيارية التي استخدمها الباحثان كانت متوسط الفرق بين الدرجة الحقيقية (الخام)، والدرجة المعادلة، وكذلك مربع متوسط الجذر root mean sqaure لذلك الفرق، ولم يجد الباحثان علاقة ارتباطية واضحة بين خطأ معادلة الاختبار، وبين حلقات التوصيل بين طرق عملية المعادلة. أما برينان وكولن (Brennan & Kolen, 1987) فيشيران إلى أن القضايا العملية الحساسة في عملية معادلة الاختبار ترتبط بالتعرف على المصادر المتعددة لخطأ معادلة الاختبار؛ ثم حساب مقداره، والتخلص منه، أو خفضه إلى الحد الأدنى. ويرى الباحثان أن القضايا الحساسة في معادلة الاختبار هي توصيف المحتوى، ومعادلة الاختبار في سياق درجات القطع، وإعادة معادلة الاختبار، وأثر خرق عنصر الأمن، والسرية في معادلة الاختبار. ويؤكدان أن هذه القضايا لا يمكن معالجتها دون التواصل الدقيق والفعال بين المتخصصين في القياس والتقويم التربويّ، والنفسيّ، وخاصة فيما يتعلق بالحدود والقيود المرتبطة بأخطاء عملية المعادلة تلك.

- **حجم العينة ومعادلة الاختبار**: هناك بعض العوامل التي تؤثر في دقة معادلة الاختبار، مثل خطأ المعاينة، وخصائص العينة، وخصائص فقرات الاختبار المشترك، وحجم العينة. وقد وجد الباحثان كوك وبيترسون (Cook & Peterson, 1987) أن المشكلة الشائعة بين برامج الاختبارات التي تطبق لأغراض الترخيص لمزاولة مهنةٍ ما هي ندرة البيانات الضرورية لعملية المعادلة. حيث إن صغر حجم العينة يفاقم من هذه المشكلة إذا كانت هذه الاختبارات طويلة. والمشكلة الأخرى المرتبطة بذلك أن بعض برامج الاختبارات تطبق اختباراتها في أوقات متعددة خلال السنة. ولم تشر الأبحاث التي أجريت حول الموضوع، كما أشار كوك، وبيترسون، إلى مدى مصداقية، وثبات نتائج عملية المعادلة

إذا أعيدت معادلة صور هذه الاختبارات، باستخدام عينات من تطبيقات متعددة لهذه الاختبارات. وتكون هذه المشكلة أكثر تأثيراً في الاختبارات التحصيلية؛ لأنها تقيس مهارات وقدرات متعددة للطلبة الذين يتقدمون لهذه الاختبارات في أوقات مختلفة.

- **تأثير قدرة المفحوص في دقة معادلة الاختبار**: البحوث التي أجريت على تطبيقات طرق النظرية الحديثة في القياس التربويّ والنفسيّ على معادلة الاختبار رأسياً كشفت عن نتائج متضاربة فيما يتعلق بدرجة اللاتباين invariance التي تنتج عن استخدام هذه الطرق بالنسبة لقدرة المفحوص، أو السِّمة المقاسة. وقد قام سكاجز وليستز (Skaggs & Lissitz, 1988) بإجراء دراسة لمقارنة عدة طرق لمعادلة الاختبار، وهي طريقة الرُّتب المئينية، وطريقة راش، وطريقة IRT ثلاثية المعلم، لفحص اللاتباين لعملية المعادلة؛ وذلك عن طريق المعادلة الرأسية لاختبارين تحت ظروف مختلفة. وجد الباحثان أن تعددية البُعْد (السِّمة) multidimensionality قد تكون السبب وراء اختفاء اللاتباين. واستخلص الباحثان أن مسالة تعددية السِّمة هي مسألة درجة degree. وفي الواقع من الصعب معرفة عند أية نقطة يمكن استبعاد عملية المعادلة عندما لا يقيس الاختباران السِّمة نفسها تماماً. ويرى برينان وكولن (Brennan & Kolen, 1987) أن مسألة اللاتباين هي قضية مرتبطة بمجموعة جزئية من المفحوصين مستقاة من المجتمع الأصلي للدراسة.
- **أثر طول الاختبارين، واختلافهما في المحتوى على معادلة الاختبار**: هذه المشكلة تصدى لها بيترسون وكوك وستوكنج (Peterson, Cook, & Stocking, 1983)، وربطها أنجوف (Angoff, 1980) بقضية وجود مجموعتين لمعادلة الاختبار لم يتم اختيارهما عشوائياً. فقد وجد بيترسون وزملاؤه أن الاختبارات التي تختلف قليلاً في الطول (عدد الفقرات)، وفي المحتوى، فإن استخدام نماذج IRT السَّوْقيَّة logistic ثلاثية الأبعاد تؤدي إلى استقرار في نتائج معادلة الاختبار. ويشترط هنا مراعاة بعض العوامل، مثل: درجة التطابق بين النموذج المستخدم، والبيانات المستخدمة في عملية المعادلة، والتذبذب العشوائي في البيانات، وعدد فقرات الربط linking items؛ لوضع تقديرات معلمات الفقرات على التدريج نفسه. وإذا استطاع الباحث الحصول على مجموعتين من المفحوصين متكافئتين عشوائيا، فإن استخدام اختبار مشترك بينهما يؤدي إلى نتائج دقيقة في معادلة الاختبار.

- **فاعلية الطرق المستخدمة قبل معادلة الاختبار**: يشير أنجوف (Angoff, 1980) إلى أن عملية معادلة الاختبار قبل تطبيقه (معادلة الصورة القديمة مع الصورة الجديدة للاختبار) تسمح بالحصول على مجموعة كبيرة من فقرات الاختبار المعايرة. وضمن السياق نفسه، أجرى كولن وهاريس (Kolen & Harris, 1990) دراسة لمعادلة الاختبار واستخدما تصميمين لهذا الغرض، هما تصميم المجموعات العشوائية، وتصميم ما قبل المعادلة للفقرات؛ وذلك في نموذج الرُّتب المئينية، ونموذج IRT السوقيّ ثلاثي الأبعاد. وجد الباحثان أن عملية المعادلة قبل تطبيق الاختبار تعطي نتائج غير ذات فائدة عملية. ومن بعض المشكلات الفنية التي أشار إليها الباحثان في هذه القضية، والتي تؤثر سلباً في عملية معادلة الاختبار، هي آثار قياس أكثر من سمة، وآثار الفروق بين المجموعات في نتائج معادلة الاختبار قبل تطبيقه، وآثار السياق الذي طُبق فيه الاختبار.

خلاصـــــة

حاول الباحث في هذه الدراسة تسليط الضوء على موضوع مهم في حقل القياس والتقويم التربويّ والنفسيّ، ألا وهو معادلة الاختبارات؛ إذ تم عرض ومناقشة مفهوم عملية المعادلة وشروطها، وأنواع المعادلة، والتصميمات الشائعة في معادلة الاختبار، وطرق معادلة الاختبار في النظرية التقليدية، وفي النظرية الحديثة في القياس والتقويم التربويّ والنفسيّ. ثم تعرض بشئٍ من التفصيل والتحليل لبعض الدراسات السابقة التي عالجت موضوع معادلة الاختبارات في النظريتين، وقارنت بينهما، كما تم عرض أهم المشكلات الفنية والتطبيقية التي تعترض عملية معادلة الاختبار، والإشارة إلى آثارها السلبية على دقة عملية المعادلة، وعلى زيادة مقدار خطأ معادلة الاختبار.

ومن الجدير بالذكر أن هناك بعض القضايا التي عرضت لها الأدبيات، والدراسات التي عالجت موضوع معادلة الاختبار، ومشكلاتها؛ ولكنها لم تعالجها بعمق، ومن مختلف الزوايا، والظروف المحيطة بتطبيق الاختبارات موضع المعادلة وشروطها. وقد استخلصنا بعض النقاط الضرورية التي تحتاج إلى معالجة وعمق أكثر؛ والتي يمكن عرضها على النحو التالي:

١. أشارت الدراسات إلى موضوع صغر حجم عينة معادلة الاختبار، وأثرها في دقة عملية المعادلة ونتائجها؛ ولكنها لم توضح ما إذا كان ذلك ينطبق على معادلة الاختبار أفقياً أو رأسياً أو الاثنين معاً.

٢. لم تشر الدراسات إلى الحد الأدنى من عينة المفحوصين، ومن عينة فقرات الاختبار في صورتيه للشروع في عملية المعادلة، وإتمامها فنياً وتطبيقياً.

٣. الكثير من الدراسات في معادلة الاختبار لم تقارن تأثير بعض العوامل في عملية المعادلة الأفقية والرأسية؛ كعامل التعب، وأثر التعلم السابق، والخبرة.

٤. هناك حاجة إلى مزيد من الدراسات، والبحوث لمعرفة تأثير نوع الاختبار المستخدم (اختبار تحصيلي، أو بطارية اختبارات) في عملية المعادلة، وفي نتائج المعادلة، ودقتها، وأثره في خفض مقدار خطأ المعادلة.

٥. هناك حاجة لمزيد من الدراسات، والبحوث لمعرفة المسار، والصيغة التي تأخذها قضية تعددية الأبعاد المشار إليها سلفا، واختلاف ذلك في النظرية التقليدية، والنظرية الحديثة في القياس التربويّ والنفسيّ، في أثناء عملية معادلة الاختبار.

٦. هناك مشكلة في إمكانية مطابقة عينات المفحوصين لعملية المعادلة في الاختبار المشترك من خلال درجاتهم الحقيقية، و لم يتم التوصل إلى حل جذري للمشكلة؛ بل تم اللجوء إلى عملية التوفيق الخطي linear combination بين درجات الطلبة، كما أشار إلى ذلك ليفنجستون ودورانز ورايت (Livingston, Dorans, & Wright, 1990) وهذه القضية، في تقديرنا، تحتاج إلى مزيدٍ من البحث والدراسة، ليس للتعرف على عملية المعادلة فقط، بل لمعرفة تأثيرها في دقة المعادلة من اختبار، أو برنامج اختبارات إلى آخر من جهة، ولمقارنة نتائج عملية المعادلة من جهةٍ أخرى إذا تم استخدام عملية المعادلة الأفقية والرأسية.

REFERENCES

Allen, M.J., & Yen, W.M. (1979). ***Introduction to measurement theory***. Monterey, CA: Brooks / Cole.

Angoff, W.H. (1967). Technical problems of obtaining equivalent scores on tests. In W.A. Mehrens, R.L. Ebel (Eds.), ***Principles of educational and psychological measurement*** (PP. 84-87). Chicago: Rand McNally.

Angoff, W.H. (1980). Technical and practical issues in equating: A discussion of four papers. ***Applied Psychological Measurement*** , 11(3), 291-300.

Baker, F.B. (1985). ***The basics of item response theory***. Portsmouth, NH: Heinemann.

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.

Brennan, R.L., & Koen, M.J. (1987). Some practical issues in equating. ***Applied Psychological Measurement , 11***(3), 279-290.

Cook, L.L., & Eignor, D.R. (1991). IRT equating methods. ***Instructional topics in educational measurement series*** (PP. 191-200). Washington, D.C.: NCME.

Cook, L.L., & Peterson, N.S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. ***Applied Psychological Measurement , 11***(3), 225-244.

Crocker, L, & Algina, J. (1986). ***Introduction to classical and modern test theory***. , New York: Holt, Rinehart, & Winston.

Dorans, N.J. (1990). Equating methods and sampling designs. ***Applied Measurement in Education*** **, 3**(1), 3-18.

Gafni, N., & Melamed, E. (1990). Using the circular equating paradigm for comparison of linear equating models. ***Applied Psychological Measurement*** , **14**(3), 247-256.

Glowacki, M.L. (1990). ***An analysis of test equating models for the Alabama high school graduation examination***. Paper presented at the annual meeting of the Mid-South Educational Research Association, Lexington. (ERIC Documents reproduction Service number ED 340720).

Hambleton, R.k., & Swaminathan, H. (1985). ***Item response theory: Principles and applications.*** Boston: Kluwer Nijhoff.

Hambleton, R.k., Swaminathan, H., & Rogers, J. (1991). ***Fundamentals of item response theory.*** Thousand Oaks: Sage.

Hills, J.R., Subhiyah, R.G., & Hirsch, T.M. (1988). Equating minimum competency tests: Comparison of Instructional methods. ***Journal of Educational Measurement*** , **25**(1), 221-232.

Keeves, J.P. (1988). Scaling achievement test scores. In J.P. Keeves (Ed.), ***Educational research, methodology, and measurement: An international handbook*** (pp. 403-419). Oxford: Pergamon.

Kenyon, D.M., & Stansfield, C.W. (1992). ***Extending a scale of language proficiency calibration and rasch model.*** Washington, D.C.: Center for applied linguistics. (ERIC Documents Reproduction service number ED 343443).

Kolen, M.J. (1988). Traditional equating methodology. ***Instructional topics in educational measurement series*** (pp. 115-122). Washington, D.C.: NCME.

_____, M.J., & Brennan, R.L. (1995). Test equating: ***Methods and practices***. New York: Springer.

_____, M.J., & Harris, D.J. (1990). Comparison of item preequating and random equating using IRT and equipercentile methods. ***Journal of Educational Measurement , 27***(1), 27-40.

Kromrey, J.D., Parshall, C.G., & Yi, O. (1998). ***The effects of content representativeness and differential weighting on test equating: A monte carlo study.*** Paper presented at the annual meeting of the American Educational Research Association, San Diego. (ERIC Documents reproduction Service number ED 421536).

Lennon, R.T. (1967). Equating non-parallel tests. In W.A. Mehrens, R.L. Ebel (Eds.), ***Principles of educational and psychological measurement*** (pp. 87-91). Chicago: Rand McNally.

Lindquist, E.F. (1967). Equating scores on non-parallel tests. In W.A. Mehrens, R.L. Ebel (Eds.), ***Principles of educational and psychological measurement*** (pp. 79-84). Chicago: Rand McNally

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. (ERIC Documents Reproduction Service Number ED 431800).

Livingston, S.A., Dorans, N.J., & Wright, N.K. (1990). What combination of sampling and equating methods works best? ***Applied Measurement in Education , 3***(1), 73-96.

Lord, F.M. (1980). ***Applications of item response theory to practical testing problems.*** Hillsdale, NJ: Lawrence Erlbaum

Lunz, M.E., & Bergstrom, B.A. (1995). ***Equating computerized adaptive testing certification examinations: The Board of Registry series of studies.*** Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco. (ERIC Documents Reproduction Service Number ED 388696).

Morrison, C.A., & Fitzpatrick, S.J. (1992). ***Direct and indirect equating: A comparison of four methods using the Rasch model.*** Austin, TX: Texas University Measurement and Evaluation Center. (ERIC Documents Reproduction Service Number ED 3375152).

Parshall, C.G., et. al . (1991). S***tatistical errors in linear equating with small samples of examinees.*** Paper presented at the annual meeting of the Florida Educational Research Association, Clearwater. (ERIC Documents Reproduction Service Number ED 339745).

Peterson, N.S., Cook, L.L., & Stocking, M.L. (1983). IRT versus conventional equating methods: A comparison study of scale reliability. ***Journal of Educational Statistics , 8***(2), 137-156.

Peterson, N.S., Kolen, M.J., & Hoover, H.D. (1989). Scaling, norming, and equating. In R.L. Linn (Ed.), ***Educational measurement*** (pp. 221-262). New York: American Council of Education and Macmillan.

Skaggs, G., & Lissitz, R.W. (1988). Effects of examinee ability on test equating invariance. ***Applied Psychological Measurement , 12***(1), 69-82.

Smith, R.M, & Gross, L.J. (1997). Validating standard settting with a modified Nedelsky procedure through common item test equating. ***Journal of Outcome Measurement , 1***(2), 164-172.

Smith, R.M, & Kramer, G.A. (1992). A comparison of two methods of test equating in the Rasch model. ***Educational and Psychological Measurement , 52***(4), 835-846.

Staples, J.G., & Luzzo, D.A. (1999). ***Measurement comparability of paper-and- pencil and multimedia vocational assessments.*** Iowa City: American College Testing program. (ERIC Documents Reproduction Service Number ED 429091).

Suen, H.K. (1990). ***Principles of test theories.*** Hillsdale, NJ: Lawrence Erlbaum.

Wright, N.K., & Dorans, N.J. (1993). ***Using the selection variable for matching or equating.*** Princeton: ETS. (ERIC Documents reproduction Service Number ED 385547).

Yang, W.L. (1997). T***he effects of content mix and equating method on the accuracy of test equating using anchor-item design.*** Paper presented at the annual meeting of the American Educational Research Association, Chicago. (ERIC Documents reproduction Service Number ED 409334).

Yang, W.L., & Houang, R.T. (1996). ***The effects of anchor length and equating method on the accuracy of test equating: Comparisons of linear and IRT-based equating using an anchor-item design.*** Paper presented at the annual meeting of the American Educational Research Association, New York. (ERIC Documents reproduction Service Number ED 401308).