7

تكييف اختبارات الإنجاز والأهلية
مراجعة موضوعات منهجية

ليندا ل. كوك
اليسيا. شميدت - كاسكالار
الجمعية الدولية للتقويم، بروكسل، بلجيكا

كاثرين براون
خدمات الاختبارات التربوية

كان الاهتمام بتكييف الاختبارات التي تم إنجازها لاستخدامها للغة وثقافة إلى استخدامها مع مجموعة ذات لغة وثقافة مختلفة سائداً بين الباحثين التربويين والنفسيين في القرن العشرين. على سبيل المثال، درس هاميلتون وبولورك (1991) الترجمات الأولى لاختبار "بينت - سيمون" للذكاء. أشاروا إلى أن الاختبار قد ترجم من الفرنسية إلى الإنجليزية في 1911 واستخدم لتقويم ذكاء الدارسين في دار المعلمين في نيوجيرسي (الولايات المتحدة). أشاروا أنه حتى عام 1916 كان قد تم ترجمة الاختبار إلى سبع لغات مختلفة (ستانلي وهوبكنز، 1972). كما أشاروا إلى أن اختبارات ذكاء أخرى ومقاييس مرافقة تم ترجمتها إلى لغة الطلاب الأولى.

أشارت مطبوعات فان دي فيفر (2002) فان دي فيفر ولونر (1995)، فان دي فيفر ولونغ (1997، 2000) إلى أنه تضاعفت حديثاً مطبوعات تعالج الاختلافات

والمقارنات عبر الثقافات، وعزوا ازدياد تلك المطبوعات إلى عدة أسباب مثل العالمية في الاقتصاد، السياحة، الهجرة المستمرة وتغيرات سياسية.

إن الغرض من هذا الفصل هو مراجعة الموضوعات المنهجية التي ترافق تكييف مقاييس الإنجاز والأهلية. يتكون هذا الفصل من ستة أقسام. يقدم القسم الأول نظرة عامة عن موضوعات التحيز والتكافؤ وعلاقتهما مع مقارنة تقويم النتائج عبر اللغات/ الثقافات. يتبع تلك النظرة العامة أقسام متفرقة تركز على موضوعات متعلقة مثل: (أ) تكافؤ البنية، (ب) تكافؤ وحدات القياس، (ج) ترجمة الاختبارات ومواد الاختبارات، (د) تفسير الدرجات واستخدام الاختبار (د) إدارة الاختبار.

نظرة عامة على الموضوعات الرئيسة

أشار فان دي فيفر ولونغ (1997) أن "الأفكار المتكررة عن التقويم في اللغات المتعددة هو التساؤل عن مدى إمكانية استخدام الوسائل التربوية المطورة في الدول الغربية من محيط ثقافات مختلفة (صفحة 61). وضعوا لائحة تتضمن أربعة من تساؤلات المهتمين بتقويم الثقافات المتعددة:

- هل يوفر الاختبار تغطية كاملة للبنية النفسية نفسها للمجموعات الثقافية التي تستخدمه؟
- هل من الممكن استخدام قواعد إدارية واحدة أو يجب تكييفها؟
- كيف تستطيع التعامل مع المتغيرات الكثيرة في مهارات اللغة الأم واللغة المضيفة لمجموعات مهاجرة؟
- هل يوجد بدائل أخرى عندما تبين أن الاختبارات الغربية غير مناسبة؟

إن تلك الأسئلة الأربعة لها علاقة بمواقف تحصل عند تكييف اختبارات الأهلية والإنجاز لاستخدامها عبر الثقافات/ اللغات المختلفة.

إن معظم المطبوعات عن التقويم عبر الثقافات خصصت لاختبار وتقويم مقارنة بين الاختبار عبر ثقافات ولغات مختلفة.

إن التساؤلات عن تأثير الإجراءات الإدارية بالإضافة إلى الاختلافات اللغوية مهمة بشكل خاص لتكييف اختبار الأهلية والإنجاز ملائم ولإثبات صدق مقارنة درجات الاختبار عبر الثقافات.

وضح فان دي فيفر وتانزر (1997) ثلاثة من أنواع التحيز التي يمكن أن تؤثر في التقويم عبر الثقافات، وعبر اللغات، التحيز في البنية، التحيز في المنهج والتحيز في بنود الاختبار. وقد ناقشوا أيضاً تأثير نوعين من التكافؤ على الأبحاث والتقويم عبر الثقافات: تكافؤ البنية وتكافؤ وحدة القياس.

إن النقطة التي أثارها فان دي فيفر وتانزر هي أن السؤال الأساسي في كل الدراسات عبر الثقافات هو عما إذا كانت الدرجات الموجودة التي أعطيت لمجموعات مختلفة يمكن تفسيرها بطريقة واحدة. وقد أكدوا أن موضوع التحيز والتكافؤ أساسيان في تلك النقطة.

عرَّف فان دي فيفر وتانزر (1997) تحيز البنية بأنها تحدث إذا كانت البنية المقاسة غير متماثلة عبر المجموعات الثقافية، وقد استعملوا كمثال اختبارات الذكاء الغربية. اعتبر هؤلاء الكتاب بشكل أساسي أن تحيز البنية ليس مصطلحاً ينطبق على أداة ما ولكنه ينطبق على خصائص مقارنة عبر الثقافات، أشاروا إلى أن الأداة التي تكشف التحيز في مقارنة موضوعات بين اليابانية والألمانية قد لا تكشف التحيز في مقارنة بين الألمانية والهولندية.

حسب فان دي فيفر وتاتزر (1997) فإن التحيز في العينات ينتج عن اختلافات في العينات المتعلقة بالمتغيرات (غير متغيرات الاهتمام) التي يمكن أن تؤثر في نتائج البحث. وقد أعطوا مثال المجموعات الثقافية التي تختلف في الخلفية الثقافية والحوافز. إذا لم تكن أي منهما واحدة في متغيرات الاهتمام فإنهم بالتأكيد

سيحدثون ارتفاعاً في مقارنة تلك المتغيرات التي هي مركز الاهتمام في تلك الدراسة.

استخدم فان دي فيفر وتانزر (1997) في مناقشة تحيز الأداة "الإثارة المألوفة" كمثال معروف جداً. وقد توسعوا في ذلك المثال بالاستشهاد بأعمال هوي وتريانديس (1989) اللذين وجدا أن الأمريكيين من أصول لاتينية يميلون إلى اختيار الحد الأقصى في مدرج قياس الشخصية ذي خمس درجات بنسبة أكبر من الأمريكيين البيض. اكتشف هوي وتريانديس أن الارتباك الحاصل بسبب تفضيل الخيار قد اختفى عندما استعمل مقياس ذو عشر درجات.

إن النوع الثالث لتحيز المنهج الذي بحثه فان دي فيفر وتانزر (1997) هو التحيز الإداري. اعتبر بعض الكتاب أيضاً موضوع التحيز الإداري كتهديد لصدق تفسير درجات الاختبارات المكيفة. على سبيل المثال ناقش كيسنجر (1994) حقيقة أن الاختلافات الثقافية لمجموعات قومية تتغير حسب مستوى معرفتهم ببنود بنية الاختبار المختلفة واقترح استخدام عدد كاف من التمارين العملية لجعل البنية الجديد مألوفاً إلى الطلاب الممتحنين.

أخيراً، بالإضافة إلى مناقشة تحيز الأسلوب والبنية فقد ناقش فان دي فيفر وتانزر (1997) تحيز البند كنوع مهم الذي يسبب الإرباك في دراسات عبر الثقافات. وقد وضعوا لائحة بالأسباب التالية لتحيز البند في تلك الدراسات: "ترجمة سيئة للبند" بنود غامضة، عوامل مزعجة (قد يسبب البند آثاراً إضافية أو قدرات)، مواصفات ثقافية (اختلافات ثانوية في معنى المفاهيم/ ملاءمة محتوى البند) (صفحة 268).

من غير الضروري القول إن موضوعات تحيز البند تُعد تحدياً هائلاً إلى صدق أي مقارنة أهلية عبر الثقافات أو درجات اختبار إنجازات. إن الموضوعات المتعلقة بتحيز البند، مثل ترجمة الاختبار والإجراءات المتبعة في كشف التفاوت الوظيفي

للبند (DIF) قد نوقشت في الفصل الرابع في هذا الكتاب؛ ولذلك لن تؤخذ بعين الاعتبار أبعد من ذلك.

في مناقشة التكافؤ في دراسات عبر الثقافات ميز فان دي فيفر وتانزر (1997) بين التكافؤ في البنية والتكافؤ بين وحدات القياس. كان تعريفهم لتكافؤ البنية أن البنية تقاس عبر كل الثقافات للمجموعات المعينة بصرف النظر عما إذا كانت أدوات القياس المستخدمة متماثلة. أضافوا أنه من الممكن للأداة نفسها قياس بنيات مختلفة في ثقافات مختلفة، أو أن البنية المقاسة بالأداة نفسها قد تتشابه جزئياً بين الثقافتين.

عرف فان دي فيفر وتانزر (1997) نوعاً ثانياً من التكافؤ له ذات أهمية تكافؤ البند، تكافؤ وحدة القياس يحصل ذلك الموقف إذا كان القياسان لهم ذات وحدة القياس مع اختلاف المصدر. وقد وضحوا ذلك التعريف بالقول: "بكلمات أخرى إن مقياس القياس الأول قد تم تحويله بتساوٍ مستمر بالمقارنة مع المقياس الآخر" (صفحة 266). وقد أشاروا إلى أن درجات المقاييس التي لها تلك المواصفات لا يمكن مقارنتها مباشرة، ولكن إذا عرف عامل التساوي فإنه من الممكن تعديل الدرجات لجعلها مناسبة للمقارنة.

أخيراً، إذا أراد أحد ما أن يقوم بمقارنة جيدة لدرجات اختبارات الإنجاز أو المهارات، يجب عليه القيام بما أشار إليه فان دي فيفر وتانزر (1997) وهو مقارنة مدرجة، وقد جرى تعريفها بأنها المستوى الأعلى من مستوى تكافؤ وحدة القياس لاثنين من المقاييس. أشاروا أن ذلك النوع من التكافؤ يمكن الحصول عليه عندما يكون المقياسان لهما ذات المصدر ووحدة القياس. أضافوا بالقول إن التكافؤ المدرج هو مطلب أساسي للقيام بمقارنة تقويم النتائج عبر التقانات. وقد أكدوا على أن أي شكل من التحيز، الأسلوب، البنية إلى ما هنالك سوف يؤدي إلى اختلال وإضعاف تكافؤ وحدات القياس.

تكافؤ البنية

يقدم هذا القسم من الفصل مناقشة أهمية تكافؤ البنية في الدراسات عبر الثقافات كما يقدم نظرة عامة لإجراءات منتقاة تستخدم لتقويم تكافؤ البنية عبر مجموعات مختلفة.

أهمية تكافؤ البنية

إن كثيراً من الدراسات النظرية والعملية سائدة في الأعمال المنشورة حول التقويم عبر الثقافات. كثير من الكتاب مثل بورتينغا (1989، 1983)، فان دي فيفر ولونغ (1997)، فان دي فيفر وتانزر (1997) قد طورا تعريفاً بليغاً ومناقشات لتكافؤ البنية.

على ذلك الخط ناقش بورتينغا (1989) ما أشار إليه "منطق المقارنة". وقد أصر على أن المقارنة بين الأفراد، أو المجموعات يمكن أن تكون مضللة لسببين. السبب الأول هو أن الموقف الذي استخدم في المقارنة قد لا يكون متشابهاً عبر الأفراد أو المجموعات (تكافؤ البنية). وقد أعطى مثال، المقارنة بين الطول والوزن، أي أنه لا معنى له أن يكون شخص أطول من شخص آخر كبير الوزن. أما السبب الثاني فهو أن وحدات القياس قد لا تكون متشابهة (تكافؤ وحدة القياس)، على سبيل المثال أن الطول المقاس بالبوصه لا يمكن مقارنته بالطول المقاس بالسنتيمتر.

توسع بورتينغا حول العلاقة بين تكافؤ البنية وتكافؤ وحدة القياس بالإشارة إلى أن إمكانية تشكيل مقياس قياس يمكن مقارنته لنسختي اختبار مختلفتين في المحتوى ضئيلة جداً، بكلمات أخرى من الواضح أن تكافؤ البنية شرط أساسي لحدوث تكافؤ القياس.

تحرى بورتينغا (1983) في ورقة بحث ثانية بتوسع العلاقة بين تكافؤ البنية وأشكال أخرى من التكافؤ وقد ناقش مضمون تحليل يظهر الاختلافات في تقويم النتائج عبر المجموعات. أشار إلى أنه ترك للباحث تقرير ما إذا كان فقدان نتائج

قابلة للمقارنة ينتج عن بنية غير متكافئة، وحدات قياس غير متكافئة أو عن اختلافات حقيقية ضمن المجموعات.

استمر بورتينغا (1983) بالقول إن تحليل مقارنة نتائج الاختبار تؤدي غالباً إلى أحد حقول صحة البنية. وقد اعترف بنظرية أن تكافؤ البنية وصدق البنية كثيراً ما تتشابك إلى حد كبير. كانت النقطة التي أثارها رئيسة في تحليل تكافؤ البنية، كان السؤال الرئيس هو عما إذا جرى قياس البنية ذاتها، بينما في صدق البنية كان الموضوع الرئيس هو أي بنية قد جرى قياسها.

أكد بورتينغا على أهمية البحث المستند على النظريات. شدد على أن "تحليل مقارنة ذات معنى يحتاج إلى إطار عمل نظري كأساس يستطيع التعهد بصراحة أية علاقة بين أية متغيرات يمكن أن تكون ثابتة على المجموعات" (صفحة 246). أشار بورتينغا كمثال إلى كتابات فان دي فيفر ودرنث (1980). حسب فان دي فيفر ودرنث فإن مناقشة قياس الطول والوزن تقدم نتائج مقارنة عبر الثقافات حتى إذا كانت العلاقة بين تلك المتغيرات قد تختلف عبر المجموعات. على كل حال أشاروا إلى أنه إذا كان الطول والوزن سيستخدم لتخمين المتغير الثالث، على سبيل المثال محيط الخصر، عندئذ لكي يجري اعتبار لقياس متكافئ عبر المجموعات التي يجري دراستها، يجب أن يكون للطول، للوزن، لمحيط الخصر علاقة متماثلة عبر تلك المجموعات. كنتيجة لذلك فإن من الصعب تقويم التكافؤ أو صدق القياس عبر المجموعات التي يجري دراستها دون نظرية مدروسة عن العلاقة بين الطول، الوزن ومحيط الخصر.

عرف فان دي فيفر وتانزر (1997) تكافؤ البنية بأنه "البنية التي يجري قياسه عبر كل المجموعات الثقافية التي تُدرس بغض النظر عما إذا كان قياس البند يستند على أدوات متماثلة عبر الثقافات أم لا" (صفحة 265). أشاروا إلى أن تكافؤ البنية يمكن أن يحدث عندما تتشابك البنيات عبر الثقافات جزئياً فقط أو عندما تترافق

البنيات مع سلوكيات أو خصائص مختلفة كنتيجة لاختلاف الثقافات وقد قدموا ملخصاً عن مصادر تحيز البنية التالي (عدم وجود التكافؤ): "تشابك جزئي فقط في البنية عبر الثقافات؛ تفاوت في ملائمة السلوك المتعلق بالبنية (مهارات لا تنتمي إلى الذخيرة الخلفية لإحدى المجموعات الثقافية)؛ نماذج سيئة لكل السلوكيات المرافقة (وسائل إيضاح قصيرة)؛ تغطية غير كاملة لكل النواحي المتعلقة بالبنية (عدم وجود نماذج لكل الميادين التعليمية) (صفحة 268).

استطرد الكتاب بتفصيل الاستراتيجية القيمة التالية لتعريف موضوعات عن تقويم تحيز البنية (أو التكافؤ) عبر الثقافات:

- اللامركزية (مثال: تطوير الأدوات الايضاحية في عدة ثقافات في وقت واحد).
- طريقة المقاربة (مثال: استقلال تطور الأداة ضمن الثقافة وتطبيق كل الأدوات عبر الثقافات لاحقاً).
- استخدام خبراء لتقديم المعلومات اللغوية عن اللغة والثقافة المحلية.
- استخدام مسح محلي (مثال: تحليل محتوى الأسئلة ذات إجابة حرة).
- تطبيق أداة غير مألوفة (التفكير بصوت عال).
- مقارنة شبكات مرتبطة مع بعضها في علم المنطق (مثال، تقارب/ صحة تميز الأبحاث، أبحاث في طرق عديدة لها سمة وحيدة، مفاهيم عبارات رئيسة) (صفحة، 272).

فيما يلي مناقشة المزيد في الاستراتيجيات الشائعة لتقويم تكافؤ البنية التي جرى تعريفها سابقاً، مع نماذج منتقاة لتطبيقات ناجمة.

تقييم تكافؤ البنية

وصف هيوي وترياندس (1985) طرقاً عديدة استخدمت لبرهان تكافؤ البنية واعتبروا استخدام طرق متعددة أساسياً في تلك المسألة. بدؤوا بوصف قواعد

تراجعية كطريق لبرهان التكافؤ عبر الثقافات. أكدوا أنه من الدقة الاستنتاج أن التقييم المتكافئ عبر المجموعات يتعلق بطريقة مماثلة بمعايير خارجية. استطردوا بالقول إن هذه طريقة بسيطة واقتصادية لمسألة التكافؤ، لكنهم أشاروا إلى أن الاختلافات في متغيرات النماذج وجدارة الأدوات يمكن أن تسبب عدم استقرار تردي تقدير الدرجات الذي هو "في الأساس إنذار كاذب".

ناقش هيوي وترياندس (1985) أيضاً نظرية الإجابة للبند (IRT)، طرق لإقامة التكافؤ (انظير سيرغي والالوف، 2003). وقد بينوا أن الأداة ذات منحى مواصفات للبند (ICCs) عبر الثقافات قد برهنت، ولو بشكل جزئي، على تكافؤ بنودها (وكنتيجة لتكافؤ البنية) وتكافؤ المدرج" (صفحة 139).

ناقش آنوف وكوك استخدام نظرية الإجابة للبند (IRT) لتطوير سلسلة من بنود تكافؤ البنية التي تستخدم لربط اختبارات قدرات متطورة أعطيت لمتحنين ناطقين في اللغة الإنجليزية وآخرين ناطقين باللغة الإسبانية. استنتجوا أن طريقة (IRT) كانت فعالة عند الاستخدام لذلك الهدف. كتب بالتفصيل عن بحث آنجوف وكوك في الفصل السادس.

بالإضافة إلى نقاش حول طريقة تردي التقدير وطريقة (IRT)، ناقش هيوي وتياندس أيضاً تطابق البنية. أشاروا إلى حقيقة أنه إذا أُريد اعتبار البنية متكافئة عبر الثقافات يجب أن تكون البنيات الداخلية والعلاقة بينهم متشابهة عبر الثقافات والاهتمامات. اقترحوا عوامل تحليلية ومعايير متعددة الأبعاد لتقنية إحصائية يمكن أن تكون مفيدة للاستخدام لفهم كيفية عمل البنية عبر ثقافات مختلفة.

استخدم سيرغي، فتزجيرالد، وزينغ (1998) مجموعة مشتركة من تحليل قواعد العناصر، تحليل العناصر التأكيدية، معايير متعددة الأبعاد لتقييم تكافؤ البنية لاختبارات أجريت عبر الانترنت للمتحنين إنجليز، فرنسيين، ألمانيين ويابانيين في لغاتهم الأم أجرى الباحثون تحليل قواعد العناصر على مستوى المعطيات،

مستوى البند الواحد ومستوى مجموعة البنود. (ينصح استخدام مجموعة من البنود في عملية تحليل العناصر لتفادي صعوبات غير منطقية التي من الممكن حدوثها إذا كان الارتباط على مستوى معطيات البند الواحد). يجري تحليل المعايير المتعددة الأبعاد على معطيات مجموعة بنود فقط، أشار عدد من الباحثين إلى إمكانية استخدام إجراءات تحليل العناصر التأكيدية لتقديم تكافؤ البند غير المجموعات (انظر. غيرل، 2000) تقصى أفرسون، غيريرو وليتوسيس تكافؤ بنية SATI، اختبار الرياضيات (SAT-M) وقسم الرياضيات في اختبار الأهلية (PAA-M) الذي يجري لطلاب مدارس ذوي لغتين في بورتوريكو. استخدم الباحثون كلاً من تقنية تحليل العناصر التأكيدية وتحليل العناصر التمهيدية. استخدم أفرسون وزملاؤه مجموعة تقنيات لتطوير المعلومات على تحليل العناصر التمهيدية.

تكافؤ وحدات القياس

إن إقامة وحدة قياس مشتركة للدرجات التي تم الحصول عليها من اختبارات الإنجاز واختبارات الأهلية التي أعطيت بلغات مختلفة للممتحنين ذوي خلفيات ثقافية مختلفة عسيرة جداً لأسباب عديدة جرى ذكرها في أوائل هذا الفصل، إن السبب الذي يجعل ذلك مشكلة جديدة هو أن أحد المفاهيم المفترضة لأكثر الطرق المستخدمة في إقامة وحدة قياس مشتركة (طرق المقاييس المترابطة) هو أن التقويمات التي ستربط ببعضها تقيس ذات البنيان أو بنيات متشابهة، يحتوي الفصل السادس نظرة عامة ومناقشة إجراءات وموضوعات متعلقة بإقامة تكافؤ بنود القياس.

ترجمة الاختبارات ومواد الاختبارات

تستخدم دراسات مقارنة الإنجاز أو الأهلية في لغات مختلفة عبر الثقافات أدوات توضيحية جرى ترجمتها وتكييفها إلى اللغة والثقافة المستهدفة. لا تتضمن عملية تطوير أدوات متكافئة في أكثر من لغة ترجمة بنود الاختبار ومواد الاختبار

ولكنه يتضمن تغيرات في هيكلية البند وإجراءات الاختبار أيضاً (تكييف الاختبار). يجب أن نأخذ بعين الاعتبار الموضوعات المتعددة التي لها علاقة بترجمة الاختبار للحصول على أدوات مناسبة للمقارنة عبر الثقافات. "يجب أن لا تعكس الترجمة الجيدة معنى البند الأصلي، لكنها يجب أن تحاول الحفاظ على صلة الاهتمام العقلي والألفة مع محتوى البند وإلا قد يتغير ما يرغب البند في قياسه" (اريكان، 1999، صفحة 2). شكلت الهيئة الدولية للاختبارات عام 1992 مجلساً مكوناً من 13 عضواً لتطوير معيار فني/ تقني لتطوير الاختبار: قدمت الهيئة مجموعة من 22 خطاً من خطوط رئيسة لتكييف اختبارات ثقافية ونفسية (هامبلتون، 1994، انظر الفصل الأول من هذا الكتاب). تناقش سبعة من تلك المجموعة بشكل خاص موضوعات متعلقة بعملية الترجمة وتأثيرها على صلاحة الأدوات التي طورت لإقامة مقارنة الإنجاز والأهلية عبر اللغات/ الثقافات بغية تنظيم ملخص عن الموضوعات المتعلقة بترجمة الاختبار، يجري تقديم كل من تلك الخطوط العريضة مع تفسير ملخص ومراجع وثيقة الصلة بتلك الموضوعات.

1- يجب على مطوري /ناشري Developers/ Publishers الأدوات التأكد أن عملية التكيف تأخذ بعين الاعتبار الاختلافات اللغوية/ الثقافية للمجموعات المقصودة لاستخدم تلك النسخ المكيفة للأداة. إن وجود مترجمين ذوي خبرة وكفاءة مهم جداً لعملية ترجمة الاختبار؛ لأن عملهم له تأثير رئيس على صدق وجدارة درجات الاختبار. تؤكد تلك الخطة على ذلك لأن الخبرة باللغتين (أو أكثر) المعرفة والخبرة في الثقافات، مضمون الاختبار وقواعد القياس أساسية ويجب استخدام فريق من الخبراء. كان هناك خطأ عام وهو استخدام أشخاص متوفرين غير مؤهلين. وحتى عند استخدام شركات للترجمة وتحديد شروط العقد على أن خبراء ثنائيي اللغة ذوي معرفة بموضوعات الاختبار يجب أن يقوموا بعملية ترجمة البنود، لا تزال هناك فروقات في الترجمة. وصل "ريكارس وكنس" (1999)

إلى قرار وهو أن الحل لتلك الفروقات والحصول على بنود تقنية أكثر لاختبار مؤهلات يجب استخدام خبراء ثنائيي اللغة لمراجعة معاني بنود مضمون الاختبار" (صفحة 16) في دراستهم عن دقة ترجمة "الاختبارات التقنية المترجمة والمقدمة طلب "ريكاس وكنس" أن يكون المترجمون ناطقين باللغة المستهدفة. وقد أوصى باستخدام المترجمين الذين لهم سيطرة على اللغة المستهدفة "ودكوك" (1985) و"هامبلتون" (1993) وبذلك تصبح الترجمة طبيعية وفعالة. إن دراسة ثنائية اللغة قد أظهرت أنه من الأسهل لأشخاص يسيطرون على اللغة المستهدفة التعرف على كلمة في لغة المصدر وتذكر بشكل فعال الكلمة المرادفة في اللغة المستهدفة وليس العكس صحيحاً (بيريز، 1975).

2- يجب على مطوري/ ناشري الأدوات توفير الدليل على أن اللغة المستعملة في الإرشادات، التعليمات والبنود ذاتها وفي كتيب الاختيار مناسبة لكل المجموعات اللغوية والثقافية المقصودة بهذه الأدوات. يمكن أن يكون الاختبار في لغة المصدر معقداً دون أي ضرورة وهذا قد يجعل الترجمة الدقيقة صعبة أو أن يكون لها مفهوم، تعابير وأفكار ليس لها مماثل في اللغة المستهدفة (هامبلتون، 1993، اريكان، 1998، 2000). يجب أن يكون مستوى صعوبة المفردات، القراءة، استخدام القواعد، أسلوب الكتابة والتنقيط متشابه عبر اللغات. إحدى الطرق لتقليص الفروقات في صعوبة المفردات هو استخدام لوائح دائمة / متكررة من المفردات (هامبلتون وكانجي، 1993).

إن المشكلة ليست فقط في عدم وجود لوائح مفردات لغوية ولكن في حال وجودها لا تكون في اللغة المستهدفة بالتحديد. أما في حال الترجمة التقنية ينصح "باريكاس وكونس" (1999) باستخدام ملحق للمصطلحات

التقنية عندما لا يكون هنالك مفردات أو تعابير خاصة في اللغة المستهدفة فإنه من الضروري إضافة بنود ثقافية ذات أهمية محددة إلى أداة المصدر (اللاتركيزية). إن المرونة في تلك الإضافة أو تغير أداة المصدر لإقامة التكافؤ عبر اللغات ممكن عندما تكون أدوات اللغتين المصدر والمستهدفة قد جرى تطويرها في وقت متزامن (هامبلتون وكانجي، 1995).

3- يجب على المطورين/ الناشرين توفير الدليل على أن محتويات البند والمواد المحفزة مألوفة لكل المجموعات المقصودة. عندما يجري تطوير الأدوات مع توقع ترجمتها إلى لغات أخرى واستخدامه في ثقافة ثانية، فمن الضروري تجنب استخدام وحدات قياس، عملات نقود أو مواد محظورة (جداول، أرقام، أشكال) التي قد تؤثر بشكل متفاوت على أداء مجموعات مختلفة. يجب أن تؤخذ تلك المسببات الممكنة التي تساعد على أداء مناسب بعين الاعتبار في مرحلة تطوير الأداة. أوصى هامبلتون وكانجي (1995) بتجنب استخدام وحدات القياس، مثل البوصة أو القدم إلى ما هنالك، لأنها تختلف من بلد إلى آخر.

4- يجب على المطورين/ الناشرين تحقيق دليل عقلاني، لغوي ونفسي، لتحسين دقة عملية التكييف وجمع الأدلة عن تكافؤ ذلك في جميع النسخ المختلفة اللغات كما يجب استخدام الطرق العقلانية قبل استخدام الأداة وتقويمها بشكل إحصائي. إن أكثر الطرق العقلانية شيوعاً لإقامة تكافؤ الترجمة هي الترجمة المبكرة والترجمة الراجعة. إن خطة الترجمة المبكرة هي أن يقوم فريق واحد من المترجمين بترجمة الأداة من لغة المصدر إلى اللغة المستهدفة بينما يقوم فريق بآخر بمراجعة تكافؤ الترجمة مع لغة المصدر. أما الترجمة الراجعة فهي إعادة ترجمة الأداة إلى لغة المصدر من قبل مجموعات مختلفة من المترجمين وتقويم تكافؤها مع الأداة الأصلية من مثل

خبير أو أكثر. مع أن الترجمة الراجعة قد جرى استخدامها بشكل مكثف (برسلين 1970) فإن هامبلتون وباتسولا (1998) يؤكدون أن تلك الطريقة لا تقيّم تشابه بنية الاختبار المقاسة في اللغتين بشكل مباشر، "إن العالم غير المساعد الرئيس في تلك الخطة هو، تقييم تكافؤ الاختيار يجرى في لغة المصدر فقط" (هامبلتون وكانجي، 1995، صفحة 151).

بالرغم من أن هذه الخطة توفر مراجعة أولية لتكافؤ الترجمة، "فإن هناك دليلاً قليلاً لإقامة الدليل على أن المترجمين والخبراء لهم المقدرة على تنبؤ تكافؤ النسخ المتعددة للأداة من مراجعة عامة حتى لو جرت بعناية (هامبلتون وكانجي، 1993، صفحة 15).

5- يجب على المطورين / الناشرين التأكد أن جمع المعطيات يسمح باستخدام تقنيات إحصائية مناسبة لإقامة تكافؤ البند في أداة النسخ المتعددة اللغات، يجب توفر مقدار كاف من النماذج في كل من اللغتين، المصدر والمستهدفة، لإقامة تقنية إحصائية (مقياس متعدد الأبعاد، تحليل العوامل) مفيدة وقد وصف هامبلتون (1993) ثلاثة خطط لجمع المعطيات:

- يجري طلاب ثنائيو اللغة اختباراً باللغتين (المصدر والمستهدفة).
- يجري طلاب أحاديو اللغة الاختبار الأصلي واختبار الترجمة الراجعة.
- يجري طلاب لغة المصدر اختبار في لغة المصدر ويجري طلاب اللغة المستهدفة اختبار في اللغة المستهدفة.

أشار هامبلتون (1993) إلى أن الخطط التي تحتاج إلى أشخاص ثنائيي اللغة صعبة التحقيق: لأنه ليس سهلاً إيجاد أشخاص ثنائيي اللغة بارعين في كلتا اللغتين. وأضاف أن الدليل الذي تم جمعه باستخدام نماذج من ثنائي اللغة لا يمكن تعميمه على مجموعة أحاديي اللغة المقصودة. وقد

انتقد خططاً تتضمن أشخاصاً أحاديي اللغة يقومون بنسخ أداة اختبار في لغة المصدر والترجمة الراجعة. أشار إلى أن أحد عيوب ذلك النوع من المقارنة هو أن أداة المصدر وأداة الترجمة الراجعة قد تظهر متشابهة بالرغم من الترجمة السيئة. من الممكن حدوث ذلك إذا استخدم المترجمون مجموعة مشتركة من قواعد الترجمة أو إذا احتفظت الترجمة بنواح غير مناسبة من لغة المصدر مثل بنية القواعد الواحدة إلى ما هنالك. فصَّل هامبلتون الخطة التي يجري فيها طلاب مجموعة أحاديي اللغة اختبار أداة لغة المصدر وطلاب مجموعة أخرى أحاديي اللغة اختبار اللغة المستهدفة ويتم "ترابط" الاختبارين بواساطة مجموعة من البنود. إن ميزة تلك الخطة هي أن نماذج المجموعتين، المصدر والمستهدفة، تستخدم في التحليل وأن النتائج تعمم غالباً في كل المجموعات المشاركة.

6- يجب على المطورين/ الناشرين استعمال تقنيات إحصائية مناسبة كي (أ) إقامة تكافؤ النسخة المختلفة للأداة و(ب) التعرف على صعوبات في عناصر الأداة والتي يمكن أن تكون غير ملائمة لواحدة أو أكثر في المجموعات المقصودة. كإضافة إلى التقنية العقلانية يمكن استخدام الطرق الإحصائية للتأكد من ملاءمة ترجمة الاختبار. نصح هامبلتون (1993) بإجراء دراسة عناصر بنية نسخ اختبارية في لغات متعددة كطريقة نافعة للتقييم عما إذا كانت ترجمة الاختبار من لغة المصدر إلى اللغة المستهدفة جيدة إلى حد كافٍ.

7- يجب على المطورين/ الناشرين توفير دليل إحصائي على تكافؤ الأسئلة لكل المجموعات المقصودة. إن الوظيفة التفاوتية للبند وتحليل تخير البند يمكن استخدامها للتقييم إذا كانت وظيفة البند متساوية عبر المجموعات بعد آخذ عامل الاستطاعة بالحسبان. جرى تقديم هذه الطرق في الفصل الرابع.

تفسير الدرجات واستخدام الاختبار

يمكن اعتبار عملية اختبار الإنجاز والأهلية وسيلة لوصول هدف. يقدم الاختبار درجات يفسرها مستخدمون مختلفون لأغراض متنوعة. في مواقف الاختبار التربوي، يجري الظن أن الدرجات لها مدلول عن إمكانية الطلاب أو عن براعتهم في حقل معرفة خاص أو عن معلوماتهم. في دراسات عبر الثقافات، يوفر الاختبار أساساً للقيام بمقارنة بين لغات ومجموعات ثقافية مختلفة وبذلك يكون هناك إدراك أفضل للاختلافات والتماثل بين تلك المجموعات (هامبلتون وبورلورك، 1991). على كل حال فإن تفسير نتائج اختبارات الإنجاز والأهلية التي تجريها مجموعات يتكلمون لغات مختلفة ليست مهمة واضحة لمطورين ومستخدمين الاختبار لأن وجود اختلافات غير مقصودة في صعوبة ومحتويات الاختبار فقط تساهم في إحداث اختلاف الدرجات بين المجموعات أو الأفراد (سيرغي، 1997).

تصرف مطورو الاختبار في السابق وكأن العامل المهم الوحيد في تكيف اختبار لثقافة مختلفة هو ترجمة اللغة الأصلية المستخدمة في الاختبار إلى لغة جديدة. كما ذكر سابقاً في قسم آخر من هذا الفصل فإن ترجمة اختبار من لغة إلى أخرى لا يضمن المساواة في الدرجات عبر اللغات المستخدمة أو الثقافات (انجوف وكوك، 1997، كيسنجر، 1994، هامبلتون، 1993، بريتو، 1992، سيرغي، 1997).

وكما صرح هامبلتون (1994) فإن استخدام طريقة غير جدية تعالج تكييف الاختبار تقود إلى مفهوم خاطئ وهو أن اختلافات الدرجات بين نماذج أو مجموعات يمكن تفسيرها وكأنها حقيقة.

جرت الإشارة إلى أن ترجمة اللغة هو عامل واحد، بالرغم من أنه عامل مهم في تكييف الاختبار، ليس كافياً لوحده لجعل نتائج نسخ اختبار مختلفة اللغات قابلة للمقارنة. إن اختلافات خاصة في أعراف لغوية تمشي يداً بيد مع اختلافات محددة

في الأفكار والسلوك، وهذا يجعل من المستحيل فصل اللغة عن الثقافة. بذلك يثبت أن اعتبار اللغة فقط ليس كافياً.

بالإضافة إلى عامل اللغة، هنالك عدة عوامل يجب أن تؤخذ بعين الاعتبار إذا أردنا أن تكون درجات الاختبار الذي جرى تكييفه للاستخدام في لغات وثقافات متعددة ذات مدلول في التفسير. تتضمن العوامل التي تؤثر على استطاعة مُستخدم الاختبار لاستنتاج تفسير جيد: شروط إدارة الاختبار، المنهاج، السياسة التربوية، دوافع الممتحنين، الحالة الاقتصادية، مستوى الحياة، القيم الثقافية، بنية بنود الاختبار غير المألوفة، قلق من الاختبار، وسرعة الاختبار (هامبلتون وكانجي، 1993؛ فان دي فيفر وبورتينغا، 1991).

وصف هامبلتون وكانجي (1993) بالتفصيل بعض العوامل المهمة التي يجب أن تؤخذ بعين الاعتبار عند تفسير نتائج اختبار الإنجاز في دراسات عبر الثقافات، التماثل في المناهج هو أحد العوامل الذي يجب الانتباه إليه بجدية. إن أي مقارنة على مستوى الإنجاز بين ثقافتين مختلفتين ستكون ضعيفة إلا إذا أخذت الاختلافات في المناهج بالحسبان. هنالك الكثير من الأمثلة عن تأثير اختلافات المناهج في المطبوعات. على سبيل المثال، في النظرة الأولى يبدو أن نتائج الدراسة الثانية للرياضيات العالمية (SIMs) تشير إلى أن أداء طلاب الولايات في بعض المستويات أقل من نظرائهم في اليابان في كل درجة أداء وكل أوجه الرياضيات، على كل عندما لوحظت الفروق في المناهج وضبطت لم تعد هناك أي فروق واضحة بين أداء طلاب الولايات المتحدة والطلاب في اليابان (وستبري، 1992).

إن تأثير حوافز الطلاب على درجات الاختبار شيء يجب التنبه له عند تفسير درجات النتائج في كل الاختبارات: على كل، إن موضوع الاختلافات الثقافية تعقد تأثير ذلك العامل بشكل أبعد. ألقى هامبلتون وكانجي (1993) الضوء على نتائج وينر (1993) التي تساءل فيه عما إذا كان يمكن فصل الخبرة الظاهرة المقاسة بالاختبار. بأي حال أشار وينر إلى دراسة تقويم التقدم التربوي العالمي (IAEP)

(لابوينت، ميد، واسكو، 1992) كدليل على ذلك. في تلك الدراسة كان أداء الطلاب الكوريين أعلى بكثير من نظرائهم الأميركيين، على كل تم إخبار الطلاب الكوريين بأنه تم اختيارهم للمشاركة في تلك الدراسة، كان ذلك شرفاً كبيراً لهم لمدارسهم ولبلدهم، وأن عليهم الأداء بأفضل ما يمكنهم. بالجانب الآخر لم يعط الطلاب الأميركيون أي تلميح لتحفيزهم؛ لذلك شاركوا في اختبار تلك الدراسة وكأنها نشاط مدرسي آخر.

إن فهم تأثير العالم السياسي/ الاجتماعي وجه مهم آخر في تفسير الدرجات لذلك يكون القيام بمقارنة الدرجات بين بلدان متطورة وبلدان غير متطورة ليس بالعملية السهلة الدقيقة. إنها تتطلب فهم المصادر المتوفرة والنوعيات المختلفة للخدمات التربوية التي يمكن أن تؤثر على أي قرارات عن القدرة الحقيقة التي تعكسها درجات الاختبار (هامبلتون وكانجي، 1993، اوليدو، 1981).

وضع هامبلتون وفان دي فيفر (1996) مع اللجنة الدولية للعلماء النفسيين وعلماء القياس السيكولوجي مجموعة من الخطوط العريضة، لتكييف اختبارات تربوية ونفسية. كان ضمنها بعض خطوط التوثيق وتفسير الدرجات المفيدة جداً للممارسين الذين يحاولون الاستفادة من تلك الدرجات.

فيما يلي الخطوط العريضة للتوثيق وتفسير الدرجات:

1- عندما يجري تكييف/ ترجمة الأداة للاستخدام في مجموعة أخرى، يجب توثيق التغييرات مع دليل التكافؤ. أصر فان دي فيفرو هامبلتون على أن فهم أي تغيرات جرت لتعزيز صدق الأداة المكيفة مهم لمستخدمي الاختبار عند تقرير عما إذا كانت أداة ما مناسبة لأغراضهم في البيئة الجديدة.

بالإضافة إلى المعلومات عن التغييرات التي جرت على الأداة، يجب أن يكون لدى المستخدمين إمكانية للحصول على معلومات عن تكافؤ نسخ الاختبار في لغة المصدر واللغة المستهدفة، مواصفات عملية الترجمة ونتائج تحليل تحيز البند أو تحليل العوامل. من المهم أيضاً لمستخدمين الاختبار معرفة عما إذا تم أخذ عوامل ثقافية معينة بالاعتبار عند وضع الاختبار.

2- يجب أن لا تُقوم الدرجات المختلفة للمجموعات التي أجرت اختبار الأداة حسب القيمة الظاهرية. تقوم على البحث مسؤولية إثبات الاختلافات بدلائل تجريبية. من الضروري الانتباه أن معنى الاختلافات بين مجموعات مختلفة يمكن تفسيره بطرق مختلفة. نبه فان دي فيفر وهامبلتون إلى أنه في حال عدم اختيار الباحث تقبل تفسير معين للدرجات، فيجب عليه توفير الدليل لدعم ذلك الخيار.

غالباً تتطلب مجموعة الدلائل مقاييس العوامل مختلفة الخيارات. إن الاختبار الذي جرى تكيفه حسب عملية تقنية موثوقة تتطلب جهداً أقل في دعم تفسير الدرجات لأن صدق الأداة جرى إثباتها إلى حد ما. أشار فان دي فيفر وهامبلتون إلى أنه حتى في أحسن الظروف فإن على الباحثين القيام بكل الجهود لإظهار تفسير دقيق لنتائج نسخ اختبار متعددة.

3- يمكن إقامة مقارنة للمجموعات المختلفة على المستوى الثابت الذي جرى إقامته للمقياس الذي بينته الدرجات. أشار فان دي فيفر وهامبلتون هنا إلى مفهوم قياس الدرجات. جرت مناقشة موضوعات متعلقة بإقامة تكافؤ قياس المقياس في هذا الفصل من قبل. من الممكن عند توفر عدد كبير من النماذج إقامة درجات نسخ اختبار للغات مختلفة على مقياس واحد لإجراء مقارنات للمفهوم. إن النقطة الأساسية لتلك الخطوط العريضة هي حث الباحثين للمرة الثانية على عدم القيام بمقارنة درجات نسخ مختلفة من الاختبار لا مبرر لها إلا إذا توفر دليل الصدق.

4- يجب على المطورين/ الناشرين توفير معلومات دقيقة عن الطرق التي يمكن أن تؤثر فيها الثقافة الاجتماعية والظروف البيئة للمجموعات على الأداء في الاختبار، كما يجب عليهم اقتراح إجراءات لبيان تلك التأثيرات على تفسير النتائج. شكل فان دي فيفر وهامبلتون في هذا الدليل من الخطوط العريضة طريقة عملية للتعامل مع العامل الثقافية/ الاجتماعية والظروف البيئية التي لها

دور في تفسير الدرجات. إن أفضل طريقة لنقل معلومات مناسبة لمستخدم الاختبار هي توفير كتيب للاختبار الذي يفصل كل المتغيرات التي بحثت عند تطوير الأداة (الصفات الثقافية)، الحالة الاجتماعية/ الاقتصادية، العمر، الجنس، التعليم للمجموعة المستهدفة.

إذا كان مثل تلك التحليلات متوفرة لمستخدمي الاختبار، فسيكون لديهم معلومات أكثر عن كيفية بيان أسباب تلك العوامل عند تفسير الدرجات (براكن وبارونا، 1991: فان دي فيفر وبورتينفا، 1991).

لكي تكون تفسيرات الدرجات ذات معنى، من الضروري إثبات أن القياسات المكيفة تُقوّم ذات البنية في اللغة أو الثقافة. جرى مناقشة موضوعات ذات علاقة بتقويم تكافؤ البنية أيضاً في أوائل هذا الفصل.

إن التحديد على أنه يجب على القياس المكيف تغطية ذات الأبعاد بذات النسب في اللغات والثقافات المختلفة للمجموعات ضروري جداً لعملية تفسير جيدة ولاستخدام درجات الاختبار (أيسينك وأيسينك، 1983). كما ذكر سابقاً، هناك تقنية واحدة لإقامة تحليل هذا العامل.

بإمكان معلومات معيارية أيضاً توفير معلومات مهمة لشخص محترف يحاول استخراج معنى من الدرجات وذلك بوضع شخص ما ضمن مجموعة من الذين يقوم بإجراء الاختبار. على كل حال من الضروري جداً إثبات أن المعايير المطورة لاختبار تم إجراؤه لمجموعة لغوية وثقافية مناسبة لتفسير درجات مجموعة لغوية وثقافية مختلفة (انظر كيسنجر، 1994، مناقشة موضوعات متعلقة باستخدام معلومات عن معايير للتقويم عبر الثقافات/ اللغات).

إدارة الاختبار

كما ذكر سابقاً في هذا القسم من الفصل، إن تحيز الطريقة مصطلح عام يمكن أن يشير إلى أي عامل تهديد للصدق المرتبطة بظروف إدارة الاختبار (فان دي فيفر وهامبلتون، 1996). إن عدم الخبرة في بنية البند، في هيكل الاختبار، أو في

وضع الاختبار، كل هذا قد يؤدي بشكل عام إلى ذلك التحيز. هناك أوجه أخرى للإدارة مثل، حالة الغرفة، حافز الاختبار، التأثيرات الإدارية، ومشكلات في التواصل بين الإداري والشخص الذي يأخذ الاختبار يمكن أن تؤدي إلى تحيز في نتائج الاختبار. قد يوجد تحيز المنهج إلى حد ما في كل مقارنات نتائج اختبار الإنجاز والأهلية عبر الثقافات ويمكن أن يقود إلى تفسير الاختلافات في النتائج بين المجموعات التي سببتها إجراءات الاختبارات وكأنها تفسير اختلافات حقيقية في المقدرة (هامبلتون وكانجي، 1993؛ فان دي فيفر وهامبلتون، 1996).

يمكن دراسة التحيز الحاصل بسبب الإجراءات الإدارية بطرق عديدة. إحدى تلك الطرق هي إعادة إدارة الاختبار. إن دراسة تغيرات تماثل الدرجات عبر الثقافات في إدارة الاختبار الأول ثم إدارة الاختبار الثاني يمكن أن تعطي دلائل مهمة عن صدق الاستنتاجات التي تم الحصول عليها.

عندما يحصل طلاب من مجموعات مختلفة على درجات اختبار متساوية في الاختبار الأول ودرجات مختلفة في الثاني، يمكن أن يكون هناك تساؤل عن صدق استنتاج الدرجات التي تم الحصول عليها من الاختبار الأول.

هناك طريقة ثانية لدراسة التحيز بسبب طريقة إدارة الأداة وهي استخدام الأداة طريقة غير قياسية. يكون ذلك ممكناً عند القيام باستخلاص كل المعلومات والإجابات من الطلاب عن تفسير الإرشادات، البنود، إجابة الخيارات، والحوافز عند اختيار أسئلة محددة (فان دي فيفر وهامبلتون، 1996).

هنالك الكثير من الطرق لتجنب مشكلات متعلقة بتحيز الإدارة. إحدى الطرق التي ينصح بها هي التأكد من أن إرشادات الاختبار نفسه واضحة وتُسيّر نفسها بنفسها، مع أقل اعتماد على التواصل الشفهي مع الإداري (فان دي فيفر وبورتينغا، 1991). في كل مواقف الاختبار من الضروري فهم الثقافات التي يأتي منها الطلاب الذين يجرون الاختبار ويجب أن يكون الإداريون قادرين على التواصل في لغة هؤلاء الطلاب على الأقل (كيسنجر، 1994؛ كيسنجر وكارلسون، 1992).

خلاصة لذلك من المفضل أن يكون إداريو الاختبار من قومية المجموعة الآخذة للاختبار ذاتها، أو أن يكون لهم معرفة بالثقافة واللغة، وأن يتمتعوا بالخبرة والمعرفة بالإدارة وأن يملكوا خبرة في القياسات (هامبلتون، كانجي، 1993). من الضروري أيضاً تشجيع ثبات إجراءات إدارة الاختبار عبر المجموعات المختلفة التي يتم اختبارها. إن أفضل طريقة لذلك هو توفير تدريب مستمر لجميع إداريي الاختبار. يجب أن يؤكد ذلك التدريب على وضوح وعدم غموض عملية التواصل بين الإداريين وآخذي الاختبار، على أهمية اتباع إرشادات الاختبار بدقة، التقيد بالوقت المحدد وعلى التأثير المحتمل لإداري الاختبار على صدق الاستنتاجات التي تم الحصول عليها من الدرجات (هامبلتون وكانجي، 1993).

قدم فان دي فيفر وهامبلتون (1996) خطوطاً عريضة واضحة ودقيقة لإداريي الاختبارات المكيفة:

1- يجب على مطوري الأداة والإداريين محاولة توقع كل أنواع المشكلات التي يمكن حدوثها واتخاذ الإجراءات المناسبة لمعالجة تلك المشكلات وذلك بإعداد مواد وإرشادات مناسبة. كما أشار فان دي فيفر وهامبلتون فإن توقع مشكلات إدارية ليس معقداً جداً. يمكن جعل تلك المهمة سهلة وذلك عند إقامة دراسة استطلاعية تستخدم الاختبار بطريقة غير قياسية لاستخلاص إجابات مختلفة من الطلاب، يمكن أن تساعد الملاحظة الدقيقة والتغذية الراجعة من الطلاب في الكشف عن التأثيرات الإدارية المحتملة.

2- يجب على إداريي الأداة إدراك بعض العوامل المتعلقة بالمواد المحفزة، الإجراءات الإدارية، طرق الإجابة التي يمكن أن تُقوم صدق الاستنتاج التي تم أخذها من الدرجات. بالرغم من أن الترجمة الحرفية كانت تُفضل دائماً فإن على الإداريين معرفة المشكلات التي يمكن أن تحدثها تلك الترجمات. على سبيل المثال يمكن أن تكون هنالك بعض الأوجه الضمنية في الإرشادات لم يجر إيصالها من خلال الترجمة.

3- يجب أن تكون أوجه البيئة المحيطة التي تؤثر على إدارة الأداة متشابهة إلى أبعد حد عبر المجموعات التي تستهدفها تلك الأداة. من المعروف أن السيطرة على وضع البيئة في البحث الميداني مستحيل. لكن على الإداريين أن يدركوا العوامل البيئية المختلفة التي يمكن أن تؤثر على صدق الدرجات لكي يقوموا بكل جهد لجعلها ثابتة.

4- يجب أن تكون إرشادات إدارة الأداة موجودة في اللغتين: لغة المصدر واللغة المستهدفة وذلك للتقليل من التأثير غير المرغوب به للاختلافات. بالإضافة إلى ذلك فإن إرشادات اختبار مطولة تحتوي على تمارين وأمثلة مختلفة تساعد على تقليل الاختلافات المتعلقة بالإدارة نفسها.

5- يجب أن يحدد كتيب الأداة كل أوجه الأداة وإدارتها التي تتطلب التدقيق في استخدام الأداة في مفهوم ثقافي جديد، أشار فان دي فيفر وهامبلتون أنه بما أن مطوري الاختبار يعملون على تكييف الاختبار للاستخدام في ثقافات ولغات مختلفة فإنهم سيكشفون عن موضوعات محددة عن استخدام ذلك الاختبار في المحيط الثقافي الجديد.

يستفيد إداريو الاختبار بمعرفة تجربة مطوري الاختبار ويجب أن يدركوا المشكلات المحتملة لكي يتجنبوا إعادتها.

6- يجب أن لا يكون الإداريون فضوليين ويجب الإقلال عن التفاعل بين الإداريين والذين يأخذون الاختبار. يجب اتباع القواعد المذكورة في كتيب الأداة. تنتج أخطاء شائعة ومهمة في مقارنة درجات عبر الثقافات بسبب التفاعل غير المحكم بين الإداريين والطلاب. يجب أن يحدد كتيب إداريي الاختبار تلك المشكلات الحالة ويقدم الحلول المناسبة.

الخلاصة

كان الهدف من هذا الفصل مراجعة الموضوعات المنهجية المتعلقة بتكييف قياسات الإنجاز والأهلية. ركزنا في هذا الفصل على ست نواح:

(1) تكافؤ البنية.

(2) تكافؤ وحدات القياس.

(3) ترجمة الاختبارات ومواد الاختبار.

(4) متغيرات البند الوظيفية.

(5) تفسير الدرجات.

(6) إدارة الاختبار.

وجدنا أن التقدم الحاصل في تلك النواحي في العقد الماضي مطمئنة جداً بحيث أصبح الباحثون المهتمون بالمقارنة عبر اللغات/ الثقافات لقياسات الإنجاز والأهلية أكثر إدراكاً لتعقيدات الموضوعات التي تؤثر على القدرة للقيام بمقارنة تقويم درجات صادقة.

نعزو نمو ذلك التطور إلى عوامل عدة. واحد منها هو أن تعدد القوى السياسية والاقتصادية التي تؤدي بدورها إلى الاقتصاد العالمي مع تدفق الهجرة جعل من الضروري إجراء الاختبار في لغات عديدة. كنتيجة لذلك أصبح كثير من الممارسين يهتمون بتلك الناحية للتقويم وأصبحوا يثيرون انتباه الباحثين إلى مشكلات عملية متنوعة. بالإضافة إلى ذلك أصبح كثير من المنهجيين مثل IRT وبعض نماذج تعادل البنية يهتمون بتلك الأمور.

نتنبأ بأن العقد التالي سوف يظهر تضاعفاً للحاجة ولاستخدام الاختبارات المكيفة. مع الاستخدام المتوسع للتقويم الذي خصص لثقافة ولغة واحدة والذي جرى استخدامه لمجموعات ذات ثقافات ولغات مختلفة سيصبح هناك معرفة متضاعفة عن المهارات والأهلية المقاسة عبر المجموعات الثقافية بالإضافة إلى طرق محسنة للقيام بالمقارنات الضرورية.

شكر

يرغب الكتّاب بالتعبير عن شكرهم لمشاركة دانيل أنجور بالعمل في هذا الفصل. له التقدير الكبير لمساعدته في التحرير.

*****

المراجع

Angoff, W.H., & Cook, L. L. (1988). *Equating the scores of the "Prueba de Aptitud Academica" and the "Scholastic Aptitude Test"* (Report No. 88-2). New York: College Entrance Examination Board.

Bracken, B. A., & Barona, A. (1991). State of the art procedures for translating, validating and using psychoeducational tests in cross-cultural assessment. *School Psychology International, 12,* 119–132.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1,* 185–216.

Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research, 29,* 543–553.

Ercikan, K. (1999, April). *Translation DIF in TIMSS.* Paper presented at the meeting of the National Council of Measurement in Education, Montreal, Canada.

Ercikan, K. (2000). Disentangling sources of differential item functioning in multilingual assessments. *International Journal of Testing, 2,* 199–215.

Everson, H. T., Guerrero, A., & Laitusis, V. (1998, April). *Preliminary evidence of construct equivalence of mathematics tests administered across languages: An analysis of findings from the SAT I and the Prueba de Aptitud Academica tests.* Paper presented at the meeting of the American Educational Research Association, San Diego.

Eysenck, H. J., & Eysenck, S. B. G. (1983). Recent advances in the cross-cultural study of personality. In J. N. Butcher & C. D. Spielberger (Eds.), *Advances in personality assessment* (Vol. 2, pp. 41–69). Hillsdale, NJ: Lawrence Erlbaum Associates.

Geisinger, K. F. (1994). Cross-cultural normative assessment: translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6,* 304–312.

Geisinger, K. F., & Carlson, J. F. (1992). *Assessing language-minority students* (Report No. EDO-TM-92-4). Washington DC: ERIC Clearinghouse on Tests, Measurement, and Evaluation.

Gierl, M. (2000). Construct equivalence on translated achievement tests. *Canadian Journal of Education, 25*(4), 280–296.

Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. *European Journal of Psychological Assessment, 9*(1), 57–58.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10,* 229–244.

Hambleton, R. K. (1996, March). *Guidelines for adapting educational and psychological tests.* Paper presented at the meeting of the National Council of Measurement in Education, New York.

Hambleton, R. K., & Bollwark, J. (1991). Adapting tests for use in different cultures: Technical issues and methods. *Bulletin of the International Test Commission, 18,* 229–244.

Hambleton, R. K., & Kanjee, A. (1993, April). *Enhancing the validity of cross-cultural studies: Improvements in instrument translation methods.* Paper presented at the meeting of the American Educational Research Association, Atlanta, GA.

Hambleton, R. K., & Kanjee, A. (1995). Translating tests and attitude scales. In T. Husen & T. N. Postlewaite (Eds.), *International encyclopedia of education* (2nd ed., pp. 6328–6334). New York: Pergamon.

Hambleton, R. K., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. *Social Indicators Research, 45,* 153–171.

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology, 16,* 131–152.

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology, 20,* 296–309.

Lapointe, A. E., Mead, N. A., & Askew, J. M. (1992). *Learning mathematics* (Report No. 22-CAEP-01). Princeton, NJ: Educational Testing Service.

Olmedo, E. E. (1981). Testing linguistic minorities. *American Psychologist, 36,* 1078–1085.

Perez, A. (1975). *Measurement of bilingual ability.* Unpublished master's thesis, University of Puerto Rico, San Juan.

Poortinga, Y.H. (1983). Psychometric approaches to intergroup comparison: The problem of equivalence. In S. H. Irvine & J. W Berry (Eds.), *Human assessment and cross-cultural factors?* (pp. 237–258). New York: Plenum.

Poortinga, Y. H. (1989). Equivalence of cross-cultural data: An overview of basic issues. *International Journal of Psychology, 24,* 737–756.

Poortinga, Y. H. (1995). Uses of tests across cultures. In T. Oakland & R. K. Hambleton (Eds.), *International perspectives on academic assessment* (pp. 187–206). Boston: Kluwer Academic.

Prieto, A. (1992). A method for translation of instruments to other languages. *Adult Education Quarterly, 43,* 1–14.

Reckase, M. D., & Kunce, C. (1999, April). *Translation accuracy of a technical credentialing examination.* Paper presented at the meeting of the National Council of Measurement in Education, Montreal, Canada.

Sireci, S. G. (1997). Problems and issues in linking assessment across languages. *Educational Measurement: Issues and Practice, 16,* 12–19.

Sireci, S. G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing, 20*(2), 147–165.

Sireci, S. G., Fitzgerald, C., & Xing, D. (1998, April). *Adapting credentialing examinations for international uses.* Paper presented at the meeting of the American Educational Research Association, San Diego.

Stanley, J. C., & Hopkins, K. D. (1972). *Educattonal and psychological measurement and evaluation.* Englewood Cliffs, NJ: Prentice-Hall.

van de Vijver, F. J. R. (2002). Cross-cultural assessment: Value for money? *Applied Psychology: An International Review, 51*(4), 545–566.

van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating test: Some practical guidelines. *European Psychologist, 1,* 89–99.

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research.* Thousand Oaks, CA: Sage.

van de Vijver, F. J. R., & Leung, K. (2000). Methodological issues in psychological research on culture. *Journal of Cross-Cultural Psychology, 31,* 33–51.

van de Vijver, F. J. R., & Lonner, W. (1995). A bibliometric analysis of the *Journal of Cross-Cultural Psychology, 26,* 591–602.

van de Vijver, F. J. R., & Poortinga, Y. H. (1991). Testing across cultures. In R. K. Hambleton & J. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 277–308). Dordrecht, Netherlands: Kluwer Academic.

van de Vijver, F. J. R., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47,* 263–279.

Van der Flier, H., & Drenth, P. J. D. (1980). Fair selection and comparability of test scores. In L. J. T. van der Kamp, W. F. Langerak, & D. N. M. de Gruijter (Eds.), *Psychometrics for educational debates.* New York: Wiley.

Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement, 30,* 1–21.

Westbury, I. (1992). Comparing American and Japanese achievement: Is the United States really a low achiever? *Educational Researcher, 21,* 18–24.

Woodcock, R. W. (1985). *Woodcock Language Proficiency Battery, Spanish Form: Technical summary* (Assessment Services Bulletin No. 9). Allen, TX: DLM.