1

موضوعات وخطط، وإرشادات تقنية لتكييف الاختبارات للغات وثقافات متعددة

رونالد. ك هامبلتون
جامعة ماستشوست/ أمهرست

يوجد اليوم عدد كبير من الإثباتات عن أن الحاجة إلى نسخ معدّة بلغات متعددة عن اختبارات الذكاء، الإنجازات، والشخصية. وتقارير مسح شاملة في ازدياد (مثال، اريسكان 2002، هامبلتون، 2002، هامبلتون دو يونغ، 2003، هاركنس، 1998). فعلى سبيل المثال، دعت الجمعية الدولية لتقييم الإنجازات التعليمية (IFA) الأبحاث التالية في الرياضيات والعلوم (TIMSS) العالمية في أكثر من 45 دولة، وكان من مهامه تهيئة اختبارات في الرياضيات والعلوم في أكثر من 30 لغة. هناك أمثلة بارزة لمشاريع جديدة في تكييف الاختبارات في الولايات المتحدة تتضمن خططاً لإعداد نسخ في اللغة الإسبانية لاختبار القبول الجامعي اختبار التقييم المدرسي، (SAT)، اختبار مجلس التعليم الأميركي "تطور التعليم العام (GED) واختيار الإدارة التعليمية في الولايات المتحدة "التقييم الوطني للتقدم التعليمي" وعدد كبير من اختبارات الإنجاز في إدارات التعليم الحكومية. وبالفعل من المتوقع القيام بعملية تكييف عدد أكثر من الاختبارات في المستقبل؛ لأن (أ) أصبح التبادل الدولي للاختبارات أكثر شيوعاً، (ب) تستعمل اختبارات أكثر للحصول على المصداقية الدولية، (ج) وازدياد الاهتمام بأبحاث عبر الثقافات.

بالرغم من أن أسباب تكييف الاختبارات من لغة وثقافة إلى أخرى واضحة فعلى سبيل المثال، تسهل دراسة مقارنة الانجازات المدرسية عبر مجموعات مختلفة الثقافة واللغة، توفير المال والوقت المتعلق بتهيئة اختبارات جديدة، والعدل في التقييم فإن الطرق والخطوط العريضة لإعداد تكييف الاختبارات ومعادلة النتائج ليست معروفة جيداً (هامبلتون، 1993، هوي وترياندس، 1985، فان دي فيفر وهاملتون، 1996). حتى إن بعض الباحثين في الدراسات عبر الثقافات علق أن نسبة كبيرة من الأبحاث في ذلك الحقل يحتوي كثيراً من الأخطاء إلى حد جعله غير صالح بسبب عملية التكييف السيئة.

إن القصد من هذا الفصل هو (أ) مراجعة عدد من المصادر عن الأخطاء وعدم الصدق المرافقة لتكييف الاختبارات واقتراح طرق لتقليل تلك الأخطاء (ب) وصف بعض الخطط لتكييف الاختبارات التي قامت بها الهيئة الدولية للاختبارات (ITC) بمساعدة سبع وكالات عالمية (هاملتون، 1994، فان دي فيفر وهاملتون، 1996).

قبل البدء يجب التمييز بين تكييف الاختبارات وترجمتها. يفضل استخدام «تكييف الاختبار» على المصطلح «ترجمة الاختبار» الذي هو أكثر شيوعاً واستعمالاً في هذا الفصل؛ لأن المصطلح الأول أوسع وأكثر انعكاساً على ما يجب القيام به في الواقع عند إعداد اختبار تم إعداده للاستخدام في لغة وثقافة واحدة للتطبيق في لغة وثقافة أخرى.

يتضمن تكييف الاختبار كل الأنشطة بدءاً من تقرير عما إذا كان باستطاعة الاختبار تقدير تركيبة الاختبار ذاتها في لغة وثقافة أخرى، اختبار المترجمين، تقرير التكييف المناسب الذي يجب القيام به لإعداد الاختبار للاستعمال في لغة ثانية، إلى تكيف الاختبار والتأكد من تطابقه مع الشكل المكيف، إن ترجمة الاختبار خطة واحدة من عملية تكييف الاختبار وحتى في تلك الحالة مصطلح التكييف مناسب أكثر من مصطلح الترجمة لوصف العملية الحقيقية التي تجري؛ ذلك لأن المترجمين يحاولون الحصول على مفاهيم، مفردات وتعابير متعادلة ثقافياً، نفسياً

ولغوياً للغة والثقافة الأخرى، بذلك تأخذ المهمة أبعاداً أكثر من ترجمة محتويات الاختبار حرفياً.

نستعمل المصطلح "اختبار" لغايتنا في إدراج كل النماذج والأدوات التربوية والنفسية، وحتى عمليات المسح والاستبيانات.

أسباب الأخطاء وعدم صدق تكييف الاختبار:

تؤمن الجمعية الأميركية للأبحاث التعليمية (AERA)، الجمعية النفسية الأميركية (APA)، والهيئة الوطنية للمقاييس في التعليم (NCME)، المعايير للاختبارات التربوية والنفسية (1985) تعليمات دقيقة للاختصاصيين في المقاييس التعليمية والنفسانيين الذين يختارون، يطورون، ويشرفون، ويستخدمون الاختبارات النفسية والتعليمية، هناك ثلاثة معايير في هذا الكتاب متعلقة بموضوع تكييف الاختبار.

المعايير 6.2 عندما يقوم مستخدم الاختبار بتغييرات أساسية في بنية الاختيار، طريقة الاستخدام، التعليمات، اللغة أو المحتوى، يجب عليه إعادة صدق استخدام الاختبار حسب حالات التغييرات أو عرض أسباب منطقية تدعم الادعاء أو مصداقية إضافة ليست ضرورة أو ممكنة.

المعايير 13.4 عندما يترجم اختبار من لغة/ لهجة إلى أخرى لاستخدامها لاختبار مجموعات ذات لغة واحدة يجب التثبت من مصداقيتها وجدارتها.

المعايير 13.6 إذا كان المقصود مقارنة نسختين لاختبارين في لغتين، يجب أن يدون دليل على مقارنة الاختبار.

توفر هذه المعايير خطوط عمل لاعتبار مصادر الأخطاء أو عدم الصدق الناتجة عن الجهود لتكييف الاختبار من لغة إلى أخرى ومن ثقافة إلى أخرى.

ولأغراضنا من الممكن تنظيم مصدر الأخطاء أو عدم صلاحيتها في ثلاث فئات عامة: (أ) اختلافات لغوية ثقافية، (ب) موضوعات تقنية، خطط، وطرق (ج) ترجمة

النتائج. إن الفشل في الاهتمام بمصادر الخطأ في كل من تلك الفئات يمكن أن ينتج عن عدم مساواة الاختبار الذي جرى تكييفه عند استخدامه في مجموعتين المختلفتين لغوياً وثقافياً. إن الاختبارات غير المتساوية، عندما يفترض أن تكون متساوية تؤدي إلى أخطاء في التفسير ونتائج مغلوطة عن المجموعات المشاركة.

مثال جيد عن خطأ في التفسير بسبب تكييف شيء لاختبار (هذا المثال قدمه ريتشارد وولف من كلية المعلمين في كولومبيا، وهو رائد في مهنته في حقل التقويم الدولي). وفي دراسة مقارنة دولية في القراءة (1990)، طلب من طلاب أميركيين دراسة أزواج من المفردات وتعريفهم كمتماثلين أو مختلفين في المعنى:

أدرجت الكلمتان "متشائم - دموي المزاج" ضمن مجموعة المفردات التي أحرز فيها الطلاب نقاطاً متوسطة (54٪) من الطلاب الأميركيين أعطوا الإجابة الصحيحة). كانت البلاد غير الناطقة بالإنجليزية الأولى في الأداء 98٪ من الطلاب أعطوا الإجابة الصحيحة. من خلال محاولة معرفة أسباب الاختلاف الكبير في الأداء اكتشف أن كلمة (دموي المزاج) ليس لها مرادف في ذلك البلد المرتفع الأداء ولذلك استعملت كلمة (متفائل)، جعل هذا التبديل السؤال أسهل وكان، من الممكن لنسبة كبيرة من الطلاب الأميركيين الإجابة عنه بالشكل الصحيح لو قدمت الكلمات المزدوجة بشكلها الجديد (متشائم - متفائل). إن الغرض من هذا المثال التركيز على الخطر من أخذ الاستنتاجات في الدراسات العالمية المقارنة للأداء دون دليل على عملية التكييف الناتجة في اختبارين متماثلين. قبل 1990 كان هناك كثير في المبادرات لدراسات عالمية تشمل أكثر من استخدام بعض المترجمين الجيدين، ويمكن مقارنة اختلاف ذلك مع تكييف الاختبارات المتطور الذي نشاهده اليوم في جمعية (TIMSS) ومنظمات التعاون الاقتصادي وبرامج التطوير للتقويم الدولي للطلاب (OECD)، (PISA انظر غريساي، 2003، هاملتون، 2002).

فيما يلي مناقشة عدة أغلاط شائعة وكيف يمكن معالجتها بشكل عملي.

الاختلافات اللغوية/ الثقافية التي تؤثر في النتائج:

إن تقويم وترجمة النتائج عبر الثقافات لا يجب أن ينظر إليها من الزاوية الضيقة لترجمة وتكييف الاختبارات (فان دي فيغر ولونغ، 1997، 2000). لكن يجب اعتبار هذه العملية ضمن كل مراحل عملية التقويم ومن ضمنها تساوي بنية الاختبار، إدارة الاختبار، بنية البنود المستعملة، وأثر السرعة على أداء الممتحن هذه العوامل الأربع سيجري مناقشتها لاحقاً، وستلقى اهتماماً أكثر في الفصول التالية:

البنية المتكافئة Construct Equivalence

يتضمن تكافؤ البنية كلاً من التكافؤ في المفهوم/ الوظيفة بالإضافة إلى التساوي في طريقة قياس البنية في عملية الاختبار في مجموعة مختلفة اللغة/ الثقافة (هاركنس، 1998). إذا فرضنا وجود البنية المتكافئة بين الثقافات المختلفة التي جرت دراستها فإن القيام بالمقارنة بين الدراسات عبر الدول. عبر الثقافات وعبر اللغات أساسي. إن استخدام اختبار غير متكافئ البنية هو أكثر الأخطاء أهمية في البحث عبر اللغات المختلفة.

على سبيل المثال؛ مقارنة أداء دولتين في الرياضيات: إذا كان اختبار المحتوى يعكس الاهتمام الأكبر للرياضيات في المناهج الدراسية في دولة وليس بذات الأهمية في الدولة الأخرى. مثال آخر: يمكن أن يكون في بنية (نوعية الحياة) يمكن أن يتضمن مفهوم الحياة كثيراً من الأمور المادية كالسيارات، المنازل، التلفزيونات، بينما لا يتضمن مفهوم الحياة في الدولة الأخرى أكثر من الطعام للبقاء وطبيب قريب من المنزل. إن مقارنة النتائج في اختبار الدولة ذات نوعية حياة جيدة وتم تكييفه للاستخدام في الأخرى له قيمة ضعيفة.

فإذا قررنا عما إذا وجود البنية المتكافئة بين ثقافتين يتضمن استراتيجية عقلانية يجب أن يبدأ الباحث باستخدام بديهته للإجابة عن أسئلة عديدة، على سبيل المثال هل من المعقول مقارنة تلك الثقافتين حسب تلك البنية؟

هل تلك البنية الذي تم دراستها لها معنى مواز في كل الثقافات التي يجري مقارنتها؟ هل تلك البنية فعالة في تلك الدراسات.

لكي نستطيع الإجابة بنعم عن تلك الأسئلة وضمان تكافؤ المفاهيم/ الوظيفة وتكافؤ فعالية تلك البنية يجب اتخاذ عدة طرق. من الممكن القيام بهذا عن طريق مقابلة وملاحظة الأشخاص في الثقافات المعنية، إجراء الأبحاث عن تلك الثقافات وطرح أسئلة على آخرين يعرفون تلك الثقافات. إن هذه الطرق موضوعية ولذلك فإن استخدام مصادر أدلة مختلفة مستحسن جداً. فان دي فيفر وبورتينغا (الفصل الرابع) لديهم الكثير ليقولوه في ذلك الفصل عن الحكم على البنية المتكافئة.

إدارة الاختبار:

تهدد صعوبات التفاهم بين الذين يجرون الاختبار وبين الذين يديرون الاختبار صدق نتائج الاختبار بشكل كبير. من الممكن أن تكون تعليمات الاختبار غير واضحة بسبب صعوبات في الترجمة. إحدى الطرق للحيلولة دون ذلك، وهي ممكنة، هو التأكد من كون التعليمات في الاختبار نفسه واضحة ومفهومة بذاتها وبأقل اعتماد على الاتصال اللفظي (فان دي فيفر وبورتينغا، 1991).

ومن المتوقع وجود بعض الصعوبات الأخرى في تعليمات تقدير مقياس الدرجات المستخدم في "قياس الموقف" أيضاً؛ لأن تلك الاختبارات ليس لها وجود في الدول الأخرى (انظر هاركنس، 1998).

إن اختيار الإداريين المناسبين للاختبار من الممكن أن يكون مفيداً أيضاً. فيجب أن تتوفر بينهم الشروط التالية (أ) أن يجري اختبارهم في موطن المجموعة التي تجري الاختبار (ب) أن يكونوا مطلعين على الثقافة، واللغة، واللهجة (ج) لديهم مهارات كافية في إدارة الاختبارات (د) أن يدركوا أهمية تطبيق الإجراءات المتبعة أثناء الاختبار.

بالإضافة إلى ذلك فإن التناسق في إدارة الاختبارات لمجموعات مختلفة يمكن أن يكون أفضل إذا توفر التدريب الأساسي لكافة الأشخاص الذي يديرون الاختبار.

يجب أن يخطط لحصص تدريبات كجزء من عملية تطوير الاختبار والتأكيد على وضوح وعدم غموض التواصل بين الإداريين والممتحنين، أهمية اتباع إرشادات، ضبط الوقت المحدد للاختبار، وتأثير مقدمي الاختبار على جدارة وصدق الاختبار وغير ذلك.

شكل ومحتوى الاختبار Test Format

إن التفاوت المألوف في بنود البنية يشكل مصدراً آخر لعدم مصداقية نتائج الاختبار في الدراسة عبر الثقافات. في الولايات المتحدة استخدمت بنود استجابة متعددة مثل بنود ذات عدة إجابات لاختيار أحدها بشكل كبير في الاختبار (مع أن هذا يجري تغييره في السنوات العشر الأخيرة واليوم نرى استعمال أسئلة تقييم الأداء أكثر). في دراسة عبر الثقافات لا نستطيع الجزم أن كل المتقدمين للاختبار مطلعين على طريقة تلك الأسئلة مثل الطلاب الاميركيين. إن الدول التي تتبع النظام البريطاني في التعليم (تاريخياً على الأقل) تؤكد أكثر على كتابة المقالات وأسئلة ذات إجابة قصيرة بالمقارنة مع الأسئلة ذات الإجابات المتعددة. وبذلك يكون الطلاب من تلك البلاد في وضع خاسر مقارنة مع نظرائهم الأميركيين. عندما يكون التأكيد في بنية الاختبارات على استجابات مثل كتابة المقالات كطريق أساسية للتقويم يكون الممتحنون ذوو المعرفة بالأسئلة ذات الإجابات المتعددة في وضع صعب. في بعض الأحيان يكون التوازن في بنية الاختبار الأفضل لتحقيق العدالة والتقليل من عدم مصداقية عملية التقويم. وقد اعتمدت هذه الطريقة في الدراسة الدولية الحديثة للأداء (TIMSS & OECD / PISA).

هناك حل آخر للتغلب على أثر احتمال التحيز المرافق لبنية أحد البنود بشكل خاص هي في أن يتضمن الاختبار الذي يقوم بتقييم المجموعات البنود المألوفة لكل تلك المجموعات. عندما يكون من المؤكد أن الطلاب ليسوا في وضع صعب وأن كافة المعطيات جرى قياسها من المفضل استخدام الأسئلة ذات الإجابات المتعددة أو مقياس درجات بسيط.

إن الميزة الكبرى لاستخدام الأسئلة ذات الإجابات المتعددة أو مقياس درجات بسيط هو أن تقديرها أكثر موضوعية وبذلك يمكن تجنب تعقيدات المقاييس التي تصاحب الإجابات ذات النهاية المفتوحة، هذا يرتبط بشكل خاص في الدراسات عبر الثقافات، حيث من الممكن أن يكون ترجمة قواعد المقاييس أكثر صعوبة من الاختبار نفسه. بالإضافة إلى ذلك فإن تعليمات كثيرة واضحة تتضمن أمثلة وتمارين تساعد في الإقلال من تفاوت الاعتياد (فان دي فيفر وبورتينغا، 1992). في ذات الوقت فإن استخدام بنية بند واحد للاختبار قد يكون من الخطورة أن يضيق بنية الاختبار المهمة إلى تلك الأجزاء التي يمكن قياسها حسب بنية البند الأوحد، أيضاً قد يحرف نتائج الدراسات المقارنة عبر الحدود القومية.

السرعة:

جرى غالباً الافتراض أن الممتحنين يعملون بشكل سريع في الاختبارات السريعة (فان دي فيفر وبورتينغا 1991) ولكن معرفة العمل السريع هي مهارة في عملية أخذ الاختبار التي من الممكن أن لا تكون معروفة أو مفهومة من قبل الممتحنين في ثقافات مختلفة في دراسة تُقارن بين طلاب هولنديين وطلاب وعرقيات مختلفة في هولندا، وجد فان ليست وبريخرودت (1995) أن عامل السرعة ضاعف الانحياز في الدرجات.

لأنه ليس، لدى كل الثقافات الخبرات في الاختبارات السريعة وكان الممتحنون الفاقدون لتلك الخبرة في وضع حرج جداً. هنالك كثير من الدراسات المختلفة التي تسلط الضوء على بنود وتميز في الاختبار بسبب دور السرعة في الاختبار (انظر دراسات حول التميز العرقي في اختبارات تقويم الطلاب في الولايات المتحدة). على سبيل المثال في (SATs) بعض البنود الأخيرة في الاختبار عادة تظهر تحيزاً أكثر من تلك التي توجد في بداية الاختبار. يكون التمييز هذا ضد القراء الضعفاء وهذا غالباً ما يكون بسبب دور السرعة في أداء الاختبار. إن الحل الأمثل هو التقليل من

السرعة كعامل في اختبار أداء المعرفة إلا إذا كانت جزءاً من البنود التي يجري قياسها. إن النقطة الأخيرة مهمة جداً لأنه في بعض الأحيان تكون السرعة في الأداء جزءاً متكاملاً من قياس بنية الاختبار كما في حالة القدرة على حل مسائل في التحليل الاستنتاجي. عندئذ تكون السرعة قسماً مهماً من الاختبار وعلى الممتحنين فهم ضرورة العمل السريع.

موضوعات تقنية، خطط، وطرق:

هنالك خمسة عوامل تقنية تؤثر في صدق الاختبارات المكيفة للاستخدام في لغات وثقافات أخرى. الاختبار نفسه، اختيار وتدريب المترجمين، عملية الترجمة، خطط عقلانية لتكييف الاختبار وخطط لجمع المعطيات وتحليلها لتثبيت التكافؤ. سيجري بحث كل من هذه العوامل بشكل مختصر. وتظهر مناقشة تلك العوامل بشكل مفصل في الفصول اللاحقة.

الاختبار:

إذا كان الباحث يعرف أنه سيستخدم الاختبار للغة أو ثقافة مختلفة، المفيد أن يضع ذلك في الحسبان في بداية عملية تطوير الاختبار. وإذا أخفق في ذلك فسينتج عن ذلك صعوبات في عملية التكييف التي تؤدي بدورها إلى تخفيض صدق الاختبار المكيف (هامبلتون وباتسولا، 1999).

إن اختيار شكل الاختبار من مواد محفزة له، المفردات، تركيب الجمل ونواح أخرى يمكن أن تشكل صعوبة في الترجمة الجيدة التي يجب أن تؤخذ جميعها بالحسبان عند إعداد مواصفات الاختبار.

يمكن لذلك العمل الوقائي التقليل من المشكلات اللاحقة. على سبيل المثال أسئلة عن النقود يمكن حذفها لأن العملات مختلفة في العالم ومن الممكن صعوبة إيجاد تكافؤ في ترجمتها لوضعها في الاختبار. كذلك نص القراءة عن موضوعات خاصة بإحدى الثقافات مثل "الهوكي" تبدو غير مألوفة في عدة ثقافات أخرى

ويمكن رفضها والاستعانة بمقاطع عن المشي في الحديقة أو نشاطات أخرى يمكن أن يكون لها معنى في لغات وثقافات مجموعات أخرى. وتنشأ صعوبة أخرى في تكييف النصوص من الإنجليزية إلى لغات أخرى وهي وجود "المبني للمجهول" في النص لأن هذا الزمن في القواعد موجود في اللغة الإنجليزية ولكنه غير موجود في لغات أخرى (الإسبانية على سبيل المثال).

أما في معيار الشخصية، فيجب أن يؤخذ الحذر في اختيار المواقف، المفردات، والتعابير التي يمكن تكييفها بسهولة عبر الثقافات/ اللغات المختلفة للمجموعات. على سبيل المثال: يمكن أن يكون بعض أنواع السلوك عادياً في العالم الغربي ولكن له معنى آخر أو ليس له أي معنى في ثقافات أخرى. عبارة "أحب أن أقوم بالمحادثة في حفلة" ليس لها معنى في ثقافة لا يكون فيها حفلات أو حيث لا تذهب النساء إلى الحفلات أو حيث المبادرة إلى الحديث يمكن أن يكون تصرفاً غير مقبول. هذا واحد فقط من الأمثلة التي من الممكن مواجهتها.

اختيار وتدريب المترجمين:

إن أهمية الحصول على خدمات مترجمين مؤهلين واضحة. حاول الباحثون متابعة عملية الترجمة لمترجم واحد تم اختياره لأنه كان من الممكن الوصول إليه/ إليها لكونه صديقاً، زوجة، أو شخصاً يمكن استخدامه بمبلغ بسيط إلى ما هنالك. إن عمل الترجمة الكفء لا يمكن أن يعتبر أمراً مفروغاً منه، كذلك فإن استخدام مترجم واحد مؤهل أو غير مؤهل لا يسمح بالحصول على تفاعل ذي قيمة بين المترجمين المختلفين لإيجاد الحلول لنقاط عديدة تنشأ عند القيام بعملية تكييف الاختبار. قد يكون لأحد المترجمين على سبيل المثال وجهة نظر في استخدام مفردات أو تعابير مفضلة لديه قد لا تكون مناسبة لتحقيق تكييف جيد للاختبار. من الممكن أن يكون استخدام مترجمين عدة حماية ضد أخطار استخدام المترجم الوحيد مع تفصيلاته وخصوصيته اللغوية.

في ذات الوقت، يجب أن يكون المترجمون أكثر من أشخاص مؤهلين ومتآلفين مع اللغات المستخدمة في الترجمة، يجب أن يعرفوا الثقافات جيداً وبشكل خاص الثقافة التي يترجم إليها (الثقافة المرتبطة باللغة التي يجري ترجمتها). إن هذه المعرفة أساسية في فعالية التكييف. كذلك من المفضل جداً معرفة مواد الموضوعات في تكييف اختبارات الأداء. إن دقة وفروق المعنى في موضوع ستخفى عن مترجم ليس له معرفة في ذلك الموضوع وغالباً يعود المترجمون الذين يجهلون المعرفة التقنية إلى الترجمة الحرفية التي تؤدي بدورها إلى إحداث صعوبات للطلاب الذين يجرون الامتحان في اللغة الثانية وتهدد صدق الاختبار. مثلاً إن جملة "Je ne suis pas une valise" في الفرنسية لها ترجمة حرفية سهلة في الإنجليزية، (أنا لست حقيبة) ولكن المعنى الحقيقي لتلك الجملة في اللغة الفرنسية هو "لست غبياً إلى هذا الحد" إن الترجمة الحرفية من الفرنسية إلى الإنجليزية قد شوه المعنى بالكامل.

أخيراً إن المترجمين سوف يستفيدون من بعض التدريب في تكوين الاختبار. مثلاً يجب أن يعرف المترجمون عندما يقومون بتكييف اختبارات الأداء والأهلية عدم إحداث قوافي لغوية تقود الممتحنين إلى الإجابات الصحيحة وترجمة البنود المتشابهة في الأسئلة ذات الإجابات المتعددة بشكل يجعلها ذات معنى واحد. إن المترجم الذي ليست لديه أي معرفة في مبادئ الاختبار وبناء المقاييس يمكن بسهولة أن يجعل الاختبار أقل أو أكثر صعوبة بدون قصد وذلك بدوره يؤدي إلى عدم صدق الاختبار في المجموعة المستهدفة.

عملية الترجمة:

قد تهدد اللهجات في لغة ما صدق تكييف الاختبارات، أي لهجة هي الأهم أو هي الهدف المستخدم في التكييف الذي يمكن تطبيقه داخل اللغة الواحدة؟ يجب البت في هذه المسألة قبل البدء في تكييف الاختبار ويجب استخدامها ضمن المواد

لتدريب المترجمين. إن إحصاء تكرار الكلمات قد يكون قيماً في الحصول على ترجمة اختبار صالح. على وجه العموم من الأفضل ترجمة الكلمات وتعابير مكونة من عدة كلمات بذات التواتر في اللغتين وذلك لمحاولة السيطرة على الصعوبات عبر اللغات. إن المشكلة هي أن لوائح تواتر الكلمات والتعابير ليست متوفرة دائماً وهذا سبب آخر لتفضيل المترجمين الذين لهم اطلاعهم الكامل على كلتا الثقافتين المصدرة والمستهدفة وليس معرفة اللغتين فقط.

تستعمل اللامركزية في بعض الأحيان في تكييف الاختبارات. من الممكن أن لا يكون لبعض الكلمات أو التعابير مرادف في اللغة المستهدفة. حتى إنه من الممكن أن لا توجد تلك الكلمات أو التعابير في تلك اللغة. إن عملية اللامركزية تتضمن مراجعة اللغة الأصلية المترجم منها الاختبار وبذلك يتم استخدام أساس لغوي مترادف في لغة لنسختين المصدر والمستهدفة. إن اللامركزية ممكنة عندما يكون الاختبار الأصلي في مرحلة التحضير في ذات الوقت الذي يتم فيه إنجاز نسخة اللغة المستهدفة. وهذا يكون موجوداً عند إعداد اختبارات التقويم العالمية. وبعض الاختبارات المعتمدة (الاختبارات المعدة من مايكروسوفت) التي أعدت للاستخدام في العالم.

خطط الحكم النقدي لتكييف الاختبارات:

إن الخطتين المفضّلتين في الترجمة هما الترجمة المبكرة والترجمة الراجعة، إن خطة الترجمة المبكرة هي أن مترجماً واحداً أو من الأفضل عدة مترجمين يقومون بتكييف الاختبار من لغة المصدر إلى اللغة المستهدفة. عندئذ يجري الحكم على تعادل النسختين المترجمتين من الاختبار من قبل مجموعة ثانية من المترجمين. يمكن إجراء مراجعة على نسخة الاختبار في اللغة المستهدفة لتصحيح بعض الأخطاء التي وجدها الفريق الثاني من المترجمين. في بعض الأحيان وكخطوة أخيرة يقوم شخص ثالث ليس بالضرورة أن يكون مترجماً بتحرير الاختبار بجعل اللغة أكثر سلاسة لأنه

في بعض الأحيان يحصل بعض التفكك في اللغة أثناء الترجمة، التي يقوم بها عدة مترجمين أو مجموعات للنسخة الواحدة.

إن الميزة الأساسية لخطة الترجمة المبكرة هو أن الحكم يصدر مباشرة على النسخة الأصلية من الاختبار والنسخة المترجمة. إن صدق الحكم على تكافؤ النسختين يُعزز بوجود مجموعة صغيرة من الممتحنين ليزودوا المترجمين بملاحظاتهم عن الاختبار والإرشادات، المحتوى أو الشكل العام من الممكن القيام بذلك في دراسات تدعى "فكر بصوت عال".

أما نقطة الضعف الأساسية في خطة الترجمة المبكرة فهي مرتبطة مع المستوى العالي من الاستنتاجات التي يقوم بها المترجمون عن التكافؤ بين نسختي الاختبار. هناك نقاط ضعف أخرى مثل (أ) قد يكون لدى المترجم مهارة في إحدى اللغات أكثر من الأخرى. (ب) أن الحكم على تكافؤ الاختبار يقوم به أشخاص ثنائيو اللغة وبهذا يمكن أن تكون نظرتهم التخمينية متركزة على معرفتهم لكلتا اللغتين، (ج) أن المترجمين قد يكونون متعلمين أكثر من الطلاب ذوي اللغة الواحدة الذين يتقدمون للاختبار وبذلك يخفق المترجمون في إدراك بعض الصعوبات التي تواجه الممتحنين، (د) أن الأشخاص الذين يطورون الاختبار ليسوا في موقع يستطيعون به أن يصدروا أحكاماً عليه بأنفسهم.

إن خطة الترجمة الراجعة هي المعروفة والأكثر شيوعاً في حفظ الحكم النقدي للاختبارات. في نسختها الأكثر شيوعاً، يقوم واحد أو أكثر من المترجمين بتكييف اختبار من اللغة الأصلية إلى اللغة المستهدفة، ثم يقوم مترجمون مختلفون بترجمة الاختبار إلى اللغة الأصلية. ويجري مقارنة النسختين، الأصلية والمعادة الترجمة ويجري تقويم التكافؤ بينهما. إذا كانت النسختان متشابهتين يجري الموافقة على التكافؤ بينهما، إن خطة الترجمة الراجعة يمكن استخدامها لاختبار عام لكل من نوعية الترجمة ولكشف بعض المشكلات التي ترافق عملية ترجمة أو تكييف غير

جيدة، يفضل الباحثون تلك الطريقة بشكل خاص لأنها تزودهم بفرصة للحكم على النسختين الأصلية والمترجمة للاختبار وبذلك يستطيعون تكوين رأيهم الشخصي عن عملية التكييف. وهذا ليس ممكناً في الترجمة المبكرة إلا إذا كانت لهم المهارة في اللغتين.

بالرغم من أن الترجمة الراجعة لها فضائلها وتستطيع تعيين المشكلات في عملية التكييف، لكنها نادراً ما تستطيع توفير الدليل الكافي لدعم صدق استخدام الاختبار المكيف. إن الدليل على أن تكافؤ الاختبار الذي توفره خطة الترجمة الراجعة هو واحد فقط من عدة أنواع من الأدلة التي يجري تصنيفها في دراسة تكييف الاختبار. أحد مواطن الضعف هو أن المقارنة بين نسخ اختبار في لغتين أو أكثر تجري في اللغة الأصلية فقط. من الممكن أن يكون تكييف الاختبار ليس جيداً بالرغم من أن دليل مقارنة الاختبار الأصلي واختبار الترجمة الراجعة يدل على غير ذلك. قد يحدث ذلك إذا استخدم المترجمون جميعاً قواعد واحدة للتأكد من أن الاختبار المترجم متشابه مع الاختبار الأصلي. هنالك نقطة ضعف أخرى هي أن كون التكييف ضعيف يعود إلى احتفاظه بأوجه غير مناسبة من اختبار اللغة الأصلي مثل بنية القواعد الواحدة والتهجئة. قد تسهل تلك الأخطاء عملية الترجمة الراجعة ولكن تلك الخطة قد تخفي نقاط ضعف مهمة في نسخة اختبار اللغة المستهدفة على سبيل المثال من الممكن الاحتفاظ بكلمة "Icehocky" عند ترجمة اختبار إلى الإسبانية وعندئذ يكون من السهل القيام بالترجمة الراجعة.

لسوء الحظ يمكن أن تكون تلك الرياضة لا معنى لها لكثير من الأشخاص الذين يتكلمون الإسبانية فقط وبذلك من الممكن حصول تدني صدق نسخة الاختبار باللغة الإسبانية.

أخيراً بالإضافة إلى ما تقدم، فإن خطط حكم نقدي أخرى لها بعض العوائق لأن بعض النماذج من المجموعة التي يجرى عليها الاختبار لا تقوم بالاختبار حقيقة

في ظروف اختبارية صحيحة (أو أي ظروف أخرى). هنالك دليل متوافر يوحي أن الذين يقومون بمراجعة الاختبار غير قادرين على تعيين الأخطاء في بنود الاختبار ولذلك يجرى اختبار ميداني بشكل دوري لبنود الاختبار قبل استخدامها. يجب أن تاختبار الاختبارات المكيفة ميدانياً أيضاً لكشف المشكلات التي لم يكتشفها المترجمون حتى عند استخدام مترجمين جيدين وخطط الترجمة المثلى معاً (انظر هاميلتون وبانسولا، 1999).

مخططات جمع المعطيات وتحليلها لإقامة تكافؤ البنود والاختبار:

هنالك ثلاث خطط شائعة الاستعمال في جمع المعطيات لتقويم التكافؤ في بنية الاختبار وبنوده في لغات مختلفة، فيما يلي تقويم تلك الخطط:

1- إجراء الطلاب "الذين يجرون الاختبار" الاختبار في اللغتين اللغة الأصلية واللغة المستهدفة. إن ميزة تلك الخطة أنه من الممكن ضبط الاختلاف في ميزات الطلاب يمكن جمع بنود وإحصائيات مختلفة عن (الاختبار من إداريين كل نسخة اختبار ومقارنتها لتقرير التكافؤ. على كل تعتمد هذه الخطة على الافتراض أن الطلاب ثنائيي اللغة لديهم مهارة متساوية في كلتا اللغتين. هذا لا يحدث غالباً في مجموعة كبيرة من الطلاب (تشيكو، 1987، روسانسكي، 1979) وبذلك يجب ملاحظة ذلك الافتراض حسب الإمكان.

لجعل خطة جمع المعطيات ثنائية اللغة صالحة من الأفضل تطبيقها مع خطة جمع معطيات ثانية وبذلك يمكن تقصي صدق متقارب للنتائج.

إن المشكلة الثانية في خطة جمع المعطيات هي أن نتائج الاحصائيات التي تم الحصول عليها من جمع المعطيات لا يمكن تعميمها ضمن مجموعة الطلاب ذوي اللغة الواحدة؛ لأن مجموعة ثنائيي اللغة (بشكل عام) مختلفين في عدة طرق عن نظرائهم (الطلاب ذي اللغة الواحدة) (هامبلتون، 1993).

في إحدى الدراسات التي أجراها هلن، دراسكو وكوموكار (1982) في "الدليل الوصفي الوظيفي" اكتشف هؤلاء الباحثون أن 4٪ من البنود في مقياس المواقف تم

تصنيفها على أنها ترجمة بشكل سيئ في أحد النماذج للطلاب ثنائيي اللغة. وتم تصنيف 30٪ من البنود على أنها ترجمة سيئة عندما استخدمها طلاب ذوو لغة واحدة (لغة المصدر واللغة المستهدفة).

هنالك خطة مختلفة عن الخطة الثنائية اللغة، التي لها ذات الحدود ولكنها أسهل في التطبيق، تتضمن طلابًا ثنائيي اللغة تم اختيارهم عشوائياً لأخذ أحد الاختبارات في تلك الحالة تظهر فاعلية تكافؤ خطة اختبار المجموعة العشوائي.

2- أخذ طلاب أحاديو اللغة الاختبارين، الاختبار الأصلي والاختبار ذا الترجمة الراجعة.

تتضمن هذه الخطة الإشراف على إجراء اختبار لغة المصدر والاختبار ذي الترجمة الراجعة على الطلاب أحاديي اللغة (لغة المصدر). يتم التعرف على تكافؤ البنود بمقارنة أداء المشاركين في كل من الاختبارين في كل بند. يمكن استخدام عملية تحليل العوامل على المعطيات المجموعة من كل اختبار ومقارنة بناء تلك العوامل. إن ميزة تلك الخطة هي أنه باستخدام نموذج واحد من المشاركين لا يكون هناك أي خلط في النتائج بسبب صفات الطلاب (هامبلتون وبولورك، 1991).

هنالك عيبان رئيسان يضعفان فائدة استخدام خطة جمع المعطيات:

أولاً : لا تجمع معطيات تجريبية من اختبار اللغة المستهدفة بمعنى أنه لا يستخدم طلاب أحاديو اللغة في اللغة المستهدفة مع أن الهدف من البحث هو تطبيق النتائج على نسخة الاختبار باللغة المستهدفة وعلى طلاب اللغة المستهدفة أحاديي اللغة.

ثانياً: لا تكون النتائج التي حصل عليها مستقلة لأنه لا يمكن استبعاد نتائج التعليم من الإشراف على الاختبار الأول في لغة المصدر الأصلية ولا تأثير التعليم على أداء الطلاب في اختبار الترجمة الراجعة. يستطيع التوازن تخفيض أهمية تأثير التطبيق ولكنها تصعب التحليل.

3- يأخذ طلاب لغة المصدر أحاديو اللغة اختباراً في لغة المصدر، ويأخذ طلاب اللغة المستهدفة أحاديو اللغة الاختبار في اللغة المستهدفة.

إن خطة جمع معطيات مناسبة تكون بأن يأخذ طلاب أحاديو اللغة اختبار لغة المصدر وأن يأخذ نموذج ثان من الطلاب الأحاديي اللغة اختيار اللغة المستهدفة. وعادة لا يمكن الاحتفاظ بافتراض مساواة توزع الإمكانيات بين المجموعتين. لحسن الحظ لا توجد ضرورة لهذا الافتراض إذا جرى القيام بالتحليلات حسب نظرية الإجابة عن كل بند (IRT) هيكلية الاختبار (إليس، 1989، 1991، اليس وكيميل، 1992، هامبلتون، سواميناثن وروجرز 1991، فان دي فيفر وليونغ، 1997، 2000) أو إذا جرى القيام بالتحليلات حسب دراسات التكافؤ باستخدام ربط الإجراءات (هولاند ووينر، 1993). إن ميزة تلك الخطة هي أن نماذج المصدر والمجموعة المستهدفة قد جرى استخدامها في التحليل، وبذلك تكون نتائج تكافؤ الاختبار في اللغتين عامة في تلك المجموعات.

إن أحد أهم الاستقصاءات لترشيح إنجاز تكافؤ البنود هي دراسات البنود الموجه (هامبلتون وغيره، 1991، سيرغي وآلالوف، 2003). إن مقارنة إحصائيات البنود في اختبارات اللغتين (أو أكثر) كانت هي المسيطرة على أي فروقات الاستطاعة في المجموعتين (انظر هامبلتون وكانجي، 1995) إن البنود التي تحتوي الفروقات قد عُرفت ودرست بدقة لتحديد التعليل الممكن لتلك الاختلافات (انظر أركيكان، 2002).

إحدى تلك التعليلات هو التكييف السيئ. لسوء الحظ فإن هذه الدراسات غير قادرة على فصل الفروقات الثقافية عن مشكلات التكييف ولكن في أكثر الأحيان يكشف بشكل عام عن مشكلات محتملة في الاختبار المكيف. إن تحليل البنود الموجه تنتج عن نظريات اختبارات قديمة وجديدة ويمكن استخدامها في كل من استجابة المعطيات الثنائية ومعطيات استجابة ملحقة (سيرغي وآلالوف، 2003).

عوامل تؤثر في شرح النتائج:

في دراسات عبر الثقافات على مقياس عالمي، فإن الهدف من الاختبار هو أن يؤمن الأساس لإجراء المقارنات بين مجموعات مختلفة الثقافات واللغات لكي نستطيع فهم المفارقات والتشابه الموجودة (هاميلتون 1990، 2002). في بعض الأحيان يكون الاهتمام في المتغيرات المتشابهة وفي أحيان أخرى يكون التركيز على تقييم متغيرات الشخصية أو على معلومات عامة (نوعية الحياة، الصحة).

تأمل تلك الدراسات أن تستخدم تلك النتائج في البحث عن طرق لمقارنة المجموعات وفهم المفارقات بينها. لا يجب أن تستعمل الدراسات عبر الثقافات لدعم المناقشات عن التفوق الفريد لقومية وكأن دراسة الفروقات الدولية تعادل سباق جياد فيها الرابحون والخاسرون (وستبيري، 1992). في أحسن الحالات توفر تلك الدراسات لمحة عن المفارقات الموجودة، وتؤمن فقط أساساً محدداً لتفسير النتائج في هذا المفهوم لكسب فهم أكثر عند تفسير النقاط، ويجب أن نأخذ بعين الاعتبار عوامل أخرى خارجية ليس لها علاقة بالاختبار أو تقدير المعايير خاصة بجنسية محددة ، المناهج الدراسية، المستويات والسياسة التعليمية، الغنى، مستوى الحياة، القيم الثقافية إلى ما هنالك. من الممكن أن يكون كل ذلك عوامل أساسية في تفسير درجات صحيح عبر الثقافات/ اللغات ومجموعات دولية. من الطبيعي أن تناقش عينات من العوامل التي يجب التفكير بها عند تفسير نتائج الاختبار لمجموعات عبر اللغات والثقافة لاحقاً.

التشابه في المناهج الدراسية:

في نطاق وجود اختلافات في المناهج الدراسية، فإن مقارنة الأداء بين ثقافات مختلفة ستكون غامضة إذا لم تؤخذ تلك اختلافات بعين الاعتبار، لاحظ وستبيري (1992) أن نتائج "الدراسة العالمية الثانية للرياضيات" (SIMS) تشير إلى أن أداء الطلاب الأميركيين كان ضعيفاً في كل المراحل في كل مادة في الرياضيات التي

جرى تغطيتها في الاختبار. عند مقارنة أداء الطلاب اليابانيين والأمريكيين لوحظت فروق كبيرة في المناهج الدراسية في البلدين. على كل حال في نطاق المنهاج الدراسي المتشابه، لاحظ وسبيري أنه ليس هناك أي اختلاف في أداء الطلاب في البلدين.

إن أهمية تحليل الاختلافات في المناهج الدراسية واضح في دراسات المقارنة الدولية للأداء، ولذلك وبالرغم من كل المعارضات (بسبب الجهد والتكلفة) صنفت معطيات استبيان مكثفة مع معطيات الاختبار في كل دولة مشاركة.

دوافع الطلاب:

تساءل وينر عما إذا كان يمكن فصل الخبرة الظاهرة المقاسة في الاختبار عن الدوافع. لاحظ أن كل الطلاب الذين تم اختيارهم (بشكل عشوائي) للمشاركة في اختبار "دراسة التقويم الدولي للتقدم التعليمي" في كل دولة قد شعروا بالفخر لأنه قد تم اختيارهم لتمثيل مدارسهم ودولتهم، وبذلك أُلقيت عليهم مسؤولية الأداء الأفضل. وفي الجانب الآخر كانت المشاركة في الدراسة المقارنة العالمية للطلاب في دولة أخرى عبارة عن نشاط آخر ليس بتلك الأهمية لأن درجاتهم لم تكن متيسرة. كان الاختبار لهؤلاء الطلاب "رهاناً ضعيفاً".

إن تفسير اختلافات الأداء بين دول ذات طلاب ذوي دوافع ودول ليس لطلابها أي دوافع دون أي اعتبار لمتغيرات الدوافع في أداء الاختبار سينتج عنها إساءة تفسير خطيرة للنتائج.

العوامل السياسية الاجتماعية:

إن معنى وتفسير الدرجات تختلف حتى وإن كانت الدرجات متشابهة، فكر في إجراء مقارنة درجات الاختبار بين طلاب من دول متطورة ودول نامية أو مجتمع صناعي ومجتمع ريفي. في ذلك المحيط، لا يكون أداء الطلاب ذا علاقة باستطاعتهم على الإطلاق، بل قد يكون الأداء انعكاساً لعدم القدرة للحصول على مصادر كافية أو النوعيات المختلفة للخدمات التعليمية المتاحة.

إن النقطة الأساسية هي أنه للحصول على تفسير ذي دلالة للنتائج يجب حسبان الحقائق الاجتماعية، السياسة والاقتصادية المختلفة التي تواجه الأمم كما يجب حساب الفرص التعليمية المتاحة في ظل تلك الحقائق (اولدا، 1981). لذلك من المهم أن يكون المسؤولون عن تطوير الاختبارات وصناعة سياستها مطلعين على تلك الموضوعات الثقافية بحد ذاتها والتي يمكن أن تؤثر في أداء الاختبار.

إرشادات عملية لتكييف الاختبارات:

من المؤكد أن الكتابات التقنية لتوجيه عملية التكييف غير كاملة (من وجهة النظر التقديرية) ومتفرقة في كثير من المطبوعات العالمية، التقارير، والكتب. لم يكن هناك أي مصدر كامل يستطيع الممارسون الرجوع إليه لبعض النصائح ولم تتكون أي مجموعة من الإرشادات لتكييف الاختبار (هامبلتون، 1994، فان دي فيفر وهامبلتون، 1996) ولم تكن طرق القياسات المعقدة (نماذج بنود الاستجابة ونماذج توازن التركيبات) التي تساعد في إنشاء عملية معادلة الدرجات التي تم الحصول عليها في الاختبارات المكيفة لاستخدامها في اللغات والثقافات المختلفة معروفة للباحثين الذين يقومون بتكييف الاختبارات حتى وقت قريب (هولن، 1987). لكن كما هو واضح في فصول هذا الكتاب (انظر هامبلتون ودي جونغ، 29003) فإن الوضع قد تحسن جوهرياً منذ أوائل التسعينيات. في الواقع كان الهدفان من مؤتمر ITC الذي عقد في جورج تاون في الولايات المتحدة عام 1999 أولاً: جمع باحثين من العالم لتبادل المعرفة والخبرة في تكييف الاختبار. وثانياً: تقديم النسخة النهائية للدليل الموجز لتكييف الاختبارات الذي أشرفت عليه الهيئة العالمية للاختبارات (ITC)، إن الهدف من هذا القسم من الفصل هو وصف الدافع الذي جعل ITC تقوم بإعداد ذلك الدليل، لتوفير بعض المعلومات عن خلفية إعداده وأخيراً وصف الخطوط الرئيسة الاثنتين والعشرين وعرض أسباب استخدام كل واحد منها على حدة.

في ظل الحقيقة أن "المراهنة القوية" ترافق غالباً نتائج الدراسات التعليمية للأداء عبر الثقافات أو المقارنات الدولية (انظر، الاهتمام على مستوى عال اليوم لدعم دراسات الأداء المقارنة الدولية مادياً). فإن الحاجة إلى دليل عملي جيد طور من قبل خبراء لاستخدامه في تكييف الاختبارات وتأسيس نظام معادلة الدرجات النهائية بدت واضحة للهيئة الدولية للاختبارات (ITC) منذ 1992. كانت المقاييس التقنية أو الخطوط الرئيسة لتقويم ممارسة تطوير الاختبارات، جدارة التقويم، إن صدق التقويم متوفرة في بلدان كثيرة (انظر، Aera, APA, & NCME, 1985, 1999)، ولكن لم يكن هناك أي اهتمام لإعداد دليل لتكييف الاختبارات وتأسيس نظام معادلة الدرجات النهائية. على سبيل المثال في مقاييس اختبارات (AERA, APA & NCME) التي نشرت في 1985 (التي كانت من أهم مقاييس الاختبارات في الولايات المتحدة حتى نشرت معايير الاختبارات عام 1999) ثلاثة مقاييس فقط تناولت موضوع تكييف الاختبارات بشكل مباشر. في كندا، بلد ثنائي اللغة، ثلاثة معايير فقط تناولت تكييف الاختبارات في "مقاييس الاختبار في الجمعية النفسية الكندية (كانت المقاييس متوفرة عام 1993).

تعالج الهيئة الدولية للاختبارات ذلك النقص بإعداد مجموعة من الخطوط الرئيسية لتكييف الاختبارات (انظر هامبلتوت، 1994، فان دي فيفر وهامبلتون، 1996)، والذي يشار إليه "دليل الهيئة الدولية للاختبارات لتكييف الاختبارات"، يحدد الجدول 1.1 هوية الهيئات الثمانية التي ساهمت في إعداد الدليل. الجدول 2 يحدد هوية أعضاء اللجنة التي قامت بالعمل لمدة ثلاث سنوات لإعداده.

إن دليل تكييف الاختيار منظم في أربعة أقسام: المحتوى، تطوير وتكييف الاختبار، إدارة الاختبار وتفسير وتوثيق الدرجات النهائية.

كان رأي اللجنة التي أعدت الدليل أنه سيكون مريحاً أكثر في الاستخدام إذا نظم في فئات ذات هدف أساسي. تناولت الخطوات الرئيسية في منشأ المحتوى

تكافؤ المفهوم في لغة المجموعة المستخدمة للاختبار. تتضمن فئة تطوير وتكييف الاختبار موضوعات تظهر في عملية التكييف، بداية باختيار المترجمين إلى طرق الإحصائيات لتحليل المعطيات التجريبية في تقصي تكافؤ الدرجات النهائية. أما الفئة الثالثة، إدارة الاختبار، فإنها تناولت طرق إدارة الاختبار مع مجموعات متعددة اللغات وبداية باختيار الإداريين بين البنود في الاختبار إلى تحديد مدة الاختبار. الفئة الرابعة تختص بتفسير وتوثيق الدرجات النهاية. كالعادة، أعد الباحثون توثيقاً قليلاً جداً عن عملية التكييف لإثبات صدق الاختبار المكيف وكانت الأخطاء في تفسير الدرجات النهائية للاختبارات المتعددة اللغات شائعة جداً. إن دليل ITC لتكييف الاختبارات يعالج كل الأمور في ذلك المجال.

الجدول ١,١

الجمعيات المشاركة في إعداد الدليل العالمي لتكييف الاختبارات:

الهيئة الدولية للاختبار	(ITC)
الجمعية الأوروبية للتقويم النفسي	(EAPA)
مجموعة ناشري الاختبارات الأوروبية	(ETPG)
الجمعية الدولية النفسية عبر الثقافات	(IACCP)
الجمعية الدولية لعلم النفس التطبيقي	(IAAP)
الجمعية الدولية لتقويم الأداء التربوي	(IAA)
الجمعية الدولية للاختبارات اللغوية	(ILTA)
الاتحاد الدولي للعلوم النفسية	(IuPsyS)

الجدول ٢,١

أعضاء اللجنة والمنظمات التي يمثلونها:

رئيس اللجنة

رونالد ك. هامبلتون	(ITC)

جامعة مستشوست، أمهرست، الولايات المتحدة

أعضاء اللجنة:

جلين بدجل (ITC)

جمعية الممرضات الكندية، كندا

روب فيلثام (ETPG)

نفرنلسن، إنجلترا

روكيو فرناندز بالستيروز (EAPA)

جامعة أوتونوما، اسبانيا

جون هـ. أ. ل دي جونج (ILTA)

سيتو/ هولندا

أنجرن مونك (IEA)

الإحصائيات السويدية/ السويد

جوزيه مونيز (ITC)

جامعة أوفيدو، إسبانيا

يب بورتينجا (IACCP)

جامعة تيلبيرغ، هولندا

اسيك سفاسير (IuPsyS)

جامعة هاستي، تركيا

تشارلز سبيلبرغر (IAAP)

جامعة جنوب فلوريدا، الولايات المتحدة

فون فان دي فيفر (ITC)

جامعة تيلبيرغ، هولندا

جانس من. زعل (ITC)

GITP الدولية، هولندا

زميل باحث

أنبيل كانجي

جامعة مساشوست، امهرست، الولايات المتحدة

وقد أقرت الهيئة الدولية للاختبارات التعريف التالي لدليل تكييف الاختبارات: "إن دليل تكييف الاختبارات هو مزاولة مهنة تعد مهمة لإدارة وتقويم التكييف أو تطور مواز للاختبارات النفسية والتربوية للاستخدام في مجتمعات مختلفة". إن الخطوط الرئيسة المقدمة من الهيئة الدولية للاختبارات يمكن تلخيصها في النقاش التالي في جدول 3.1 (طبعت كمسودة من قبل في هامبلتون، 1994 وفان دي فيفر وهامبلتون، 1996). تلك الخطوط العريضة موجودة في هذا الفصل مع بعض التعديلات البسيطة في التقرير الأخير للجنة (ITC, 2001) تم وصف كل خط من تلك الخطوط (أ) الأسباب المنطقية لحصر تلك الخطوط، (ب) الخطوات لتطبيق تلك الخطوط، (ج) لائحة الأخطاء الشائعة، (د) ومجموعة من المراجع. هناك نموذج كامل لأحد تلك الخطوط في الجدول (4.1) ويليه وصف مقتضب لكل من الخطوط والأسباب المنطقية لوجودها ضمن اللائحة.

المضمون:

1- ج. أ: يجب تقليل تأثير الاختلافات الثقافية غير الضرورية في أسباب الدراسة الأساسية إلى الحد الأدنى.

الأسباب/ الشرح: هنالك الكثير من العوامل المؤثرة في المقارنة عبر اللغات/ الثقافات التي يجب أخذها بعين الاعتبار عند مقارنة مجموعتين أو أكثر من ذوي خلفية لغوية/ ثقافية مختلفة، خاصة عند تطوير اختبار أو تكييفه أو عند تفسير الدرجات النهائية. على كل حال، من الضروري أن لا يتم التفكير بهذه العوامل فقط بل يجب القيام بخطوات عملية، إما بالإقلال منها أو حذف تأثيرات العوامل غير المرغوب بها في أي مقارنة عبر اللغات/ الثقافات. على سبيل المثال المستويات المختلفة لدوافع المشاركين في الاختبار في بحث حديث للتقويم الدولي للتطور التربوي هو أحد الأسباب لاختلاف أداء المشاركين في الاختبار في تلك الدول (وينر، 1993).

2- ج. د: يجب تقييم تداخلات التراكيب التي يجري تقديرها في الاختبار ضمن المجموعات التي تجري الاختبار.

الأسباب/ الشرح: لا تتوقف الاختلافات الموجودة بين الثقافات واللغات المختلفة للمجموعات على اختلافات تقاليد، قواعد السلوك والقيم ولكن على رؤية العالم وترجماتها، وبذلك من الممكن أن يفسر التركيب ذاته ويفهم بطرق مختلفة كلياً في ثقافتين مختلفين. على سبيل المثال: إن مفهوم "الذكاء" موجود في كل الثقافات تقريباً ولكن في الثقافات الغربية يرتبط هذا المفهوم مع التقديم السريع للإجابات بينما يرتبط في الثقافات الشرقية مع التفكير، الاستجابة، وقول الشيء الصحيح (لونر، 1995)، يجب على الباحثين التأكد من أن التركيب الذي يقاس في اختبار مجموعة من ثقافة/ لغة المصدر الأساسي يمكن أن يوجد بذات الشكل والتواتر في الثقافات الأخرى التي تجري دراستها.

جدول (3-1)

دليل الهيئة الدولية للاختبار ITC لتكييف الاختبار:

ج. 1 (1) يجب تخفيض تأثيرات الاختلافات الثقافية التي ليس لها أهمية للهدف الأساسي للدراسة إلى أقل حد ممكن.

ج. 2 (2) يجب تقويم مقدار التشابك في البنية المقاسة في اختبار المجموعات المطلوبة.

تطور الاختبار وتكيفه

د. 1 (3) يجب على الذين يقومون بعملية التطوير والناشرين التأكد من أن عملية التكييف تأخذ بعين الاعتبار الاختلافات اللغوية الثقافية للمجموعات المقصودة.

د. 2 (4) يجب على الذين يقومون بعملية التطوير والناشرين إقامة الأدلة بأن اللغة المستخدمة في تعليمات الاختبار، إرشادات الدرجات، وفي البنود مناسبة للغة وثقافة جميع المجموعات التي ستقوم بالاختبار.

د. 3 (5) يجب على المطورين/ الناشرين إقامة الدليل على أن اختيار أسلوب الاختبار، هيكلية البنود، قواعد الاختبار وإجراءات أخرى مألوفة للمجموعات المقصودة.

د. 4 (6) يجب على المطورين/ الناشرين إقامة الدليل على أن محتوى البنود والمواد الأخرى (المنبهة) مألوفة للمجموعات المقصودة.

د. 5 (7) يجب على المطورين/ الناشرين جمع دليل النقد العقلاني، اللغوي والنفسي، لتحسين دقة عملية التطوير وجمع الدليل على تكافؤ كل النسخ في اللغات المختلفة.

د. 6 (8) يجب على المطورين/ الناشرين التأكد من أن خطة جمع المعطيات تسمح باستخدام أساليب إحصائية مناسبة لإقامة تكافؤ البند والبنية في نسخ الاختبار في اللغات المختلفة.

د. 7 (9) يجب على المطورين/ الناشرين استخدام أساليب إحصائية مناسبة كي يستطيعوا (أ) إقامة التكافؤ في لغة النسخ المختلفة للاختبار.

(ب) التعرف على العناصر التي يمكن أن تحدث مشكلات أو تكون غير مناسبة لإحدى المجموعات المشاركة.

د. 8 (10) يجب على المطورين/ الناشرين توفير معلومات عن صدق الاختبار المكيف للمجموعة المقصودة.

د. 9 (11) يجب على المطورين/ الناشرين توفير الدليل الإحصائي عن تكافؤ البنود لكل المجموعات المقصودة.

د. 10 (12) يجب عدم ربط بنود الاختبار المكيف غير المتكافئ للمجموعة المقصودة مع الدرجات العامة للمقياس. على كل حال يمكن أن تكون تلك البنود مفيدة لإعطاء تقرير عن درجات كل مجموعة على حدة.

الإدارة

أ. 1 (13) يجب أن تكون أوجه المحيط/ البيئة التي تؤثر في إدارة الاختبار متشابهة إلى أقصى حد عبر المجموعات التي تجري الاختبار.

أ. 2 (14) يجب على مطوري الاختبار والإداريين محاولة توقع المشكلات التي يمكن حدوثها واتخاذ الإجراءات المناسبة لمعالجة كل المشكلات وذلك بإعداد مواد وإرشادات مناسبة.

أ. 3 (15) يجب على الإداريين أن يدركوا العناصر المتعلقة بالمواد المحفزة، الإجراءات الإدارية وطرق الاستجابة التي قد تخفض صدق الاستنتاجات التي تم الحصول عليها من الدرجات.

أ. 4 (16) يجب أن تكون لغة إرشادات الإداريين في كلتا اللغتين، لغة المصدر واللغة المستخدمة في الاختبار مقللة من المتغيرات غير المرغوب بها عبر المجموعات.

أ. 5 (17) يجب أن يعطي كتيب الاختبار وصفاً دقيقاً لكل أوجه الاختبار وطريقة إدارته التي تتطلب الدقة في تطبيق الاختبار في محيط ثقافي جديد.

أ. 6 (18) يجب أن لا يكون الإداريون فضوليين، ويجب أن تكون العلاقة بين الإداريين والذين يجرون الاختبار قليلة إلى أدنى حد. يجب اتباع القواعد الواضحة التي جرى وصفها في كتيب الاختبار.

التوثيق/ تفسير المقياس

1 . 1 (19) عندما يكيف الاختبار للاستخدام في مجموعة أخرى، يجب توثيق التغيرات مع الدليل الذي يدعم تكافؤ النسخة المكيفة للاختبار.

1 . 2 (20) يجب أن لا تؤخذ اختلافات الدرجات لنماذج المجموعات التي قامت بالاختبار كقيمة ظاهرية. على الباحث مسؤولية إقامة الدليل على معنى الاختلافات من الأدلة التجريبية.

1 . 3 (21) يمكن إقامة المقارنات عبر المجموعات فقط على مستوى الثوابت التي أقيمت للمقياس الذي تنقله الدرجات.

1 . 4 (22) يجب على المطورين توفير معلومات محددة عن الطرق التي يمكن أن تؤثر فيها المفاهيم الاجتماعية/ الثقافية والبيئية للمجموعات على أداء الاختبار، كما يجب عليهم اقتراح إجراءات لتحليل تلك التأثيرات في تفسير النتائج.

الجدول 1- 4

نموذج للخطوط الأساسية د. 1 في شكله العام

الخط الرئيس د. 1: شروط أساسية عامة ومهنية

يجب على المطورين/ الناشرين التأكد أن عملية التكييف تأخذ بعين الاعتبار الاختلافات اللغوية والثقافية للمجموعات المقصودة.

الأسباب/ التفسير:

إن خبرة وتجربة المترجمين يمكن أن تكون أكثر الأوجه أهمية في عملية تكييف الاختبار بأجمعها لأنهم يؤثرون بشكل كبير على جدارة وصدق الاختبار. (براكون وبارونا، 1991). على سبيل المثال: عمد المترجمون الذين يتمتعون بالمعرفة التقنية أو تفاصيل الحقل المترجم إلى الترجمة الحرفية التي يمكن أن تشكل فهمًا خاطئًا

المجموعة المستهدفة وتهدد صدق الاختبار (هامبلتون وكانجي، 1995). بناء على ذلك فإن اختيار مترجمين مؤهلين جيدين عامل مهم في عملية تكييف الاختبار. بالرغم من أن الخبرة في اللغتين شرط أساسي فإن المعرفة والتجربة (أ) للثقافتين، (ب) محتويات الاختبار، (ج) ومبادئ تطوير الاختبار خاصة كتابة البنود، يجب أن تكون أحد الشروط الأساسية في اختيار تدريب المترجمين. هنالك حاجة لوجود فريق عمل للقيام بعملية تكييف دقيقة؛ لأنه من الواضح أنه ليس من المعقول أن يتمتع فرد واحد من المترجمين بكل تلك الشروط الأساسية.

1- التأكد كحد أدنى أن المترجمين مؤهلون وذوو خبرة في كل من اللغتين، لغة المصدر واللغة المستهدفة وثقافتهما (بتشر وغراسيا، 1987). إن الشهادات أو الخبرة السابقة شروط أساسية مهمة. على سبيل المثال: ليس بالإمكان افتراض أن ثنائيي اللغة متمكنون في اللغتين أو مطلعون على كلتا الثقافتين بشكل متساو.

2- إن معرفة موضوع المادة شرط أساسي لأي مترجم يقوم بتكييف اختبارات إذا لم تكن له معرفة بموضوع المحتوى على الأقل فإن دقة موضوع المحتوى قد تفقد. يجب أن يكون للمترجمين الذين ليس لهم اطلاع على معرفة معينة في حقل الترجمة الاطلاع المسبق على موضوع المحتوى كجزء من عملية التكييف.

أين يمكن أن يعيش طائر له أقدام كفيره (كأقدام البط)؟

أ - في الجبال.

ب- في الغابات.

ج - في البحر.

د - في الصحراء.

عندما تُرجم هذا السؤال من الإنجليزية إلى اللغة السويدية، أصبحت "الأقدام الكفين" "أقدام سابحة" وبذلك أصبحت الإجابة واضحة للطلاب السويديين عن

مكان عيش الطائر. إن مترجماً ذا معرفة بقواعد كتابة البنود كان سيلاحظ دون شك الخطأ في الترجمة ويقوم بتعديلها.

4- من المفضل أن يقوم فريق من الاختصاصيين بمشروع تكييف الاختبار (انظر، غريسي، 2003). يجب أن يشارك المترجمون في فريق المشروع وأن يكون لهم دور في تقرير عملية الترجمة، وأن تطلب آراؤهم وأفكارهم وتقبل. إن تلك الطريقة حسب "برسلين" (1986) تستطيع تحسين نوعية التكييف. إن طريقة فريق عمل تساعد في (1) تمكن استخدام طريقة الترجمة الراجعة (انظر الخطة 5 في الأسفل)، (2) تسمح للمترجمين بمقارنة ومناقشة أعمالهم وبذلك تُحسن الصلة والنوعية للترجمة، (3) تساعد في التأكد من أن المعرفة المختصة في كل الحقول المطلوبة يمكن الوصول إليها.

5- إن استخدام فريق عمل من المترجمين الذين يعملون إما منفردين أو ضمن مجموعات صغيرة لتكييف الاختبار هو خطه عمل ممكنة. من الممكن لاحقاً القيام بمقارنة التقويم الفردي للاختبار وحل الاختلافات لتقديم الترجمة الأفضل، هنالك إجراء آخر وهو استخدام مطورين ومترجمين أحاديي اللغة في وقت واحد، يقوم المترجمون بترجمة/ تكييف الاختبار ثم يقوم المطورون أحاديو اللغة بتحرير الاختبار في اللغة المستهدفة، ومن ثم يقوم مترجم/ مطور ثنائي اللغة بتقويمه (برسلين، 1986). لاحظ "برسلين" (1986) أن ميزة تلك الطريقة هي أن مطوري الاختبار أحاديي اللغة يستطيعون إعادة كتابة الاختبار ويجعلونه أكثر وضوحاً واستحساناً لطلاب اللغة المستهدفة، كما أن تلك الخطة تقلل من المواقف، حيث تكون النسخة الجديدة ضعيفة، ولكن من الممكن إغفال تلك المشكلة لأن وجود مترجم عالي الخبرة يستطيع تقديم نسخة ترجمة راجعة ممتازة من نسخة اختبار سيئة في اللغة المستهدفة. أما في حالة وجود مترجم واحد فقط فمن المفضل استخدام مترجم من مجموعة اللغة المستهدفة. في

ذلك الموقف يستطيع المترجم على الأقل مناقشة نسخة اللغة المستهدفة مع شخص آخر من مجموعة اللغة المستهدفة الذي يستطيع الإشارة إلى مواطن المشكلة وربما يقترح التنقيح أيضاً.

أخطاء شائعة:

1- اختيار المترجمين أو الأشخاص من معارف مطور الاختبار (أصدقاء، جيران) لأنهم ثنائيو اللغة قد أثبت أنه اختيار غير ناجح (برسلين، 1986).

2- الفشل في التأكد من اختيار المترجمين ذوي الاطلاع على محتوى الاختبار والذين لهم خبرة في تطوير الاختبار، لقد تم الإبلاغ عن حدوث تلك الأخطاء في كثير من الدول.

3- لم يعط المترجمون الوقت الكافي للقيام بأعمالهم، وقد تم الإبلاغ عن هذه الأخطاء أيضاً.

تطوير الاختبارات وتكييفها:

1. د. 1: يجب على الذين يقومون بتطوير الاختبارات ونشرها التأكد أن عملية التكييف تأخذ بعين الاعتبار الاختلافات اللغوية الثقافية للمجموعات التي تجري الاختبار.

الأسباب/ الشرح إن أسباب تلك الخطة مع الأجزاء الأخرى لتوصيف هذا الدليل تظهر في الجدول (4.1) الذي يستخدم كنموذج للمعلومات المتوفرة عن كل خطة في التقرير النهائي (انظر ITC، 2001).

2. د. 2: يجب على القائمين على تطوير الاختبار ونشره توفير الدليل على أن اللغة المستعملة في إرشادات الاختبار، قواعد الدرجات النهائية والبنود الموجودة في الاختبار كلها مناسبة لجميع الثقافات ولغات المجموعات التي يستهدفها الاختبار.

الأسباب/ الشرح: إن أحد أسباب سوء تكييف اختبار الدراسة عبر الثقافات هو وجود خطأ في نسخة الاختبار في لغة المصدر؛ وذلك يسبب صعوبة في التكييف. وهناك سبب آخر هو أنه من الممكن أن تكون المفاهيم، والتعابير والأفكار المستخدمة في لغة اختبار المصدر ليس لها مرادف في اللغة المستهدفة. إن أحد الأسباب الكثيرة لنجاح الدراسات الحديثة التي قام بها TIMSSو OECD / PISA هو الجهد الفعلي الذي بذل في تطوير الاختبار في لغة المصدر مع وضوح التنظيم ووجود مواصفات الاختبار وبعض بنود للتطوير والاختبار الميداني، وبعض الفعاليات المرافقة لتطوير اختبار مناسب.

من الضروري أيضاً التأكد أن المفردات المستعملة في اختبار متعدد اللغات متشابه من حيث درجة صعوبة الكلمات، نصوص القراءة، استعمال القواعد، أسلوب الكتابة، التنقيط، يجب الحذر عند استعمال الاختبار لتقويم مقدرة المشاركين في الكتابة والقراءة (الصغار والبالغين).

3. د. 3: يجب على المطورين والناشرين توفير الدليل على أن اختيار الأسلوب، الشكل العام، وكل الخطوات المتبعة في الاختبار مألوفة للمجموعة المستهدفة.

الأسباب/ الشرح بعض الأشكال العامة (أسئلة عديدة الإجابات، المقالة، 5 درجات مقياس التقدير) والاصطلاحات والإجراءات في إعطاء الإرشادات الاختبارية وفي تقديم بنود الاختبار قد لا تكون مألوفة في كل المجتمعات. هنالك اختلاف في الاستعمال اللغوي في إرشادات الاختبار، التنسيق واستعمال الرسوم البيانية، طريقة التقديم (الأقلام، الأوراق، الكمبيوتر). لتحقيق المساواة من الضروري أن يكون كل ما سبق مألوفاً للمجتمعات التي يُجري تكييف الاختبار لها. وهذا يتضمن تطوير مواد علمية مكثفة لتقليص التحيز الناتج عن عدم الألفة في بعض وجوه عملية التقويم.

4. د. 4: يجب على المطورين والناشرين إثبات أن كل بنود المحتوى والمادة المحفزة مألوفة لدى المجموعات المستهدفة.

الأسباب/ الشرح. إذا ثبت أن أي اختبار مكيف هو أسهل أو أصعب قراءة أو فهماً بسبب بعض البنود في المحتوى فإن هذا سيكون مصدر تحيز آخر. في بعض الدول في العالم تستعمل وحدات مختلفة للمقادير، على سبيل المثال الوزن، الطول والنقود. يمكن أن يكون تكييف الاختبار أكثر صعوبة للمجموعات المستهدفة إذا كانت الوحدات المستخدمة غير مألوفة أو إذا كانت هنالك بعض العمليات الحسابية (انظر هامبلتون، يون وسليبتر، 1999) وإذا كان هناك مواد محفزة (أشكال، أرقام، جداول، أو علامات حدود) غريبة عليهم.

5- د. 5: يجب على المطورين الناشرين جمع الدلائل العقلانية، اللغوية والنفسية، لتحسين الدقة في عملية التكييف وجمع الدلائل على تساوي كل نسخ الاختبارات في اللغات المختلفة.

الأسباب/ الشرح. يجب تقييم الأسئلة، التمارين، تقدير المقاييس المرادفة في اللغات والثقافات المختلفة. إن الطرق العقلانية في القيام بالترجمة المترادفة تقوم على قرارات المترجمين أو مجموعات المترجمين. إن الطريقتين المستخدمتين في الترجمة: الخطة المقدمة والخطة الراجعة، اللتين جرى ذكرهما في بداية هذا الفصل، فيهما بعض الأخطاء ولذلك من الصعب جداً أن توفر الخطط العقلانية الدلائل الكافية لصدق الاختبار المكيف.

6. د. 6: على المطورين/ الناشرين التأكد أن خطة جمع المعطيات تسمح باستخدام تقنيات إحصاء مناسبة لإقامة التساوي في الشكل والمحتوى في نسخ الاختبار في لغات مختلفة.

الأسباب/ الشرح. إن خطة جمع المعطيات تعود إلى الطريقة التي جمعت بها تلك المعطيات كي يجري التساوي في نسخ الاختبارات المكيفة. إن أول شروط جمع المعطيات هو أن النماذج يجب أن تكون كثيرة بشكل كبير كي يكون هناك إمكانية الحصول على معلومات إحصائية متوازنة. مع أن هذا الشرط ضروري لأي نوع من الأبحاث إلا أنه مهم بشكل خاص في تقدير صدق الاختبار المكيف لأن الأعداد الكبيرة الكافية للنماذج قد يكون لها دور في جمع المعطيات اللازمة لإثبات التعادل في الاختبار.

إن الخطة في الدراسة التجريبية هي مجموعة وظائف (أ) طبيعة المشاركين (أحادي أو ثنائي اللغة)، (ب) نسخة الاختبار المستخدمة (الأصلية، المكيفة، المكيفة الراجعة) (ج) خطة جمع المعطيات المحددة (تناقش بشكل مفصل في د. 7). قدم سيرغي (1997) نقاشاً عن المعضلات والموضوعات في ربط اختبارات متعددة اللغات مع مقياس عام. قدم ودكوك ومونوز ستاندوفل (1993) نموذج ربط مقياس اختبار مع اختبار عبر اللغات مستخدماً IRT انظر في فصل لاحق في هذا الكتاب).

7. د. 7: على المطورين/ الناشرين تطبيق تقنيات إحصائية مناسبة في (أ) إقامة التساوي في لغة الاختبار المستخدم (ب) تعيين المعضلات أو العناصر التي قد تكون غير مناسبة للاستخدام ضمن إحدى المجموعات.

الأسباب/ الشرح. تقدم التقنيات الإحصائية معلومات مفيدة لتقويم تساوي الاختبارات المطورة في أكثر من لغة (فان دي فيفر ولينونغ، 1997، 2000، فان دي فيفر ونانزر، 1997، انظر فصل لاحق). يجب أن تستخدم تلك التقنيات في توفير إضافات إلى تقنيات المقارنة لكونها قادرة على تعيين بنود الاختبار غير المتوافقة التي لم يتم اكتشافها عند استعمال تقنيات المقارنة. هناك ميزة أخرى وهي أن التقنيات الاحصائية تستخلص

المعلومات مباشرة من المشاركين في الاختبار، من مضمون إدارة الاختبار العقلية وبذلك تكون تلك التقنيات مفيدة جداً في تعيين البنود التي قد تشكل بعض الصعوبات في التطبيق.

8. د. 8: يجب على المطورين/ الناشرين توفير المعلومات عن صدق الاختبار المكيف ضمن المجموعة المستهدفة.

الأسباب/ الشرح. إن الاختبارات الموجودة يجري تطويرها وتوحيدها غالباً للاستخدام في ثقافة واحدة وتُكيف للاستخدام في ثقافة أخرى. يمكن توفير الوقت والنفقات إذا جرى تكييف الاختبارات الموجودة (برسلين، 1986). على كل حال فإن كثيراً من التراكيب تكون غير مفهومة دون بعض التعديلات الأساسية للاستخدام في الثقافات الأخرى. طرحت عدة نماذج في هذا الفصل، الذكاء، نوعية الحياة اليومية، والإنجازات الرياضية. في بعض الأحيان من الممكن أن يكون الاختبار غير جدير بالترجمة وبذلك يمكن توفير الوقت، الجهد والمال، حتى في حال وجود ذلك التركيب في اللغة أو الثقافة الثانية من الممكن وجود تفاوت في المظاهر السلوكية والتفسيرات بشكل واضح (لونر، 1990). يجب جمع الأدلة على صدق التركيب في كل مجموعة تجري الاختبار. كما هو معروف فإن استقصاء صدق التركيب يستغرق وقتاً للتخطيط والتطبيق لكونه شاملاً ويتضمن دراسات ومنهجيات متنوعة منها اختبارات متداخلة، متعلقة بالمقياس، تجريبية، متعددة السمات والطرق (انظر. فان دي فيفر وتانزر، 1997).

9. د. 9: على المطورين/ الناشرين توفير إثبات إحصائي عن تساوي البنود في كل المجموعات المستهدفة.

الأسباب/ الشرح. إن أحد أهم التحليلات الإحصائية في صدق اختبار للاستخدام في لغة/ ثقافة واحدة أو أكثر في دراسة بنود متحيزة أو كما يشار إليها حالياً "دراسة تفاوت أداء البند DIF" (هولند ووينر، 1993، سيرغي وآلالوف، 2003، وعدة فصول في هذا الكتاب). يتطلب مساندة تكافؤ اختبار لمجموعتين ثقافتين مختلفتين وجود إثبات على أن أعضاء المجموعتين لهما الكفاءة الواحدة، يجب عليهم الأداء المتكافئ في كل بند. عندما لا يكون الأداء متساوياً يجب أن يكون هناك سبب وجيه أو يلغى البند من الاختبار. هذا لا يعني عدم وجود اختلاف أداء عام في الاختبار. بشكل عام من المتوقع وجود اختلافات. هذا يعني عندما تجري مقارنة المجموعتين في التراكيب المقاسة في الاختبار، في حال وجود الاختلافات، عندئذ تكون دراسة تفاوت DIF أداء البند موجودة ويجب دراسة خواص البند بحذر قبل استخدامه في أي اختبار. إن البنود المشار إليها "DIF" قد تكون مسببة للمشكلات بسبب سوء الترجمة أو بسبب استعمال مصطلح، موقف، أو تعابير غير معروفة أو مألوفة إلى إحدى المجموعات الثقافية. هناك أسباب أخرى أيضاً ربما المهارة التي يجري اختبارها في تلك البنود ليست جزءاً من ذخيرة ثقافة مجموعة اللغة المستهدفة أو ربما يكون شكل البند غير مألوف، إن تقرير أسباب الاختلاف مهم لأنه يوثر في تقرير ما يجب القيام به بشأن ذلك البند.

يكون لتقديم الخطوط الرئيسة معنى عندما يكون هناك إثبات أن التركيب له صلة مع المجموعة الحضارية المستهدفة، وأن هناك إثباتاً أنه قد جرى التحقق من الترجمة والتكييف بحذر (ربما بواسطة خطة الترجمة المقدمة). هناك ثلاث منهجيات يمكن استخدامها بشكل أساسي لإجراء عدة نماذج للتحليلات المطلوبة في تلك الخطوط الرئيسة:

(أ) إجراءات (IRT انظر اليس، 1989، 1999، اليس وكيميل، 1992).

(ب) إجراء مانتل - هانزل (MH) وما يتبعه (انظر، هامبلتون، كلوس، يزور، وجونز، 1993، هولند وثابر، 1988، هولند ويونر، 1993، سيرغي وآلالوف، 203).

(ج) إجراء منطقي ارتدادي (LR سوامينثان وروجرز، 1990) إن كل هذه المنهجيات شرطية بمعنى أن المقارنة تجري بين مجموعتي أشخاص مثلاً: (إنجليز - فرنسيين) الذي جرى الافتراض أنهم متماثلون في المهارات المقاسة في الاختبار. في إجراء IRT يُطابق الممتحنون باستخدام درجات المهارة المقدرة (تقدر باستخدام نموذج درجات البنود). كل تلك الإجراءات مجتمعة تستخدم الدرجات النهائية للاختبار لمقارنة الممتحنين. وكل تلك المنهجيات تعطي نتائج ثابتة وصادقة في حال كانت النماذج كثيرة ومتوفرة وأنه قد تم إنجاز الإجراء بشكل صحيح وأن النتائج قد فسرت بعناية.

يحتاج إجراء LR و MH إلى عينة حجمها 200 من كل مجموعة ثقافية على الأقل. بشكل عام تحتاج إجراءات IRT حقيقة إلى عينات أكبر.

10. د 10: يجب عدم استخدام البنود غير المتكافئة للمجموعة الثقافية المستهدفة في ربط الاختبار المكيف مع مقياس الدرجات النهائية العامة المقدمة. على كل حال تلك البنود قد تكون مفيدة لتقديم الدرجات في كل مجموعة بشكل منفصل.

الأسباب/ الشرح. قد تعتبر بنود في الاختبارات المكيفة بعض الأحيان غير متكافئة بسبب التكييف السيئ أو كونه غير مناسب في تلك الثقافة (هلن، 1987). لا يمكن استخدام تلك البنود في ربط النسخة المكيفة للاختبار مع مقياس الدرجات النهائية العامة لأنها توفر معلومات مختلفة عن المجموعات التي جرت مقارنتها. على كل حال تقدم البنود المكيفة بشكل جيد والتي جرى تعريفها بأنها غير متكافئة (غير مناسبة ثقافياً) معلومات مفيدة عن تلك الثقافات والحضارات. إن معرفة مصدر عدم التكافؤ لتلك البنود قد يقدم معرفة أدق عن ثقافة ولغة تلك المجموعة وهذا يؤدي إلى زيادة فهم تلك المجموعة (اليس، 1991).

الإدارة:

1. أ. 1: إن ظروف البيئة التي تؤثر في إدارة الاختبار يجب أن تكون متشابهة قدر الإمكان عبر المجموعات الثقافية التي يستهدفها الاختبار.

الأسباب/ الشرح. يختلف عدد الصعوبات المتوقعة في إدارة الاختبار باختلاف الفروق اللغوية والثقافية بين المجموعات التي يجري اختبارها أو بين ثقافة المجموعة التي جرى اختبارها أولاً وثقافة المجموعة التي ستقوم بالاختبار. هناك حاجة لمعرفة ثقافة ولغة المجموعة المستهدفة لكي تستطيع العمل في هذا الدليل. من المتوقع أن يواجه المطور بصراحة الصعوبات التي تؤثر في عملية المقارنة وأن يدرس الإجراءات الضرورية، يجب تقديم الدلائل التجريبية لدعم مطالبة المقارنة. إذا لم يكن ذلك ممكناً يمكن تقديم مناقشة عقلانية لتبرير استخدام الاختبار المكيف عبر الثقافات.

2. أ. 2: يجب على المطورين والإداريين محاولة توقع أنواع الصعوبات واتخاذ الإجراءات المناسبة لمعالجتها وذلك بإعداد إرشادات ومواد مناسبة.

الأسباب/ الشرح. يجب أن يكون لدى مطوري الاختبار معلومات جيدة عن تطور الاختبار عبر الثقافات، إضافة إلى ذلك يجب أن يتمتعوا بالخبرة الكافية كي يستطيعوا إدارك تعقيدات وخاصية إدارة الاختبار عبر الثقافات. إحدى الطرق العلمية هي إعداد جدول عن الصعوبات التي تحدث غالباً وقد تهدد صدق الاختبار.

إن معرفة اللغة والثقافة الدقيقة للمجموعة المستهدفة ضروري جداً؛ على سبيل المثال فإن درجة 3 أو 4 في مقياس الدرجات في تركيا هو الأفضل، وتشكل درجات أعلى مشكلات في الدلالات اللغوية.

3. أ. 3: يجب أن يكون لدى الإداريين إحساس دقيق بعدد من العوامل المتعلقة بالمواد المحفزة، الإجراءات الإدارية، وأساليب الاستجابات التي قد تؤثر على صدق الاستنتاجات التي تم الحصول عليها من الدرجات النهائية.

الأسباب/ الشرح. قد تكون شروط إدارة الاختبار مصدر متغيرات غير مقصودة في الدرجات النهائية. كي نستطيع مضاعفة صدق وعملية مقارنة درجات الاختبار عبر مجموعات ثانية، يجب وصف الأسباب التي قد تؤدي إلى اختلافات في الدرجات النهائية.

4. أ. 4: يجب وجود التعليمات الإدارية في لغة المصدر وفي اللغة المستهدفة لتقليص تأثير أسباب الاختلافات (غير المرغوب فيها) عبر المجموعات المختلفة.

الأسباب/ الشرح. تخاطب الدراسات عبر الثقافات غالباً مجموعات ذوي خلفية مختلفة جداً. عندما يبدأ الطلاب المشاركون في الاختبار بالإجابة عن الأسئلة/ التمارين يجب تقليص تأثير مصادر الاختلافات غير المرغوب فيها بقدر المستطاع. إحدى الطرق للقيام بهذا هو إرشادات الاختبار الواضحة.

5. أ. 5: يجب أن يحتوي كتيب الاختبار على كل تفاصيل الاختبار وإدارته التي تحتاج إلى تدقيق في تطبيق الاختبار في محيط ثقافي جديد.

الأسباب/ الشرح. كثير من السمات المتعلقة بإدارة الاختبار ضمن مجموعة لغوية أخرى يمكن أن يتوقعها الذين يقومون بتطوير الاختبار؛ لذلك يجب على المطورين جمع معلومات عن موضوعات معينة والتي من المحتمل أن تكون متعلقة بالاختبار المكيف في أثناء عملية تطوير صدق الاختبار. في بعض الحالات يحصل المكيف على معطيات من الأقليات الثقافية أو التطبيقات عبر الثقافات الموجودة، يجب وجود المعلومات المتعلقة بإدارة اختبارات تلك المجموعات في كتيب الاختبار.

6. أ. 6 يجب أن يكون الإداري فضولياً كما يجب تقليل العلاقة بين الإداريين والطلاب. تناقش قواعد واضحة عن إدارة الاختبار لاحقاً في الكتيب.

الأسباب/ الشرح. من الممكن أن يكون تأثير الإداري على نتائج الاختبار حقيقياً. إن الهدف هو تقليص ذلك التأثير وذلك بأن يتعهد الإداريون باتباع الإرشادات والإجراءات المعتمدة في إجراء الاختبار؛ من ناحية ثانية قد يكون للإداريين تأثير غير واضح، وغير مرغوب. إن صفات الإداري مثل الجنس، العمر، العرق وحتى طريقة اللباس وأشياء أخرى يمكن أن تؤثر في نتائج الاختبار خاصة إذا كان هنالك إداري واحد فقط. إذا جرى استخدام اختبار جديد مكيف في مجموعة ثقافية ما يمكن أن يكون من الأسهل نسبياً، إذا كان الإداري ينتمي إلى ذات المجموعة، أن نستطيع تحديد صفات الإداريين التي يمكن أن تهدد صدق حصيلة نتائج الاختبار. عندئذ يمكن القيام ببعض الخطوات (كدراسة تجريبية) في حالة عدم تماثل الخلفية الثقافية للإداريين والطلاب الممتحنين خاصة، يجب التحقق من التأثير السلبي المحتمل للإداري واتخاذ الخطوات لتقليل المشكلات المحتملة.

التوثيق/ تفسير الدرجات:

1. 1. 1 عندما يكيف الاختبار لاستخدامه ضمن مجموعة مختلفة، يجب توثيق التغيرات في الاختبار مع البراهين الداعمة لتكافؤ النسخة المكيفة للاختبار.

الأسباب/ الشرح. من الممكن أن توفر معلومات عن تفاصيل في تكييف اختبار فكرة ثاقبة إذا كان من المناسب استخدام الاختبار ضمن بيئة محددة. على سبيل المثال: معرفة أن عوامل ثقافية، اجتماعية في ثقافة ما قد تم أخذها بعين الاعتبار - أثناء عملية تكييف اختبار لمتكلمي اللغة الإسبانية في أميركا الجنوبية يمكن أن يكون ذا قيمة عند تقرير عما إذا كان الاختبار مناسباً

لاستخدام المتكلمين بالإسبانية في الولايات المتحدة. يجب توثيق الإجراءات المتبعة في تكييف الاختبار بكاملها في كتيب الاختبار لتسهيل تقويم الاختبار من قبل مستخدمين محتملين. يجب أن يتضمن التوثيق بيان خطوات مفصلة عن الإجراءات بكاملها بما فيها المحاكمة العقلية المستخدمة، الطرق المستخدمة في تقويم البنود وتكافؤ الاختبار المكيف ونتائجه. تفاصيل عن اختيار المترجمين واستخدامهم، أسباب وتبرير استخدام وإدخال بعض البنود ومعلومات عن البنود التي تم تعديلها أو استبعادها، بعض المشكلات الرئيسة التي واجهت سير عملية التكييف وكيف أمكن حلها، كل النواحي المتعلقة بإدارة الاختبار بما فيها اختبار وتدريب الإداريين وتفسير النتائج.

2. 1. 2: يجب أن لا يكون لاختلاف درجات اختيار عينات المجموعات التي أجرت الاختبار قيمة ظاهرية. تقع على الباحث مسؤولية إثبات معنى الاختلاف بدلائل تجريبية (إمبريقية).

الأسباب/ الشرح. يبدو أن الأخطاء العامة في التطبيق هي التي تعطي أهمية محدودة العملية تكييف الاختبار، والتي تفسر اختلافات الدرجات لمجموعة كأنها انعكاس اختلافات حقيقية للتركيب الذي يجري قياسه بواسطة الاختبار. إن تجاهل مشكلات تكييف الاختبار التي تحصل بشكل دوري في التطبيق والحاجة إلى تأكيد صدق الاختبارات في الثقافات التي تجريها قد شوهت صحة نتائج الدراسات الكثيرة عبر الثقافات، إن وجود عملية تكييف موثوقة أساسية لإثبات صدق الاختبار المكيف. في الوقت ذاته حتى في وجود اختبارات مكيفة ممتازة يجب على الباحثين القيام بجهود لتفسير نتائجهم. بشرط معرفتهم الكاملة للثقافات المستخدمة للاختبار. هذا يعني، على سبيل المثال أنه يجب جمع دلائل ثابتة كلما كان ذلك

ممكناً، وإذا لم يكن ذلك ممكناً يجب الحذر التام في تفسير النتائج التي تم الحصول عليها من مجموعات مختلفة.

3. 1. 3: يمكن إجراء المقارنة عبر المجموعات في حالة ثبات المقياس الذي تستند عليه الدرجات.

الأسباب/ الشرح. في بعض الأحيان من الممكن تصنيف درجات اختبار بلغات مختلفة حسب مقياس عام وذلك بغية تسهيل عملية المقارنة للدرجات. عند الحصول على نماذج كثيرة، نماذج إحصائية فعالة مثل إحصائيات IRT (انظر هامبلتون وآخرين، 1991).

يمكن الحصول على معدل مترابط لدرجات اختبار مكيف إذا كان بناء الاختبار متعادلاً في النسخ المختلفة وإذا كانت المعطيات الصحيحة لذلك التعادل متوفرة (انظر د. 6). عند حصول ذلك يمكن القيام بكل أشكال مقارنة الدرجات بما فيها متوسط الدرجات، الانحراف القياسي والتوزيع. غالباً تكون درجات نسخ الاختبار المترجمة إلى لغات مختلفة لم تعدل جيداً وبذلك لا يمكن مقارنة الدرجات مباشرة. ومع ذلك يمكن القيام بمقارنة دور بناء الاختبار في كل مجموعة لغوية. على سبيل المثال في اختبار المهارات المكيف من الإنجليزية إلى الإسبانية، يمكن أن يكون الباحث مهتماً بمقارنة صدق مهارة التنبؤ في الاختبار في كل مجموعة لغوية. إن الغرض الأساسي لهذا الكتيب هو التأكد أن الباحثين لا يقومون بمقارنات غير مبررة لدرجات اختبار نسخ مترجمة للغات متعددة وأنهم يقتصرون في عمليات التفسير على المقارنات التي توفر دلائل صدقها.

4. 1. 4: يجب على مطور اختبار توفير معلومات محددة عن الطرق التي يمكن أن تؤثر فيها الثقافة الاجتماعية، البيئة لمجموعة على الأداء في الاختبار، كما يجب عليه اقتراح إجراءات لبيان أسباب تلك التأثيرات على عملية تفسير النتائج.

الأسباب/ الشرح. إن العوامل المختلفة في أي دراسة عبر ثقافات/ القوميات المتعلقة بأسباب الاختبار يجب اعتبارها عاملاً للحصول على فهمٍ كاملٍ للنتائج (براكون وبراونا، 1991). إن العوامل الاجتماعية/ السياسية المختلفة التي تؤثر على الأداء في الاختبار بشكل ثابت لا تؤخذ بعين الاعتبار (فان دي فيفر وبورتينغا، 1991) غالباً. مثلاً عند مقارنة الأداء الأكاديمي لطلاب من دول نامية وطلاب من دول متطورة، يمكن أن تكون فروق الأداء عائدة إلى عدم إمكان الحصول على مراجع لا عن عدم وجود الإمكانيات أو ربما تكون انعكاساً لنوعية الخدمات التربوية المتوفرة.

الخاتمة

كي يجري تقدير معنى وفائدة أبحاث عبر الثقافات، من الضروري أن يكون الباحثون حذرين في اختيار الإداريين، وأن يستخدموا بنود اختبار مناسبة وأن يسيطروا على عامل السرعة. بالإضافة إلى ذلك فإن المترجمين المتآلفين مع المجموعة المستهدفة ومع ثقافتهم، الذين يعرفون محتوى الاختبار والذين تم تدريبهم على تطوير الاختبار، هم الأشخاص الأكثر مقدرة على تقديم اختبار مكيف صادق. إن اختبار مخطط عقلاني مناسب، مخطط جميع معطيات وتحليل إحصائيات يمكن أن يوفر معطيات قيمة متعلقة ببنود الاختبار وتكافؤ الاختبار عبر مجموعات لغوية وثقافية مختلفة. أما ما يتعلق بتفسير الدرجات، فيجب أن نفكر بحذر بالتفاصيل الخلفية للمتغيرات المؤثرة على الأداء، المناهج المختلفة، مستوى الدوافع والعوامل الاجتماعية/ السياسية التي يمكن أن تكون مهمة بشكل خاص. يجب أن لا تقوم المقارنة بالتركيز على الاختلافات فقط. يمكن أن تؤمن التشابهات بين المجموعات المختلفة اللغات والثقافات معلومات مفيدة ووثيقة الصلة بالبحث.

إن دليل ITC لتكييف الاختبار الذي جرى وصفه في هذا الفصل يؤمن خطوط عمل عريضة للباحثين للقيام بدراسات في تكييف الاختبار ويتوقع أن يكون ذلك

الدليل مع التوصيات المرافقة مفيداً لعدد كبير من الهيئات وأن يحسن نوعية تكييف الاختبارات في العالم وبذلك يساهم في صدق أبحاث عبر اللغات وعبر الثقافات (انظر ITC، 2001). هنالك عدد لا بأس به من المراجع للقراء: قدم كيسنجر (1994). هامبلتون وباتسولا (1999) خطوات مفصلة لمشاريع تكييف الاختبارات، هامبلتون وغيره (1999) تقدموا بنتائج واحد من الاختبارات الأولية الميدانية، تناول هاركنس (1998) موضوعات وطرق مرتبطة بتكييف الاختبار مع التركيز على مقياس التقدير، فان دي فيفر وبورتينغا (1997) قدموا خطة عمل لاستقصاء التهديدات لصدق تفسير درجات اختبار عبر الثقافات.

كما قدم فان دي فيفر وتانزر (1997) سيرغي وآلالوف (2003) لوائح شاملة للإجراءات الاحصائية.

شكر

يظهر هذا الفصل أيضاً في "تقرير البحث التقويمي والقياس السيكولوجي المخبري" رقم 353، جامعة ماستشوست، كلية التربية، امهرست.

يود المؤلف أن يشكر مجلس الكلية لتقديمه الدعم المادي للبحث، على كل حال مجلس الكلية ليس مسؤولاً عن أي خطأ كما أنه لا يجب افتراض مصادقة المجلس على الآراء المقدمة.

يشكر المؤلف فون فان دي فيفر ويب بورتينغا من جامعة تيلبرغ لمساعدتهم في إعداد هذا الفصل كما يشكر يان باتسولا واكتبيل كانجي لمساعدتهم التقنية.

*****

المراجع

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

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. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural psychology* (pp. 137–164). Newbury Park, CA: Sage.

Butcher, J. N., & Garcia, R. E. (1978). Cross-national application of psychological tests. *The Personnel and Guidance Journal, 56*(8), 472–475.

Cziko, G. (1987). Review of the Bilingual Syntax Measure I. In J. C. Alderson & K. J. Krahnke (Eds.), *Reviews of English language proficiency tests*. Washington, DC: Teachers of English to Speakers of Other Languages.

Ellis, B. B. (1989). Differential item functioning: Implications for test translation. *Journal of Applied Psychology, 74,* 912–921.

Ellis, B. B. (1991). Item response theory: A tool for assessing the equivalence of translated tests. *Bulletin of the International Test Commission, 18,* 33–51.

Ellis, B. B., & Kimmel, H. D. (1992). Identification of unique cultural response patterns by means of item response theory. *Journal of Applied Psychology, 77,* 177–184.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing, 2*(3), 199–215.

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

Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing, 20*(2), 225–240.

Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. *European Journal of Psychological Assessment, 9,* 54–65.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10,* 229–240.

Hambleton, R. K. (2002). Adapting achievement tests into multiple languages for international assessments. In A. C. Porter & A. Gamoran (Eds.), *Method-*

ological advances in cross-national surveys of educational achievement (pp. 58–79). Washington, DC: National Academy Press.

Hambleton, R. K., & Bollwark, J. (1991). Adapting tests for use in different cultures: Technical issues and methods. *Bulletin of the International Test Commission, 18,* 3–32.

Hambleton, R. K., Clauser, B. E., Mazor, K. M., & Jones, R. W. (1993). Advances in the detection of differentially functioning test items. *European Journal of Psychological Assessment, 9*(1), 1–18.

Hambleton, R. K., & de Jong, J. (Eds.). (2003). Advances in translating and adapting educational and psychological tests. *Language Testing, 20*(2), 127–240.

Hambleton, R. K., & Kanjee, A. (1995a). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. *European Journal of Psychological Assessment, 11,* 147–160.

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

Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Applied Testing Technology, 1*(1), 1–16.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hambleton, R. K., Yu, J., & Slater, S. C. (1999). Field-test of the ITC guidelines for adapting educational and psychological tests. *European Journal of Psychological Assessment, 15*(3), 270–276.

Harkness, J. (Ed.). (1998). *Cross-cultural equivalence.* Mannheim, Germany: ZUMA.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel–Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P. W., & Wainer, H. (Eds). (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Hulin, C. L. (1987). A psychometric theory of evaluations of item and scale translations: Fidelity across languages. *Journal of Cross-Cultural Psychology, 18,* 115–142.

Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Application of item response theory to analysis of attitude scale translation. *Journal of Applied Psychology, 67,* 818–825.

International Test Commission. (2001). *International Test Commission guidelines for test adaptation.* London: Author.

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

Lonner, W. J. (1990). An overview of cross-cultural testing and assessment. In R. W. Brislin (Ed.), *Applied cross-cultural psychology* (Vol. 14, pp. 56–76). Newbury Park, CA: Sage.

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

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

Rosansky, E. J. (1979). A review of the Bilingual Syntax Measure. In B. Spolsky (Ed.), *Some major tests: Advances in language testing (Series 1).* Arlington, VA: Center for Applied Linguistics.

Sireci, S. G. (1997). Problems and issues in linking tests 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), 148–166.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361–370.

van Leest, P. F., & Bleichrodt, N. (1990). Testing of college graduates from ethnic minority groups. In N. Bleichrodt & P. J. D. Drenth (Eds.), *Contemporary issues in cross-cultural psychology.* Amsterdam: Swets & Zeitlinger.

van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: 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., & Poortinga, Y. H. (1991). Testing across cultures. In R. K. Hambleton & J. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 277–308). Boston: Kluwer Academic.

van de Vijver, F. J. R., & Poortinga, Y. H. (1992). Testing in culturally heterogeneous populations: When are cultural loadings undesirable? *European Journal of Psychological Assessment, 8,* 17–24.

van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13,* 29–37.

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.

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., & Munoz-Sandoval, A. F. (1993). An IRT approach to cross-language test equating and interpretation. *European Journal of Psychological Assessment, 9,* 233–241.