لجنة الاختبارات الدّوليّةITC

إرشادات لجنة الاختبارات الدّوليّة لترجمة ونقل الاختبار لثقافة جديدة (الإصدار الثاني)

نسخة 2.4

قام بالترجمة فريق من طلاب الماجستير بقسم علم النفس بجامعة الكويت تحت أشراف أ.د. هشام فتحي جادالرب:
ميسون الناصر، فجر العيسى، نور الجاسر، وسمية العجمي.

Translated by Prof. Hesham F. Gadelrab
Maysoon Alnaser, Fajer Aleisa, Nour Aljaser, Wasmiah Alajmi
Psychology Department, Kuwait University

عند الاقتباس من النسخة المترجمة إلي اللغة العربية:

اللجنة الدولية للاختبارات. (2017). *مبادئ ITC الإرشادية لترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني)*. ترجمة هشام فتحي جادالرب، ميسون النصر، فجر العيسى، نور الجاسر، وسمية العجمي. [www.intestcom.org]

شكر وتقدير:

يتقدّم مجلس إدارة لجنة الاختبارات الدّوليّة بخالص الشّكر من اللجنة المكوّنة من ستة أشخاص، والتي عملت لعدّة سنوات لإنتاج الإصدار الثّاني من المبادئ الإرشادات للتّرجمة ونقل الاختبار لثقافة جديدة: ديفيد بارترام David Bartram (المملكة المتحدة)، جيراي بربروجلو Giray Berberoglu (تركيا)، جاك جريجوار Jacques Grégoire (بلجيكا)، رونالد هامبلتون Ronald Hambleton؛ رئيس اللجنة (الولايات المتحدة الأمريكية)، خوسيه مونيز Jose Muniz (إسبانيا)، وفونس فان دي فيجفر Fons van de Vijver (هولندا).

كما تود لجنة الاختبارات الدّوليّة أن تشكر تشاد بوكيندال Chad Buckendahl (الولايات المتّحدة الأمريكيّة)، آن هيرمان Anne Herrmann؛ وزملائها في شركة OPP للخدمات المحدودة (المملكة المتّحدة)، وأبريل زنيسكاي April Zenisky (الولايات المتحدة الأمريكيّة) للمراجعة الدّقيقة للمسودّة الأولية لهذه الوثيقة. وتعبّر لجنة الاختبارات الدّوليّة عن امتنانها لكل من ساهم في مراجعة الإصدار الثّاني من وثيقة ITC لترجمة ونقل الاختبارات لثقافة جديدة بشكل مباشر وغير مباشر.

ملخّص:

تم إعداد الإصدار الثّاني من إرشادات لجنة الاختبارات الدّوليّة لترجمة ونقل الاختبار لثقافة جديدة ما بين عامي 2005، و2015 لتطوير وتحسين الإصدار الأول، ولمواكبة التّقدّم التكنولوجي للاختبارات وممارساتها. تم تنظيم الإرشادات (عددها 18) ضمن (6) فئات لتسهيل استخدامها: شروط مُسبقة Pre-condition (3)، تطوير الاختبار Test Development(5)، التّحقّق Confirmation (4)، التطبيق والإدارة Administration (2)، رصد الدّرجات وتفسيرها Score Scales and Interpretation (2)، التقرير (التوثيق) Documentation (2). لكل واحد من هذه الفئات شرحً مفصّلاً واقتراحات للتطبيق والممارسة الفعلية. تم إضافة قائمة مراجعة في نهاية هذه الوثيقة، لتحسين استخدام الإرشادات المنصوص عليها.

قائمة المحتويات:

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

خلفية نظرية

شهد مجال ترجمة الاختبارات ومنهجيّة نقلها إلى ثقافة جديدة تقدمًا سريعًا في آخر 25 سنة أو نحو ذلك، حيث تم نشر العديد من الكُتب والدّراسات والأبحاث وأمثلة على نقل الاختبارات إلى ثقافة جديدة، أنظر على سبيل المثال، (van de Vijver & Leung, 1997, 2000: Hambleton, Merenda, & Spielberger, 2005; Grégoire & Hambleton, 2009; Rios & Sireci, 2014). كانت هذه التّطوّرات ضروريّة بسبب الاهتمام المتزايد بـ: (1) علم النّفس عبر الثّقافات، (2) الدّراسات الدّوليّة المقارنة واسعة النّطاق للتحصيل الدراسي (على سبيل المثال، TIMSS، PISA، OECD)، (3) امتحانات الاعتماد المستخدمة في جميع أنحاء العالم (على سبيل المثال: في مجال تكنولوجيا المعلومات من قِبل شركات مثل Microsoft، Cisco)، و(4) عدالة الاختبارات؛ من خلال السّماح للمستجيبين باختيار اللغة التّي يستجيبوا بها (على سبيل المثال: في اختبارات القبول الجامعي في بعض الدول التي لديها مستجيبين يجيدون لغات مختلفة، يتم السماح لهم باختيار لغة من بين عدة لغات لتطبيق الكثير من الاختبارات).

تم إحراز تقدّم تقني في مجالات الأساليب الكيفيّة (النّوعيّة)، والكميّة لتقييم التّحيّز في التكوين الفرضي موضع القياس، وفي المنهجية، وفي البنود المكونة للاختبارات والاستبانات التي تم نقلها من ثقافة لأخرى، بما في ذلك استخدامات الإجراءات الإحصائيّة المعقّدة مثل نظريّة الاستجابة للبند IRT، ونمذجة المعادلة البنائية SEM ونظرية إمكانية التّعميم (انظر Hambleton et al., 2005; Byrne, 2008). تم تطوير تصميمات جديدة للترجمة من قِبل منظّمة OECD/PISA (انظر Grisay, 2003). تم تقديم خطوات لاستكمال مشاريع نقل الاختبارات لثقافات جديدة (انظر على سبيل المثال: Hambleton & Patsula, 1999؛ والمشاريع النّموذجيّة

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

المتاحة لتوجيه ممارسات نقل الاختبارات لثقافات جديدة مثل OCED/PISA، وTIMSS)، وغيرها من التطورات.

الطبعة الأولى من هذه الإرشادات (أنظر van de Vijver & Hambleton, 1996; Hambleton, 2005) قد بدأت من منظور مقارن، حيث كان الهدف من نقل الاختبارات لثقافة جديدة هو السماح بـ (أو تسهيل) عملية المقارنة بين مجموعات مختلفة من المستجيبين. تم استخدم النموذج الضمني الذي تم تصميم الإرشادات من أجله لتطوير أداة القياس بشكل متكرر في سياق مقارن (يجب تكييف أداة القياس الأصلية لاستخدامها في سياق ثقافي جديد). ومع ذلك، أصبح من الواضح بشكل متزايد أن تكييف الاختبارات لها مجالات أوسع من التطبيقات. قد يكون الأمثلة الأكثر أهمية لهذه التطبيقات هي: استخدام أداة جديدة أو موجودة بالفعل مع مجموعات متعددة الثقافات، مثل العملاء الذين يخضعون للإرشاد النفسي Counseling، والذين لديهم مجموعات عرقية مختلفة، والتقييم التعليمي في مجموعات متنوعة اثنياً؛ لأفراد لديهم مستويات مختلفة في إتقان اللغة التي أُعد بها الاختبار، والإجراءات الاختبارية المتعلقة بالتوظيف في الشركات متعددة الجنسيات. هذا التغيير في مجال استخدامات الاختبارات كان له آثار على تطوير وتطبيق وصدق وتقرير إجراءات تطوير الاختبار. على سبيل المثال، قد تشمل إجراءات نقل الاختبار للغة وثقافة جديدة؛ تعديل بنود الاختبار من أجل زيادة قابلية فهمه لغير الناطقين بلغة البند الاصلية (مثلاً، من خلال تبسيط اللغة). هناك امتداد أخر مهم للإرشادات يجب أن تهتم به، وهو استيعاب التطوير المتزامن (أي التطوير المشترك للغة أداة القياس الاصلية، واللغة التي تم نقل الأداة لها). تضع المشاريع الدولية واسعة النطاق هذا التطوير المتزامن في اعتبارها بشكل متزايد. بطبيعة هذه المشاريع

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

يتم تطبيق أدوات القياس على أفراد من دول وثقافات مختلفة، لهذا فهي تسعلا لتجنب مشكلة أن لغة الاختبار الذي يتم تطويره، يصعب نقلها إلى لغات أخرى تقع ضمن نطاق المشروع.

تم نشر الطّبعة الأولى من إرشادات لجنة الاختبارات الدّوليّة لترجمة الاختبارات ونقلها إلى ثقافة جديدة بواسطة "فان دي فيجفر" و"هامبلتون" عام 1996، و"هامبلتون" عام 2002، و"هامبلتون" و"ميريندا" و"سبيلبيرق" عام 2002 (de Vijver & Hambleton, 1996; Hambleton, 2002; Hambleton, Merenda and Spielberger, 2005). هناك فروق طفيفة في الإرشادات التي تم نشرها بين عامي 1996، و2005، في حين حدثت العديد من التّطورات منذ عام 1996 والذي ظهرت فيه الإرشادات للمرة الأولى: أولًا: كان هناك عدد من المراجعات الجيّدة لإرشادات لجنة الاختبارات الدّوليّة. وتشمل هذه الأوراق التي كتبها "جانري وبرتراند" (Jeanrie & Bertrand, 1999)، و"تانزر وسيم" (Tanzer & Sim, 1999)، و"هامبلتون" (Hambleton, 2002). سلط جميع المؤلفين الضوء على قيمة الإرشادات، ولكنهم قدموا بعد ذلك سلسلة من الاقتراحات لتحسينها. قام "هامبلتون" و"ميريندا" و"سبيلبرجر" (Hambleton, Merenda, & Spielberger, 2005) بنشر الجلسات الرئيسية للمؤتمر الدولي للجنة الاختبارات الدّوليّة الذّي عقد عام 1999 في جامعة جورج تاون بالولايات المتحدة الأمريكية. قدم العديد من مؤلفي الفصول نماذج جديدة لنقل الاختبار لثقافة جديدة وقدموا منهجيات جديدة منها الفصل الذي كتبه (Cook & Schmitt-Cascallar 2005) و (Sireci 2005). في عام 2006 عقدت لجنة الاختبارات الدّوليّة مؤتمرًا دوليًا في بروكسل/بلجيكا، كان مركزاً على الإرشادات التي أعدها لترجمة الاختبارات ونقلها إلى ثقافة جديدة. كان محور اهتمام أكثر من 400 شخص من أكثر من 40 دولة، هو موضوع نقل الاختبارات إلى ثقافة جديدة، كما تم تقديم العديد من

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

الأفكار المنهجية الجديدة، وتم اقتراح عدد من الإرشادات الجديدة، كما تم مشاركة أمثلة للتطبيق الناجح لهذه الإرشادات. كانت الأوراق المقدمة في الندوات في الاجتماعات الدّوليّة من عام 1996 إلى عام 2009 وفيرة (انظر على سبيل المثال، Grégoire & Hambleton 2009)، وانظر (Hambleton, Elosua & Muniz, 2013) للحصول على نسخة أولية من الإصدار الثاني من إرشادات لجنة الاختبارات الدّوليّة باللغة الإسبانية.

في عام 2007، شكلت لجنة الاختبارات الدّوليّة لجنة من ستة أشخاص وتم تكليفهم بمهمة تحديث إرشادات اللجنة للتأكيد على المعرفة الجديدة التي تم تطويرها والخبرات العديدة التي اكتسبها الباحثون في هذا المجال. تشمل هذه التطورات (1) تطوير نمذجة المعادلة البنائية لتحديد التكافؤ العاملي للاختبار عبر المجموعات اللغوية، (2) مناهج موسعة لتحديد الأداء التفاضلي/التمييزي للبند DIF، مع مقاييس تصنيف الاستجابة المتعددة عبر المجموعات اللغوية، و(3) تصميمات نقل الاختيارات الجديدة والتّي ابتكرتها مشاريع التقييم الدّوليّة مثل OECD / PISA و TIMSS. قدمت اللجنة أيضًا عروضًا ومسودات للإرشادات الجديدة في الاجتماعات الدولية لعلماء النفس في براغ (في عام 2008) وأوسلو (في عام 2009) وتلقت تعليقات جوهرية بشأنها.

تم الاحتفاظ بقسم إرشادات تطبيق الاختبارات في الإصدار الثاني، ولكن تم دمج الإرشادات المتداخلة وتم تقليل العدد الإجمالي من ستة إلى اثنين. "تقرير وتوثيق الإجراءات / تفسيرات الدرجات" كان القسم الأخير في الطبعة الأولى. في الإصدار الثاني، قمنا بتقسيم هذا إلى قسمين منفصلين – أحدهما يركز على عمليات تقييس الدرجات والتفسيرات، والآخر يركز على التقرير والتوثيق. بالإضافة إلى ذلك، تم تنقيح اثنين من المبادئ التوجيهية الأصلية الأربعة في هذا القسم بشكل كبير.

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

كما في الإصدار الأول، نريد أن يكون من الواضح للقراء تمييزنا بين ترجمة الاختبار test translation ونقله إلى ثقافة جديدة test adaptation. ربما تكون ترجمة الاختبار هي المصطلح الأكثر شيوعًا، لكن نقل الاختبار لثقافة جديدة هو المصطلح الأوسع ويشير إلى ترجمة الاختبار من لغة ونقله من ثقافة إلى أخرى. يشير نقل الاختبار إلى ثقافة جديدة إلى جميع الأنشطة بما في ذلك: تقرير ما إذا كان اختبار بلغة وثقافة ثانية يمكنه قياس نفس البنية في اللغة الأولى أم لا، واختيار المترجمين، واختيار تصميم لتقييم عمل مترجمي الاختبار (على سبيل المثال، الترجمات الأمامية forward translations والعكسيّة backward translations)، واختيار أي وسائل التوفيق اللازمة بين النسخة الأصلية والجديدة للاختبار، وتعديل صيغة format الاختبار، وإجراءات الترجمة، والتحقق من التكافؤ بين الاختبار باللغة الاصلية والاختبار باللغة والثقافة الثانية وإجراء دراسات الصدق اللازمة. من ناحية أخرى، فإن مصطلح ترجمة الاختبار له معنى محدود مقارنة بمصطلح نقل الاختبار لثقافة جديدة، حيث يقتصر على الاختيار الفعلي للغة والثقافة التي سينقل الاختبار لها؛ للحفاظ على المعنى اللغوي. ترجمة الاختبار ليست سوى جزء من عملية نقله إلى ثقافة جديدة، ولكن يمكن أن تكون، بمفردها نهجًا مبسطًا للغاية لنقل اختبار من لغة إلى أخرى دون أي اعتبار للتكافؤ التربوي أو النفسي.

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

المبادئ الإرشادية

مقدمة

يتم تعريف الإرشادات في نطاق عملنا على أنّها ممارسة مهمّة لإجراء وتقييم عملية نقل الاختبار إلى ثقافة جديدة (ويطلق عليه أحيانًا "التكييف المحلّي" localisation، أو التطوير المتزامن simultaneous development للاختبارات النّفسيّة والتّربوية للاستخدام على مجتمعات مختلفة. في النص التالي، تمّ تنظيم (18) مبدأ إرشادي حول ست مواضيع عامّة: الشروط المُسبقة (3)، تطوير الاختبار (5)، التحقق [التّحليلات الإمبريقية] (4)، التطبيق والإدارة (2)، رصد الدّرجات وتفسيرها (2)، والتقرير (التوثيق) (2).

يسلّط الموضوع الأول (الشروط المُسبقة) على حقيقة أن يجب اتخاذ بعض القرارات قبل أن تبدأ عمليّة التّرجمة/نقل الاختبار إلى ثقافة جديدة، وينصب الموضوع الثاني (تطوير الاختبار) على العمليّة الفعليّة لنقل الاختبار إلى ثقافة جديدة، ويتضمّن الموضوع الثالث (التحقق) على الإرشادات المتعلّقة بتجميع الأدلّة الإمبريقية (التجريبية) لمعالجة تكافؤ وصدق وثبات الاختبار بلغات وثقافات متعدّدة. تتعلّق المواضيع الثّلاث الأخيرة بالتطبيق والإدارة، ورصد الدّرجات وتفسيرها، و"التقرير أو التوثيق". "التقرير (التوثيق)" بشكلٍ خاص كان موضوعًا مُهمَلًا في المبادرات السابقة لنقل الاختبار إلى ثقافة جديدة في علم النّفس والتّربية، ونود أن نرى محررّي المجلّات وجهات التّمويل يطلبون المزيد من التفاصيل من الباحثين عند تقرير وتوثيق إجراءات نقل الاختبارات إلى ثقافة جديدة. لكل مبدأ إرشادي، قدّمنا شرحًا واقتراحاتٍ لتطبيقه وممارسته عمليًا.

المبادئ الارشادية للشّروط المُسبقة:

(1) PC-1 الحصول على الإذْن اللازم من صاحب الحقوق الملكيّة الفكريّة المتعلّقة بالاختبار قبل إجراء أي نقل للاختبار للثقافة الجديدة.

الشّرح: تشير الحقوق الملكيّة الفكريّة إلى مجموعة من الحقوق التّي يتمتّع بها الأشخاص على إبداعهم واختراعاتهم أو منتجاتهم. وهي تحمي مصلحة المبدعين من خلال منحهم حقوقًا معنويّة ومادّية على إبداعهم. ووفقًا للمنظّمة العالميّة للملكيّة الفكريّة (www.wipo.int)، "تتعلّق الملكيّة الفكريّة بأي وحدة من وحدات المعلومات أو المعرفة، والتي يمكن دمجها مع أشياء ماديّة ملموسة في نفس الوقت في عدد غير محدود من النّسخ في مواقع مختلفة في أي مكان في أنحاء العالم".

هناك فرعان للملكيّة الفكريّة: الملكيّة الصّناعيّة industrial propery، وحقوق التأليف والنّشر copyright، الأول: يشير إلى براءات الاختراع، وهو الذّي يحمي الاختراعات والرّسوم والنماذج الصّناعيّة والعلامات والأسماء التّجاريّة. بينما يشير حق التأليف والنّشر إلى الابداعات الفنيّة والقائمة على التّكنولوجيا. للمبدع (المؤلّف) حقوق محدّدة على إبداعه (على سبيل المثال، منع بعض التشويهات أو التّعديلات عند نسخ المحتوى أو نقله من ثقافة إلى ثقافة أخرى)، يمكن ممارسة الحقوق الأخرى (على سبيل المثال: عمل نسخ) من قِبل أشخاص آخرين (على سبيل المثال: النّاشر) حصلوا على ترخيص من المؤلّف أو صاحب حقوق النّشر. بالنّسبة للعديد من الاختبارات كما هو الحال في الأعمال المكتوبة الأخرى، يتم تحديد حقوق النّشر من قِبل المؤلّف للنّاشر أو الموزّع.

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

بما أنّ الاختبارات التّربويّة والنّفسيّة ما هي إلّا إبداع من ابداعات العقل البشري، فهي مشمولة بحقوق الفكريّة. في معظم الوقت لا تشير حقوق النشر إلى المحتوى المحدّد للعناصر (على سبيل المثال: لا يمتلك أي شخص حقوقًا في عناصر مثل "1+1=..، أو "أنا أشعر بالحزن")، ولكن للتّنظيم الأصلي للاختبار (هيكل المقاييس، ونظام التّسجيل، وتنظيم المواد، وما إلى ذلك). وبالتّالي فإنّ محاكاة هيكلية اختبار ما، بمعنى الحفاظ على بناء الاختبار الأصلي ونظام التّسجيل الخاص به مع إنشاء بنود جديدة.، يعد انتهاكًا لحقوق الملكيّة الفكريّة الأصليّة. عندما يحصل مطور اختبار ما على تصريح بإجراء تعديل على الاختبار من قِبل المؤلّف نفسه، فيجب على المطوّر الحفاظ على الخصائص الأصليّة للاختبار (الهيكل أو البناء الاساسي، والمادّة، والشّكل أو الصيغة، والتسجيل...)، ما لم يسمح الاتفاق بإجراء تعديلات على هذه الخصائص من صاحب الملكيّة الفكريّة.

اقتراحات للتطبيق: يجب على مطوري الاختبار احترام أي قانون واتفاقيّات لحقوق النّشر الموجودة للاختبار الأصلي. يجب أن يكون لدى المطّورين اتّفاقيّة موقّعة من صاحب الحقوق الملكيّة الفكريّة (سواءً مؤلّف أو ناشر) قبل البدء في نقل الاختبار للغة والثقافة الجديدة. ويجب أن تحدّد الاتفاقيّة التّعديلات المسموح إجراؤها في الاختبار، والتّي ستكون مقبولة فيما يتعلّق بخصائص الاختبار الأصلي، ويجب أن توضّح من سيمتلك حقوق الملكيّة الفكريّة في النّسخة المعدّلة.

(2) PC-2 القيام بالتأكد من أن مقدار التّداخل في التّعريف ومحتوى التكوين الفرضي construct المقاس بواسطة الاختبار، ومحتوى البنود في المجتمعات التي سينقل منها وإليها الاختبار، كافية وفقاً للغرض المقصود من استخدام درجات الاختبار.

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

الشّرح: يتطلّب هذا المبدأ الإرشادي أن يتم فهم ما يتم قياسه بواسطة الاختبار؛ بنفس الطّريقة عبر المجموعات اللغوية والثّقافيّة، وهذا هو الأساس لمقارنات صحيحة عبر الثّقافات. في هذه المرحلة لم يتم حتى نقل الاختبار أو أداة القياس إلى ثقافة جديدة بعد. لذا من المفضل، قبل القيام باي جهد تجميع الأدلّة الإمبيريقيّة (التّجريبيّة) السّابقة مع اختبارات مماثلة، والتي تتعلق بمدى مطابقة بنود التكوين الفرضي موضع الاهتمام عبر المجموعات اللغوية التي تهتم بها الدراسة. ومع ذلك في النّهاية، يجب أم يظل تقييم هذا المبدأ الإرشادي المهم باستخدام البيانات التّجريبية الميدانية عملية جمع الأدلّة المطلوبة في (10) C–2. لا يتمثّل الهدف من أي تحليلات في تحديد بنية الاختبار، على الرّغم من أن هذا هو من النتائج التي ستحصل عليها بأي حال من هذه التحليلات، ولكن الهدف الرئيس هننا من هذه التحليلات، هو التأكد من تّكافؤ بنيّة الاختبار عبر إصدارات اللغة الاصلية واللغة المنقول إليها.

اقتراحات للتطبيق: يجب اختيار عدد من الخبراء في التكوين الفرضي الذي يقيسه الاختبار، بحيث يكون لديهم دراية كافية بالمجموعتين الثّقافيّة الاصلية والمنقول إليها الاختبار، لتقييم مطابقة البناء المقاس في كل مجموعة من المجموعات الثّقافيّة/اللغويّة. عند قيامهم بهذا، فإنهم يحاولون الإجابة على السؤال التّالي: هل التكوين الفرضي يبدو منطقي make sense في ثقافات كلا المجموعتين؟ على سبيل المثال، لقد شهدنا عدّة مرّات في الاختبارات التربوية، أنّ لجنة الخبراء قد لاحظت أن التكوين الفرضي الذي يتم قياسه بواسطة الاختبار قد أفتقر إلى المعنى عند نقله لثقافة جديدة، أو أنّ المعنى الأصلي اصبح مبهم وغير واضح في الثقافة الجديدة (مثل التكوينات الفرضية: جودة الحياة أو الاكتئاب أو الذّكاء). يمكن استخدام طرق مثل مجموعات العمل، والمقابلات والاستطلاعات، للحصول على معلومات منظّمة وأدلة حول درجة تداخل البناء.

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

(3) PC-3 التقليل من تأثير أي اختلافات ثقافية ولغويّة لا علاقة لها بالاستخدامات المقصودة للاختبار في المجتمع المنقول له.

الشّرح: يجب تحديد الخصائص الثّقافيّة واللغويّة غير المتصلة بالمتغيّرات التّي يهدف الاختبار إلى قياسها في مرحلة مبكرة من المشروع. يمكن أن تكون هذه الخصائص غير المتصلة تتعلق بتنسيق البند، أو طبيعة الاختبار (مثل استخدام الكمبيوتر أو الصّور أو الأشكال...)، أو الحدود الزّمنيّة للاختبار وغيرها من الخصائص.

أحد المقاربات للتعامل مع هذه المشكلة هو تقييم "المسافة اللغوية والثقافية linguistic and cultural distance" بين لغة الاختبار الأصلي، واللغة والمجموعات الثّقافيّة المستهدفة. قد يشمل تقييم المسافة اللغوية والثّقافيّة اعتبارات الاختلافات في اللغة، وبُنية الأسرة، والدّين، ونمط الحياة، والقيم (Van de Vijver & Leung, 1997).

يعتمد هذا المبدأ الإرشادي بشكل أساسي على الأساليب النّوعيّة/الكيفيّة والمتخصّصين المطّلعين على البحث حول الاختلافات الثّقافيّة واللغويّة المحدّدة. إنّه يضع اهتماماً خاصًا على اختيار المترجمين ويتطلّب أن يكون المترجمون متحدثين للغة المستهدفة كلغة أم ومن أبناء الثّقافة المستهدفة؛ لأنّ معرفة اللغة المستهدفة فقط لا يكفي لتحديد المصادر المحتملة لتحيّز في الأسلوب. على سبيل المثال، في الدراسة المقارنة الصّينيّة-الأمريكيّة للإجابة عن أسئلة الرياضيات للصّف الثّامن التي أجراها كلا من هامبلتون ويو وسلاتر (Hambleton, Yu, & Slater, 1999)، تم تحديد مشكلات التّنسيق وطول مدة الاختبار، جنبًا إلى جنب مع مجموعة من السّمات الثّقافيّة المرتبطة باختبار الرّياضيّات للصف الثّامن.

اقتراحات للتطبيق: يصعب التحقق من هذا المبدأ الإرشادي من خلال البيانات التّجريبيّة (الإمبريقية) بصفة عامة، وبشكل خاص يكون من الأصعب التحقق منه بواسطة البيانات التّجريبيّة (الإمبريقية) في المراحل الأولى من عمليّة نقل الاختبار إلى ثقافة جديدة. في الوقت نفسه، غالبًا ما يمكن جمع الأدلّة النّوعيّة/الكيفية عن طريق:

- يمكن تحديد المستويات التحفيزيّة والدافعية للمشاركين، إمّا عن طريق الملاحظة أو المقابلة أو المجموعة المركّزة أو عن طريق المسوح survey؛ كما يمكن تقييم فهمهم للتّعليمات وخبراتهم في أداء الاختبارات النّفسيّة والسّرعة المرتبطة بتطبيق الاختبار والإلمام بطبيعة الاختبار والاختلافات الثّقافيّة (ولكن حتى في هذه النّقطة قد تكون المقارنات غير مفيدة بما يكفي، بسبب الاختلافات الثّقافيّة في فهم المتغيّرات موضع القياس نفسها). عندما يكون هناك صعوبة في جمع مثل هذه البيانات البحثيّة من المشاركين أنفسهم، يمكن الحصول على أكبر قدر ممكن من المعلومات من المترجمين؛ يمكن القيام ببعض من هذا العمل قبل أن تُقدم على نقل الاختبار إلى ثقافة جديدة.
- قد يكون من الممكن ضبط هذه "المتغيّرات المزعجة" غير المرتبطة بالهدف من الاختبار، في أي تحليل ميداني لاحق بعد أن يتم نقل الاختبار إلى ثقافة جديدة. عندما يكون الاختبار جاهزًا لدراسة التّحقّق من صدقه، يمكن استخدام تحليل التّباين المشترك (تحليل التغاير ANCOVA) أو التّحليلات الأخرى التّي تضبط المتغيرات لدى المشاركين عبر المجموعات اللغوية/الثقافية المختلفة. يمكن ضبط بعض المتغيّرات مثل مستوى الدافعية لدى المستجيب أو الألفة بطبيعة الاختبار باستخدام الأساليب الاحصائية (على سبيل المثال Johnson, 2003; Javaras & Ripley, 2007).

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

المبادئ الإرشادية لتطوير الاختبار

(4) TD-1 التأكد من أن عمليات الترجمة ونقل الاختبار لثقافة جديدة تأخذ في الاعتبار الاختلافات اللغوية والنفسية والثقافية في المجتمعات المستهدفة من خلال اختيار الخبراء ذوي الخبرة المناسبة.

الشرح: يعد هذا المبدأ الإرشادي، على مر السنين، واحداً من أكثر المبادئ تأثيرًا لأن هناك أدلة كثيرة تشير إلى أهمية اهتمام المؤسسات الاختبارية بالبحث عن مترجمين ذوي مؤهلات تتجاوز معرفة اللغتين المشاركتين في نقل الاختبار لثقافة جديدة (انظر، على سبيل المثال، Grisay, 2003). وقد أصبحت معرفة الثقافات المتعلقة بعملية نقل الاختبار، والمعرفة العامة على الأقل بموضوع الاختبار والمعلومات عن بناء الاختبارات، جزءًا من معايير اختيار المترجمين. ويبدو أيضًا أن هذا المبدأ الإرشادي كان لها تأثير في تشجيع المؤسسات على لاستخدام اثنين على الأقل من المترجمين عند ترجمة الاختبارات ونقلها لثقافة جديدة أي كان نوع التصميم المستخدم (على سبيل المثال تصميمي الترجمة الأمامية والترجمة العكسية). وقد حذفت من قائمة الممارسات المقبولة اليوم، الممارسة القديمة المتمثلة في الاعتماد على مترجم واحد في جميع القرارات، مهما كان ذلك المترجم مؤهلا تأهيلًا جيدًا.

تنتج المعرفة والخبرة في الثقافة المستهدفة من استخدام مترجمين يتحدثون اللغة المستهدفة كلغة أم، والذي يعد شرطاً أساسياً في المترجم، ويعيشون في المنطقة الثقافية المستهدفة، والذي يعد شرطاً مرغوبًا فيه للغاية. فلن ينتج عن تحدث المترجم للغة المستهدفة كلغة أم ترجمة دقيقة فحسب، بل أيضًا ترجمة تقرأ بطلاقة وتبدو أصلية. بالإضافة إلى ذلك، فإن العيش في المنطقة الثقافية المستهدفة سيضمن معرفة حديثة للاستخدام الحالي للغة.

ومن هنا فإن تعريفنا لمصطلح "الخبير" في هذا السياق، على أنّه هو الشخص أو الفريق الذي يتمتع بمعرفة مشتركة كافية عن: (1) اللغات المتضمنة في العملية، (2) الثقافات المستهدفة، (3) محتوى الاختبار، (4) المبادئ العامّة للاختبارات؛ لإنتاج ترجمة احترافية ونقل الاختبار لثقافة جديدة بأعلى مستوى من الجودة. من الناحية العملية قد يكون من الأفضل استخدام فرق من الأشخاص بمؤهلات مختلفة (على سبيل المثال، مترجمين ذوي خبرة في موضوع الاختيار، مترجمين بدون خبرة في موضوع الاختبار، خبير في علم الاختبارات، إلخ) من أجل تحديد المجالات التي قد يغفلها الآخرون. في جميع الأحوال يجب أن تشكل المعرفة بالمبادئ العامة للاختبار، بالإضافة إلى معرفة محتوى الاختبار، جزءًا من التدريب الذي يتلقاه المترجمون.

اقتراحات للتطبيق. إليكم بعض الاقتراحات لتنفيذ هذا المبدأ الإرشادي:

- قم باختيار مترجم من مواطنين اللغة المستهدفة ولديه معرفة متعمقة بالثقافة التي تستهدف نقل الاختبار إليها، ويفضل ان يكون من سكان المكان المستهدف. من الأخطاء الشائعة التي نقع فيها أحيانا هي اننا نختار مترجمين الذين يعرفون اللغة، ولكنهم لا يعرفون الثقافة التي تنتمي إليها اللغة؛ في حين أن عملية التعمق في الثقافة المستهدفة ذات أهمية قصوى لتحقيق التكافؤ الثقافي في الاختبارات. ان معرفة المترجمين بالثقافة سوف يؤهلهم الى التعرف على المعاني الخاصة للرموز الثقافية (على سبيل المثال لعبة الكريكت، وبرج ايفل، والرئيس لينكولن، وحيوان الكنغر، إلخ). هذه الرموز لها معاني ثقافية يدركها فقط الشخص الذي يعرف الثقافة، ويستطيع أن ينقلها للثقافة الجديدة.
- قم باختيار مترجمين يتحلون، إذا أمكن بالخبرة في محتوى الاختبار، ومعرفة مبادئ القياس والتقييم (على سبيل المثال على معرفة بأنه في أسئلة الاختيار من متعدد يجب أن

ألا تكون الاجابة الصحيحة أطول أو أقصر من البدائل الأخرى؛ وأن المستجيب الذي لا يعرف الإجابة الصحيحة، قد يستخدم بعض الدلائل النحوية للمساعدة في تحديد الإجابة الصحيحة. وفي أسئلة الخطأ والصواب يجب أن يعرف أن العبارة الصحيحة لا يجب ان تكون أطول من العبارة الخاطئة بشكل ملحوظ).

- قد يكون من الصعوبة إيجاد مترجمين ذو خبرة ومعرفة بمبادئ تطوير الاختبار، لذلك يجب تدريب المترجمين على مبادئ كتابة البنود بالصيغ (مثلاً اختيار من متعدد) التي سوف يقومون بالترجمة منها. بدون هذا التدريب المترجمين، سوف يقوم بعض المترجمين الحريصين على الالتزام الشديد في الترجمة، بتقديم ترجمات تحتوى على مصادر أخطاء، مما قد يؤدى إلى خفض في صدق الاختبار المترجم. على سبيل المثال في بند من بنود الاختيار من متعدد، قد يستخدم المترجم كلمة توضيحية في البديل الصحيح أثناء الترجمة، تشير هذه الكلمة بشكل ما إلى أن هذه هي الإجابة الصحيحة، مما يجعل هذا البند أسهل من البند المقابل في الاختبار باللغة الأصلية. في بند الصواب والخطأ، قد تكون العبارة الصحيحة المترجمة للغة الجديدة، أطول من العبارات الأخرى، مما قد يعطى دليل لبعض المستجيبين أن هذه العبارة صحيحة.

(5) TD-2 استخدام تصاميم وإجراءات الترجمة المناسبة لتحقيق أقصى قدر من ملاءمة عملية نقل الاختبار للثقافة جديدة في المجتمعات المستهدفة.

<u>الشرح:</u> يتطلب هذا المبدأ الإرشادي أن تكون القرارات التي يتخذها المترجمين أو فرق المترجمين تحقق أقصى قدر من الملائمة للمجتمع المستهدف. يعنى هذا أنه يجب أن تكون اللغة مقبولة وطبيعية وتركز على التكافؤ الوظيفي وليس الحرفي بين النسخة الأصلية للاختبار والنسخة المنقولة للثقافة الجديدة. من الأنماط الشائعة لتصميمات الترجمة لتحقيق هذه الأهداف هي الترجمة الأمامية forward translations، والترجمة العكسية backward translations. قدم بريسلين

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

(Brislin,1986) وهامبلتون وباتسولا (Hambleton and Patsula,1999) مناقشات مفصلة حول التصميمين، بما في ذلك تعريفهما ومواطن القوة والضعف فيهما، ولكن يجب التنويه الى ان كلا من التصميمين له عيوبه الخاصة، ونادرًا ما يزود هذين التصميمين من الترجمة بأدلة كافية للتحقق من صدق تكافؤ النص الأصلي، والنص المنقول لثقافة جديدة. والعيب الرئيسي للترجمة العكسية هو أنه اذا تم تنفيذ هذا النمط في أضيق صوره، فلن يتم مراجعة النسخة المنقول لها الاختبار، وغالبًا ما ينتج عن استخدام هذا التصميم، اصدار نسخة باللغة الجديدة المستهدفة للاختبار بحيث تعطي اقصى درجة ممكنة من المقاربة مع الاختبار الأصلي، إذا ما تم أعادة ترجمتها للغة الأصلية، بصرف النظر عن الإخلال في المعنى الذي قد ينتج من هذه العملية.

يهدف إجراء الترجمة المزدوجة double-translation والمقاربة بين النسختين إلى معالجة أوجه القصور والمخاطر الناجمة عن الاعتماد على خصوصيات الترجمة من لغة إلى لغة أخرى فقط. وفي هذا النهج، يقوم مترجم مستقل ثالث أو فريق خبراء بتحديد أي تناقضات بين الترجمات البديلة وحلها، حيث يتم المقاربة والتوافق بينها ودمجها في نسخة واحدة. في برامج التقييم عبر الثقافات واسعة النطاق مثل PSIA، يمكن استخدام نسختين مختلفتين من اللغات (على سبيل المثال الإنجليزية والفرنسية) كمصادر منفصلة للترجمة، ثم يتم التوفيق بينهما في نسخة واحدة باللغة المستهدفة (Grisay, 2003). ويوفر هذا النهج مزايا هامة، مثل تحديد التناقضات المحتملة ومراجعتها بشكل مباشر باللغة المستهدفة. بالإضافة إلى ذلك، يساعد استخدام أكثر من لغة مصدرية (اصلية) واحدة على تقليل تأثير الخصائص الثقافية للمصدر.

يمكن أن تتسبب الاختلافات في بنية اللغة في حدوث مشكلات في ترجمة الاختبار، على سبيل المثال، في المقياس المعروف الذي طوره روتر ورافرتي (Rotter and Rafferty, 1950) لإكمال الجمل باللغة الإنجليزية، يطلب من المستجيبين إكمال الفراغات في البنود غير مكتملة مثل

" أحب. ...."، "أنا نادم على...."، " لا أستطيع ....."، غير أن الصيغة نفسها كانت غير ملائمة في اللغة التركية، حيث يجب أن يأتي المفعول به قبل الفعل والفاعل، وبالتالي فإن استخدام الجمل غير المكتملة كما هو الحال في النسخة الإنجليزية من شأنه أن يغير نمط الإجابة تمامًا لأن الطلاب الأتراك يجب أن ينظروا أولًا إلى نهاية الجملة قبل تعبئة البداية. أية حلول بديلة لهذه المشكلة، ستؤدي إلى نسخة مترجمة تختلف بطريقة أو بأخرى، عن النسخة الأصلية من حيث مواصفات التنسيق.

اقتراحات للتطبيق. يبدو أن تجميع البيانات الصادرة عن المراجعين قد تكون ذات قيمة خاصة للتحقق من استيفاء هذا المبدأ الإرشادي:

- استخدم مقاييس التقدير التي قدمها بريسلين (Brislin, 1986)، جينري وبرتراند (Jeanrie, & Bertrand, 1999)، أو هامبلتون وزينيسكي (Hambleton, & Zenisky 2010). قدم هامبلتون وزينيسكي قائمة مكونة من 25 خاصية مختلفة يجب فحصها في الاختبار المترجم، يجب التحقق من استيفاؤها أثناء عملية نقل الاختبار لثقافة جديدة. من أمثلة بنود هذه القائمة (Hambleton, & Zenisky 2010): هل لغة البند المترجم تتسم بصعوبة مماثلة وقواسم مشتركة فيما يتعلق بالكلمات الموجودة في البند في الإصدار الأصلي؟ هل تضمنت الترجمة إحداث تغييرات في النص (الحذف أو الاستبدال أو الإضافات) التي قد تؤثر على صعوبة البند في نسختي الاختبار؟
- استخدم تصميمات ترجمة متعددة إذا كان ذلك ممكناً عملياً، على سبيل المثال، يمكن استخدام تصميم الترجمة العكسية للتحقق مرة أخرى من الإصدار المستهدف الذي تم إنشاؤه من خلال الترجمة المزدوجة والتسوية التي تمت بواسطة لجنة من الخبراء.

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

- إذا كان من المقرر استخدام الاختبار عبر الثقافات، ففكر في التطوير المتزامن لإصدارات لجميع اللغات المستهدفة من البداية، لتجنب المشاكل المستقبلية في ترجمة ونقل الاختبار لثقافة جديدة في النسخة الأصلية. يمكن العثور على مزيد من المعلومات حول تطوير الاختبار المتزامن simultaneous test development، في (Solano-Flores, Trumbull, & Nelson-Barber, 2002)، وعلى أقل تقدير صمم النسخة الأصلية منذ البداية، بحيث تتيح الفرصة في الحصول على ترجمات مستقبلية، وتتجنب المشاكل المحتملة قدر الإمكان. على وجه التحديد عند تطوير النسخة الأصلية تجنب قدر الإمكان الركوز الثقافية الخاصة، وتنسيقات البنود التي لا تخص ثقافة محددة دون أخرى.
- وبالنظر إلى الاختلافات في بناء الجملة بين اللغات، ينبغي تجنب استخدام الصيغ التي تعتمد على البنية الجامدة للجمل، خاصة في الاختبار الدولية الواسعة النطاق، وربما أيضا في الاختبارات النفسية بسبب مشاكل الترجمة التي قد تنشأ من استخدام هذه الصيغ.

(6) TD-3 تقديم أدلة على أن تعليمات الاختبار ومحتوى البنود لها معنى مماثل لجميع المجتمعات المستهدفة.

<u>الشرح:</u> يمكن جمع الأدلة التي يتطلبها هذا المبدأ الإرشادي من خلال مجموعة متنوعة من الاستراتيجيات، (انظر، على سبيل المثال،van de Vijver and Tanzer, 1997)، وتشمل هذه الاستراتيجيات (1) استخدام المراجعين الناطقين باللغة المستهدفة كلغة أم ويعيشون في نفس الثقافة المستهدفة، (2) استخدام عينات من المستجيبين ثنائي اللغة bilingual، (3) استخدام الاستبيانات المحلية لتقييم الاختبار، (4) تطبيق الاختبار بالطرق غير المعيارية (التعليمات المعدلة) لزيادة القبول والصدق للاختبار في الثقافة الجديدة.

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

يعد إجراء تجربة استطلاعية صغيرة للنسخة المعدلة من الاختبار فكرة جيدة، يمكن أن تستخدم التجربة الصغيرة ليس فقط في تطبيق الاختبار وتحليل البيانات، ولكن أيضاً، والأهم من ذلك، المقابلات مع المسؤولين والمستجيبين للحصول على انتقاداتهم للاختبار نفسه. من الممكن أيضاً استخدام تصميمات أخرى باستخدام خبراء المحتوى من خلفيات لغوية مختلفة، أو خبراء المحتوى ثنائي اللغة، وعلى سبيل المثال يمكن أن يطُلب من خبراء المحتوى ثنائي اللغة تقييم التشابه في صعوبة صيغ البنود ومحتوى الاختبارين. من التصميمات الجديدة المبشرة، إجراء المقابلات المعرفية cognitive interviewing (Levin, et al., 2009).

<u>اقتراحات للتطبيق</u>. تم تقديم العديد من الاقتراحات لتنفيذ هذا المبدأ الإرشادي؛ على سبيل المثال:

- استخدم المراجعين الناطقين باللغة المستهدفة كلغة أم ويعيشون في نفس الثقافة المستهدفة، لتقييم ترجمة الاختبار ونقل الاختبار لثقافة جديدة.
- استخدم عينات من المستجيبين ثنائي اللغة لتقديم بعض الاقتراحات حول تكافؤ نسختين من الاختبار، سواء في تعليمات الاختبار أو بنود الاختبار.
- استخدم الاستطلاعات المحلية لتقييم الاختبار، ويمكن أن تكون هذه التجارب الصغيرة ذات قيمة كبيرة. تأكد من إجراء مقابلة مع المطبقين للاختبار والمستجيبين بعد تطبيق الاختبار؛ لأنه غالبًا ما تكون تعليقات المطبقين والمستجيبين أكثر قيمة من ردود المستجيبين الفعلية للبنود الموجودة في الاختبار.
- استخدم التعليمات المعدلة للاختبار المنقول لزيادة القبول والصدق، فإن اتباع تعليمات اختبار مماثلة لا معنى له إذا أسيء فهمها من قبل المستجيبين في اللغة الثانية/ المجموعة الثقافية المستهدفة.

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

(7) TD-4 تقديم أدلة على أن صيغ البنود، ومقاييس التقدير، وفئات التسجيل وتعليمات الاختبار، وطرق التطبيق، وغيرها من الإجراءات مناسبة لجميع المجتمعات المستهدفة.

الشرح: صيغ البنود مثل مقاييس التقدير ذات النقاط الخمس أو صيغ البنود الجديدة مثل السحب والإفلات" أو "الإجابة على كل ما هو صحيح" أو حتى "الإجابة على خيار واحد وفقط إجابة واحدة"، يمكن أن تكون مربكة بالنسبة للمستجيبين الذين لم يروا صيغ البنود هذه من قبل. حتى تخطيطات البنود item layouts، أو استخدام الرسومات أو صيغ البنود المحوسبة التي تظهر بسرعة يمكن أن تكون مربكة للمستجيبين في الثقافة الجديدة. هناك أمثلة كثيرة على هذه الأنواع من الأخطاء التي وجدت في الولايات المتحدة، مع مبادرتها بنقل الكثير من الاختبارات المقننة للأطفال إلى الحاسوب. لكنه وُجد أنه من خلال التدريبات التطبيقية على الصيغ الغير مألوفة، يمكن التغلب على المشاكل بالنسبة لمعظم الأطفال. يجب أن تكون صيغ البنود الجديدة مألوفة للمستجيبين، وإلا فإن هذه الصيغ ستكون مصدر من مصادر التحيز في الاختبار يمكن أن يشوه درجات الاختبار الناتجة عن التطبيق الفردي أو الجماعي له.

من المشاكل الحديثة نسبياً، تلك المتعلقة بإصدارات الاختبار التي يديرها الحاسوب. إذا لم يكن المستجيبون على دراية بمنصة الاختبار المعتمدة على الحاسوب، فهناك حاجة إلى بعض التدريب للتأكد من أن هؤلاء المستجيبين يكتسبون الألفة التي يحتاجونها للاختبار الذي يديره الحاسوب لتقديم درجات ذات معنى.

اقتراحات للتطبيق. يمكن استخدام كل من الأدلة الكيفية والكمية للتحقق من تنفيذ هذا المبدأ الإرشادي، فهناك العديد من الخصائص في الاختبار المنقول إلى الثقافة الجديدة، يجب التحقق منها فيما يتعلق بهذا المبدأ الإرشادي:

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

- تأكد من أن التدريب على طريقة الاستجابة على الاختبار كانت كافية لرفع المستجيبين إلى المستوى المطلوب لهم لتقديم إجابات صادقة أو ردود تعكس مستوى إتقانهم للمادة موضع الاختبار.
- تأكد من أن المستجيبين على دراية بصيغ البنود الجديدة، أو إدارات اختبار غير التقليدية (مثل إدارة الحاسوب)، والتي تعد جزء من بنية الاختبار.
- تحقق من أن تعليمات الاختبار الخاصة مفهومة بشكل واضح للمستجيبين (على سبيل المثال كيفية الاستجابة الصحيحة على البند).
- مرة أخرى، قوائم التقدير التي قدمها جانري وبرتراند (Jeanrie, & Bertrand, 1999) وهامبليتون وزينيسكي (Hambleton, Zenisky, 2010) تعد مفيدة فيما يتعلق بهذا المبدأ الإرشادي. على سبيل المثال، قام كل من هامبليتون وزينيسكي بوضع أسئلة في قائمتهما مثل: "هل صيغ البنود، بما في ذلك التخطيط الفعلي للبند، هي نفسها في نسختي الاختبار؟"، و"إذا كان هناك نوع من التأكيد على الكلمة أو العبارة (غامق، مائل، تحته خط، إلخ)، في أحد البنود في الاختبار الأصلي، هل تم استخدام نفس هذا التأكيد في البند المترجم؟

(8) TD-5 جمع البيانات الاستطلاعية من النسخة الاختبارية الجديدة حتى يمكن تحليل البنود، وتقييم الثبات ودراسة الصدق (على نطاق صغير) بحيث يمكن إجراء أية مراجعات ضرورية للنسخة الاختبارية الجديدة.

<u>الشرح:</u> قبل الشروع في إجراء أي نوع من التحقق من الصدق والثبات أو الحصول على معايير للاختبار على نطاق واسع، من المهم أن تكون هناك أدلة تثبت جودة نقل الاختبار من الناحية السيكومترية؛ حيث أن إجراءات التحقق من الصدق والثبات والحصول على المعايير على

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

نطاق واسع هي عملية مكلفة مادياً وتأخذ وقت طويل، ولا يجب إجراؤها قبل التأكد من جودة عملية نقل الاختبار للثقافة الجديدة. هناك العديد من التحليلات السيكومترية التي يمكن إجراؤها لتقديم أدلة أولية على ثبات الدرجات والصدق. على سبيل المثال في مرحلة تطوير الاختبار، يمكن لتحليل البنود باستخدام حجم عينة مناسب (على سبيل المثال، 100) أن يوفر بيانات أولية نحتاجها بشدة في هذه المرحلة، حول أداء بنود الاختبار موضع الاهتمام. يمكن مراجعة البنود، التي تكون سهلة جداً أو صعبة مقارنة بالبنود الأخرى، أو التي تظهر قوى تمييز منخفضة أو حتى سلبية، بحثاً عن عيوب محتملة في الصياغة الجديدة للبنو. إذا كانت البنود من نوع الاختيار من متعدد، سيكون من المناسب التحقيق في فعالية المشتتات (البدائل الخاطئة). يمكن رصد المشاكل مبكراً، وإجراء المراجعات اللازمة. أيضاً مع نفس البيانات التي تم تجميعها لتحليل البنود، يمكن الحصول على معامل ألفا أو معامل أوميغا (McDonald, 1999)، والتي تمد مطور الاختبار بمعلومات قيمة يمكن استخدامها لدعم القرارات المتعلقة بالطول المناسب لإصدارات الاختبار الأصلي والجديد.

في بعض الحالات، قد تظل هناك أسئلة حول جوانب معينة من نقل الاختبار: هل تم فهم تعليمات الاختبار تماماً؟ هل يجب أن تكون التعليمات مختلفة لتوجيه المستجيبين للاختبار بشكل فعال في اللغة والثقافة الجديدة؟ هل سيتسبب الاختبار الذي يتم إجراؤه بواسطة الحاسوب في حدوث مشكلات لدى بعض المستجيبين (على سبيل المثال، المستجيبون ذوو الحالة الاجتماعية والاقتصادية المنخفضة) في المجتمع المستهدف نقل الاختبار له؟ هل عدد الأسئلة لا يتناسب مع الزمن المتاح للاستجابة؟ يمكن الإجابة على كل هذه الأسئلة وغيرها من خلال دراسات الصدق المصغرة. إن الهدف هنا هو تجميع بيانات كافية لاتخاذ قرار بشأن المضي قدماً في إجراءات نقل الاختبار أم لا، وإذا كان القرار هو المضي قدماً، فيمكن تخطيط وتنفيذ سلسلة من الدراسات على

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

نطاق أوسع (على سبيل المثال، دراسات مستوى الأداء التفاضلي للبند DIF، ودراسات للتحقق من البناء العاملي للاختبار).

اقتراحات للتطبيق. هناك عدد من التحليلات الأساسية التي يمكن إجراؤها:

- إجراء دراسة التحليل الكلاسيكي للبنود للحصول على معلومات حول متوسطات البنود (مؤشر الصعوبة) ومؤشر تمييز البند، ومع بنود الاختيار من متعدد، أو البنود المماثلة التي تعتمد على الاختيارات، قم بإجراء تحليل المشتتات أيضاً.
- إجراء تحليل الثبات (على سبيل المثال، استخدام معادلة كيودر–ريتشاردسون (20) KR–20 مع البنود ثنائية الدرجة (لها إجابة واحدة صحيحة)، أو معامل ألفا أو معامل أوميغا مع البنود متعددة الاستجابات).
- عند الضرورة، قم بإجراء دراسة أو اثنتين لتفحص صدق الاختبار بعد نقله للثقافة الجديدة. على سبيل المثال، افترض أن نقل الاختبار سيتم إجراؤه عبر الحاسوب؛ قد يكون من المفضل في هذه الحالة إجراء دراسة لتقييم طريقة إدارة وتطبيق الاختبار (أي تطبيق الاختبار باستخدام الورق والقلم في مقابل تطبيق الاختبار بواسطة الحاسوب). بافتراض أن التعليمات تتطلب من المستجيبين الإجابة على جميع الأسئلة؛ قد يكون من الضروري إجراء بعض الدراسات لتحديد أفضل التعليمات لتحقيق هذا الهدف. توصل الباحثون أنه من الصعب بشكل مفاجئ إقناع بعض المستجيبين بالإجابة على كل سؤال، وأن هذا الاجراء قد يشجع اللجوء للتخمين عندما لا يكون لدى المستجيبين المعلومات المطلوبة.

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

المبادئ الإرشادية للتحقق

المبادئ الارشادية للتحقق confirmation، هي تلك المبادئ التي تستند إلى التحليلات التجريبية (الأمبريقية) للدراسات الشاملة للصدق.

(9) C–1 اختيار عينة ذات خصائص متعلقة بالاستخدام المقصود للاختبار وبحجم كافي للتحليلات الأمبريقية التي ستقوم بها.

الشرح: يشير تصميم جمع البيانات إلى الطريقة التي يتم بها جمع البيانات لوضع المعايير (إذا لزم الأمر) والتكافؤ بين النسخ المختلفة اللغة للاختبار، ولإجراء دراسات الصدق والثبات، ودراسة الأداء التفاضلي للبنود. الشرط الأول فيما يتعلق بجمع البيانات هو أن العينات يجب أن تكون كبيرة بما يكفي للسماح بالتوصل لنتائج إحصائية مستقرة. على الرغم من أن هذا المطلب ينطبق على أي نوع من الدراسات، إلا أنه أكثر أهمية في سياق دراسة التحقق من صدق نقل الاختبار. ذلك أن الأساليب الإحصائية اللازمة لإنشاء اختبار وتكافؤ البنود (على سبيل المثال، التحليل العاملي التوكيدي CFA، وتحليلات نظرية الاستجابة للبند IRT لتحديد البنود المحتمل تحيزها) تعطي نتائج ذات معنى عند استخدامها مع عينات كبيرة الحجم بما يكفي لتقدير معلمات (بارامترات) النموذج بشكل موثوق ومستقر (يعتمد حجم العينة الموصي به على مدى تعقيد النموذج وجودة البيانات).

كما ينبغي أن تكون عينة الدراسة الشاملة للصدق ممثلة للمجتمعات المستهدفة من نقل الاختبار، ونوجه الانتباه إلى الورقة الهامة التي أعدها فان دي فيفر وتانزر (van de Vijver & Tanzer, 1997)، والمساهمات المنهجية الموجودة في فان دي فيفر وليونغ (van de Vijver & Leung, 1997)، وهامبلتون وميريندا وسبيلبيربيرغر (Hambleton, Merenda, &

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

(Spielberger, 2005، وبايرن (Byrne, 2008)، وبايرن وفان دي فيفر (Byrne, & van der Vijver, 2014)، لإرشادات اختيار التصاميم والتحليلات الإحصائية المناسبة، وقدم سيرتشي (Sireci, 1997) مناقشة للمشاكل والقضايا المتعلقة بربط الاختبارات متعددة اللغات بمقياس مشترك.

في بعض الأحيان، من الناحية العملية، قد يحصل المجتمع المستهدف على درجات درجات أقل أو أعلى بكثير من المجتمع الأصلي، أو يكون المجتمع المستهدف أكثر أو أقل تجانساً من المجتمع الأصلي. قد يخلق هذا مشاكل كبيرة فيما يتعلق ببعض طرق التحليل، خاصة تلك التحليلات المتعلقة بدراسات الثبات والصدق. يتمثل أحد الحلول في اختيار عينة فرعية من المجتمع الأصلي تماثل عينة مجموعة اللغة المستهدفة، وباستخدام العينات المتطامثلة، يمكن التخلص من أي اختلافات في نتائج العينات المتماثلة، والتي قد تكون ناجمة عن الاختلافات في أشكال التوزيعات في المجموعتين (انظر Sireci & Wells, 2010) . على سبيل المثال، تشتمل مقارنات بنية الاختبار عادة على متغيرات مصاحبة، وستختلف هذه المتغيرات كدالة لتوزيعات الدرجات. عند استخدام المجموعات المتماثلة، فإن توزيع الدرجات في العينتين يكون متماثل، مما يعني تحييد الدور الذي ممكن أن يلعبه توزيع الدرجات في النتائج، كتفسير للاختلاف في النتائج.

ربما يساعد مثال آخر في شرح مشكلة توزيع الدرجات المختلفة في مجموعات اللغة الأصلية واللغة المستهدفة. افترض أن ثبات درجة الاختبار هي (80.) في مجموعة اللغة الأصلية، في حين أن الثبات في مجموعة اللغة المستهدفة هو (60.) فقط. يبدو هذا الاختلاف مقلقاً ويثير تساؤلات حول مدى ملاءمة إصدار الاختبار باللغة المستهدفة. ومع ذلك، غالباً ما يتم التغاضي عن أن الثبات هي خاصية مشتركة للاختبار والمجتمع الذي يطبق فيه الاختبار (McDonald,

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

(1999 لأن الثبات يعتمد على كل من تباين الدرجة الحقيقية (خاصية في المجتمع) وتباين الخطأ (خاصية في الاختبار). لذلك يمكن أن يؤدي تباين الخطأ نفسه إلى ثبات أعلى ببساطة بسبب التباين الأكبر في الدرجة الحقيقية في مجموعة اللغة الأصلية. يوضح ماكدونالد (McDonald, 1999) أن الخطأ المعياري للقياس (وهو الجذر التربيعي لتباين الخطأ) هو في الواقع الاحصاءة الأكثر ملاءمة للمقارنة بين العينات في هذه الحالة، وليس الثبات. بديل آخر باستخدام معاملات الثبات هو سحب عينة مماثلة من المستجيبين من مجموعة اللغة الأصلية وإعادة حساب ثبات درجات الاختبار.

تسمح الأساليب الحديثة لاختبار تكافؤ القياس measurement invariance باستخدام التحليل العاملي التوكيدي متعدد المجموعات (multi-group CFA) بتقييم العينات ذات التوزيعات المختلفة للسمات الكامنة. في هذه النماذج الاحصائية، يمكن افُتراض أن معلمات (بارامترات) القياس مثل تشبعات البنود item-factor loadings ومتوسطات البنود متساوية عبر المجموعات، وفي نفس الوقت، يمكن السماح بعدم تساوي المتوسطات والتباينات والتغايرات للسمات (المتغيرات) الكامنة عبر المجموعات. هذا الاجراء يسمح باستخدام العينات كما هي، حيث يستوعب السيناريو الأكثر واقعية والذي يتمثل في اختلاف توزيعات السمات المقاسة عبر المجتمعات المختلفة.

<u>اقتراحات للتطبيق</u>. في جميع الأبحاث تقريباً، هناك اقتراحان يتم تقديمهما عند وصف العينة (أو العينات):

جمع عينة كبيرة بقدر معقول نظرا لأن الدراسات التي تستهدف تحديد بنود الاختبار التي يحتمل أن تكون متحيزة تتطلب ما لا يقل عن 200 شخص لكل نسخة من الاختبار (Mazor, Clauser & Hambleton, 1992; Subok, 2017) . لإجراء تحليلات ونماذج نظرية الاستجابة للبند

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

المناسبة، يلزم عينة من 500 مستجيب على الأقل (Hulin, Lissak & Drasgow, 1982; Hambleton, Swaminathan & Rogers, 1991)، وفي حين أن الدراسات التي تستهدف التحقق من البناء العاملي للاختبار تتطلب أحجام عينات كبيرة إلى حد ما، ربما 300 أو أكثر من المجيبين (Wolf, Harrington, Clark & Miller, 2013). من الواضح أنه من الممكن القيام بهذه التحليلات باستخدام عينات اصغر حجماً، لكن القاعدة الذهبية هنا هي استخدام عينات كبيرة من المشاركين كلما أمكن ذلك.

- اختر عينات تمثل مجتمع المستجيبين كلما أمكن ذلك. تعميمات النتائج من عينات غير ممثلة تكون محدودة. للتخلص من الاختلافات في النتائج بسبب العوامل المنهجية مثل الاختلافات في توزيع الدرجات، عادة ما يكون سحب عينة من مجموعة اللغة الأصلية مماثلة لمجموعة اللغة المستهدفة فكرة جيدة. قد تكون مقارنات الأخطاء المعيارية للقياس أكثر ملاءمة من مقارنة معاملات الثبات.

(10) C–2 تقديم أدلة إحصائية ذات صلة حول التكافؤ البنائي، والتكافؤ المنهجي، والتكافؤ على مستوى البند لجميع المجتمعات المستهدفة.

<u>الشرح:</u> يعد التحقق من التكافؤ البنائي لنسخ الاختبار باللغة الأصلية واللغة المستهدفة أمراً مهماً، ولكن ليس فقط التحليل التجريبي (الإمبريقي) فقط. تناولت هذه القائمة الارشادية بصورة موجزة المناهج المتبعة في مجال التكافؤ البنائي (PC–2) و التكافؤ المنهجي (PC–3).

يحتاج الباحثون إلى معالجة التكافؤ على مستوى البند، وليس على مستوى الاختبار فقط. وتتم دراسة التكافؤ على مستوى البند تحت عنوان " تحليل الأداء التفاضلي/التمييزي للبند". بشكل عام، يحدث الأداء التفاضلي/التمييزي للبند إذا كان اثنين من المتقدمين للاختبار، من مجموعتين

مختلفتين (ثقافياً، لغوياً)، ولديهما نفس المستوى في السمة المقاسة ولكن لديهما احتمال استجابة مختلف على بند الاختبار. يمكن أن تحدث اختلافات عامة في أداء الاختبار عند المجموعات، ولكن هذه الاختلافات لا تمثل مشكلة بمفردها. تظهر المشكلة عندما يتم مطابقة المجتمعات ذات العلاقة على البنية المقاسة بالاختبار(عادة ما يكون مجموع درجات الاختبار، أو مجموع درجة الاختبار ناقص درجة البند الذي تتم دراسته)، ومع ذلك توجد اختلافات في الأداء على البند عبر المجموعات، عند يحدث ذلك نقول أن هذا البند لديه أداء تفاضلي أو تمييزي. هذا النوع من التحليل يتم تنفيذه لكل بند في الاختبار. وفي وقت لاحق، يتم إجراء محاولة لفهم أسباب ظهور بعض البنود بأداء تفاضلي (تمييزي). بناءً على نتائج هذا التحليل، قد يتم تحديد بعض البنود على أنها معيبة، ويتم تغييرها، أو إزالتها تمامًا من الاختبار.

ثمة مصدران محتملان ومهمان لتقييمهما لظهور الأداء التفاضلي للبند؛ هما مشاكل الترجمة والاختلافات الثقافية. وبشكل أكثر تحديدًا، قد يكون سبب الأداء التفاضلي/التمييزي للبند هو (1) عدم تكافؤ الترجمة التي تحدث من اللغة الأصلية إلى نسخة اللغة المستهدفة للاختبار، مثل عدم إلمام المستجيبين بالمفردات اللغوية المستخدمة في الاختبار المترجم، والتغيير في صعوبة البند، والتغيير في تكافؤ المعنى، إلخ. و (2) الاختلافات الثقافية السياقية. (Scheuneman & Grima, 1997; van de Vijver & Tanzer, 1997; Ercikan, 1998, 2002; Allalouf, Hambleton, & Sireci, 1999; Sireci & Berberoğlu, 2000; Ercikan, et al., 2004; Li, Cohen, & Ibera, 2004; Park, Pearson & Reckase, 2005; and Ercikan, Simon, & Oliveri, 2013).

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

أثناء الترجمة، هناك إمكانية استخدام مفردات أقل شيوعًا في اللغة المستهدفة. يمكن أن تكون المعاني هي نفسها في النسخ المترجمة، ولكن في ثقافة ما، يمكن أن تكون الكلمة أكثر شيوعًا مقارنة بشيوعها في ثقافة أخرى. من الممكن أيضًا تغيير مستوى صعوبة البند نتيجة الترجمة بسبب طول وتعقيد الجملة واستخدام المفردات السهلة أو الصعبة كذلك. علاوةً على ذلك، قد يتغير المعنى في اللغة المستهدفة: من خلال حذف بعض من أجزاء الجمل، الترجمات غير الدقيقة، وجود أكثر من معنى واحد في المفردات المستخدمة في اللغة المستهدفة، والانطباعات غير المتكافئة لمعنى بعض الكلمات عبر الثقافات، وإلخ. قبل كل شيء، قد تتسبب الاختلافات الثقافية في سلوك البنود بشكل مختلف عبر اللغات. على سبيل المثال، قد لا تكون كلمات مثل "همبرغر" أو "سجل نقدي cash register" مفهومة، أو لها معنى مختلف في ثقافتين.

هناك أربع مجموعات على الأقل من التحليلات للتحقق مما إذا كانت البنود تعمل بشكل مختلف عبر اللغة و / أو المجموعات الثقافية. وتشتمل على (أ) إجراءات نظرية الاستجابة المفردة (انظر على سبيل المثال، Ellis, 1989; Thissen, Steinberg, & Wainer, 1988; 1993; Ellis & Kimmel, 1992)، (ب) اجراء مانتل-هانزل Mantel-Haenszel (MH) (انظر أمثلة، Dorans & Holland, 1993; Hambleton, Clauser, Mazor, & Jones, 1993; Holland & Wainer, 1993; Sireci & Allalouf, 2003)، (ج) إجراءات تحليل الانحدار اللوجستي Logistic Regression (LR) (Swaminathan & Rogers, 1990; Rogers & Swaminathan, 1993)، و(د) اجراء تحليل العاملي المقيد restricted factor analysis (RFA) (Oort & Berberoğlu, 1992).

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

في إجراءات نظرية الاستجابة المفردة، يتم مطابقة المتقدمين للاختبار عبر لغتين استنادًا إلى درجات السمة الكامنة. وفي منهجيتي مانتل–هانزل، وتحليل الانحدار اللوجستي تستخدم درجة الاختبار الملاحظة أو المقدرة كمحك للمطابقة قبل مقارنة أداء البنود للمستجيبين في المجموعتين. على الرغم من أن مجموع الدرجات على الاختبار هو محك المطابقة الأكثر شيوعاً في هذه الإجراءات، يمكن أيضا استخدام درجات أخرى مقدرة، على سبيل المثال الدرجات العاملية من التحليل العاملي. يتم أيضًا "تنقية" هذه الدرجات بشكل متكرر عن طريق حذف البنود ذات الأداء السئ. يجب أن يكون محك المطابقة صادقًا وثابتًا بدرجة كافية لتقييم الأداء التفاضلي/التمييزي للبند بشكل صحيح. عند إجراء التحليل العاملي المقيد، يتم انحدار كل بند على متغير الانتماء للمجموعة وكذلك على السمة الكامنة موضع القياس. يتم تقدير تشبع كل بند على المتغير الكامن، ومن ثم تقييم ملاءمة النموذج مقارنةً بالنموذج المرجعي (النموذج الذي لا يتشبع فيه أي بند على متغير الانتماء للمجموعة، حيث يمثل الحالة المثالية لعدم وجود أي أداء تفاضلي لأي بند). إذا حقق النموذج ملاءمة أفضل بشكل دال إحصائياً، فسيتم وضع علامة على البند على أن له أداء تفاضلي.

عندما تكون أبعاد الاختبار معقدة، فإن العثور على محك مطابقة مناسب يمثل مشكلة (Clauser, Nungester, Mazor & Ripkey, 1996). قد يؤدي استخدام محكات المطابقة متعددة المتغيرات multivariate matching criteria، مثل الدرجات العاملية المختلفة التي تم الحصول عليها نتيجة للتحليل العاملي، إلى تغيير مستوى تفسيرات البند في الأداء التفاضلي للبند. وفقًا لذلك، يشير هذا المبدأ الاسترشادي إلى أنه إذا كان الاختبار متعدد الأبعاد، فقد يستخدم الباحثون محكات مختلفة لتمييز البنود التي لها أداء تفاضلي، وتقييم البنود التي يتم تمييزها باستمرار

على أن لها أداء تفاضلي عند استخدام محكات المطابقة المختلفة. يمكن أن تقلل المطابقة المتعددة المتغيرات من عدد البنود التي تظهر أن لها أداء تفاضلي عبر المجموعات اللغوية والثقافية.

قد تتطلب هذه المناهج أحجام عينات مختلفة. فإن اجراء مانتل–هانزل، تحليل الانحدار اللوجستي، والتحليل العاملي المقيد هي من تلك النماذج التي يمكن تعمل بشكل ثابت وصادق باستخدام عينات صغيرة نسبيًا مقارنة بالتقنيات المستندة إلى نظرية الاستجابة المفردة، والتي تتطلب عينات أكبر لتقديرات المعلمات بشكل صادق. وثمة اعتبار آخر هو نوع البيانات التي تنتج من الاستجابة على البند. يمكن تطبيق اجراء مانتل–هانزل، وتحليل الانحدار اللوجستي، والتحليل العاملي المقيد على البيانات الثنائية. وهناك حاجة إلى مناهج أخرى، مثل اجراء مانتل–هانزل المعمم، مع بيانات الاستجابة المتعددة.

يتطلب هذا المبدأ الإرشادي من الباحثين تحديد المصادر المحتملة للتحيز المنهجي في الاختبار المنقول إلى ثقافة أخرى. وتشمل مصادر التحيز المنهجي (1) المستويات المختلفة لدافعية المشاركين المستجيبين على الاختبار (2) الحبرة التفاضلية من جانب المستجيبين الذين خضعوا لاختبارات نفسية من قبل (3) سرعة أكبر في تطبيق الاختبار في مجموعة لغوية واحدة عن المجموعة الأخرى (4) الإلمام المتباين بتنسيق وطريقة الاستجابة عبر المجموعات اللغوية، و(5) عدم التجانس في أسلوب الاستجابة، وغيرها. كانت التحيزات في الاستجابات، على سبيل المثال، مصدر قلق كبير في تفسير نتائج PISA، وخضعت لبعض الإجراءات البحثية.

وأخيرًا، وليس أقل أهمية، يتطلّب هذا المبدأ الإرشادي من الباحثين التحقق من التكافؤ البنائي. هناك أربعة نهج إحصائية على الأقل لتقييم التكافؤ البنائي عبر النسخ المختلفة اللغات والثقافات من الاختبار: التحليل العاملي الاستكشافي EFA، والتحليل العاملي التوكيدي CFA،

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

والتقييس متعدد الأبعاد multidimensional scaling (MDS)، ومقارنة الشبكات comparison of nomological networks (Sireci, Patsula, & Hambleton, 2005).

وفقًا لما ذكره "فان دي فيفر وبورتينيغا" (van der Vijver & Poortinga, 1991)، فإن كلا من التحليل العاملي الاستكشافي والتحليل العاملي التوكيدي هي التقنيات الإحصائية الأكثر استخدامًا لتقييم ما إذا كان يوجد بناء في ثقافة واحدة بنفس الشكل والتواتر في ثقافة أخرى. هذه النتيجة من عام 1991، لكنها ما زالت صحيحة إلى اليوم، على الرغم من أن منهجيات النمذجة الإحصائية قد تقدمت بشكل كبير (انظر على سبيل المثال، Hambleton & Lee, 2013, Byrne, 2008). بما أنه من الصعب مقارنة أبنية عاملية منفصلة، في التحليل العاملي الاستكشافي، ولا توجد قواعد متفق عليها عموما لتحديد متى يمكن اعتبار الأبنية العاملية متكافئة، فإن النهج الإحصائية مثل التحليل العاملي التوكيدي (انظر على سبيل المثال، Byrne, 2001, 2003, 2006, 2008) والقياس متعدد الأبعاد الموزون weighted multidimensional scaling، قد يكون مرغوباً فيهما بشكل أكبر؛ حيث يمكنهما استيعاب مجموعات متعددة في نفس الوقت (Sireci, Harter, Yang, & Bhola, 2003).

هناك العديد من الدراسات التي تم فيها استخدام التحليل العاملي التوكيدي لتقييم ما إذا كانت البنية العاملية في النسخة الأصلية للاختبار متسقة عبر النسخ المنقولة إلى ثقافات أخرى (على سبيل المثال ،Byrne & van de Vijver, 2014). يعد التحليل العاملي التوكيدي من ضمن التحليلات التي تتلاءم بشكل كبير مع تقييم التكافؤ البنائي عبر الاختبارات المنقولة لأنه يمكنه التعامل مع مجموعات متعددة في وقت واحد. وتتوفر الاختبارات الإحصائية، والمؤشرات الوصفية لملائمة النموذج (Sireci, Patsula, & Hambleton, 2005). إن القدرة على التعامل

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

مع مجموعات متعددة مهمة بشكل خاص حيث أصبح من الشائع نقل الاختبارات إلى العديد من اللغات على سبيل المثال، يتم الآن ترجمة/نقل بعض مقاييس الذكاء إلى أكثر من مائة لغة مختلفة، وفي TIMSS، و OECD/PISA يتم تقنين الاختبارات إلى أكثر من 30 لغة. وبما أن الشرط الصارم المتمثل في عدم تشبع أي بند على مكونين كامنين مختلفين في التحليل العاملي التوكيدي، فإنه غالبًا ما لا يتناسب بشكل جيد مع البيانات المتعلقة بالأدوات المعقدة المتعددة الأبعاد. في هذه الحالات فإن نمذجة المعادلة البنائية الاستكشافية Exploratory Structural Equation Modeling (ESEM) أصبحت أكثر إنتشاراً، خاصة مع بيانات الشخصية أو المتغيرات المترابطة، والأكثر تعقيدًا (Asparouhov & Muthén, 2009).

القياس متعدد الأبعاد الموزون (WDMS) هو نهج أخر ملائم لتقييم التكافؤ البنائي عبر نسخ اللغة المختلفة لاختبار ما. وعلى غرار التحليل العاملي الاستكشافي، لا يتطلب تحليل القياس المرجح متعدد الأبعاد، تحليل بنية الاختبار مسبقًا، ولكنه يتشابه أيضاً مع التحليل العاملي التوكيدي، في أنه يسمح بتحليل مجموعات متعددة (مثال، Sireci, et al., 2003).

اقترح "فان دي فيفر وتانزر" (van de Vijver & Tanzer, 1997) أن يدرس الباحثون المهتمون بالدراسات عبر الثقافات ثبات كل نسخة من الاختبار موضع الاهتمام عبر المجموعات الثقافية المختلفة، وأن يبحثوا عن أدلة الصدق التقاربي والتمييزي في كل مجموعة ثقافية. قد تكون هذه الدراسات في كثير من الأحيان عملية أكثر من دراسات بنية الاختبار التي تتطلب أحجام عينة كبيرة جدًا.

مع ذلك، يجب الاعتراف بأن مقارنة أداء المتقدم للاختبار عبر نسختين لغويتين من الاختبار ليست دائما هدف ترجمة/نقل الاختبار لثقافة جديدة. الهدف ببساطة هو أن يكون قادرًا

على تقييم المتقدمين للاختبار في مجموعة لغوية مختلفة في تكوين فرضي معين. في هذه الحالة، يعد الفحص الدقيق لصدق الاختبار في المجموعة اللغوية الثانية أمرًا ضروريًا، لكن البحث لإيجاد دليل على تكافؤ الصيغتين ليس بالغ الأهمية. يعتمد أهمية هذا المبدأ الارشادي على الغرض أو الأغراض المقصودة من الاختبار باللغة الثانية (أي مجموعة اللغة المستهدفة)، حيث تتطلب اختبارات مثل تلك المستخدمة في PISA أو TIMSS أدلة على تداخل المحتوى العالي لأن النتائج تستخدم لمقارنة تحصيل الطلاب في العديد من البلدان. إن استخدام قائمة الاكتئاب المترجمة من اللغة الإنجليزية إلى اللغة الصينية للباحثين لدراسة الاكتئاب أو للمرشدين لتقييم اكتئاب عملائهم لا يتطلب تداخلًا كبيرًا في المحتوى. وبدلًا من ذلك، سوف تكون هناك حاجة إلى الصدق لدعم قائمة الاكتئاب في الصين.

يمكن التحقق من هذا المبدأ الارشادي أيضاً بالطرق الإحصائية بعد نقل الاختبار إلى ثقافة أخرى. على سبيل المثال، إذا كان يُعتقد أن المجموعات الثقافية تختلف بشأن متغيرات هامة لا صلة لها بالبناء المقاس، يمكن استخدام تصاميم شاملة وتحليلات إحصائية للسيطرة على هذه المتغيرات "المزعجة". ويمكن استخدام تحليل التباين المصاحب، والتصميمات العاملية، وغيرها من الأساليب الإحصائية (تحليل الانحدار، والارتباط الجزئي، وغيرها) للسيطرة على آثار المصادر غير المرغوب فيها في التباين بين المجموعات.

اقتراحات للتّطبيق. هذا المبدأ الارشادي في غاية الأهمية، وهناك العديد من التحليلات التي يمكن إجراؤها. لتحليلات التكافؤ، نقدم الاقتراحات التالية:

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

- إذا كانت أحجام العينات كافية، قم بإجراء دراسة لمقارنة التكافؤ البنائي للنسخة باللغة الأصلية، والنسخة باللغة المنقول لها الاختبار. هناك الكثير من حزم البرامج لتسهيل هذه التحليلات (انظر، Byrne, 2006).
- إجراء تحليل عاملي استكشافي (يفضل أن يكون بالتدوير إلى البناء المستهدف – ما يسمى بالدوران المستهدف target rotation) أو التحليل العاملي التوكيدي، و/أو تحليل التقييس متعدد الأبعاد الموزون، لتحديد مستوى الاتفاق في بنية الاختبار محل الاهتمام عبر اللغة و/أو المجموعات الثقافية. إن اشتراط أحجام عينات كبيرة (10 أشخاص لكل متغير/بند) يجعل من الصعب إجراء العديد من الدراسات بين الثقافات. يمكن الاطلاع على مثال ممتاز لدراسة من هذا النوع قام بها "بيرن وفان دي فيفر" (Byrne & van de Vijver, 2014).
- ابحث عن أدلة عن الصدق التقاربي والصدق التمييزي (بشكل أساسي، ابحث عن دليل من خلال الارتباطات بين مجموعة من الأبنية المختلفة، وتحقق من استقرار هذه الارتباطات عبر اللغة و/أو المجموعات الثقافية. (انظر، van de Vijver & Tanzer, 1997).

لتحليلات الأداء التفاضلي/التمييزي للبند، نعرض بعض الاقتراحات أدناه. للمهتمين بمناهج أكثر تطورًا، نشجع الباحثين لقراءة الدّراسات السّابقة المختصّة عن الأداء التفاضلي/التمييزي للبند:

- اجراء تحليل للأداء التفاضلي للبند باستخدام أحد الإجراءات المعيارية (إذا كان البند ثنائي التقسيم، فقد يكون إجراء مانتل–هانزل هو الأكثر مباشرة؛ أما إذا كان البند من البنود متعددة الاستجابة، فإن إجراء مانتل–هانزل المعمم يعد من الخيارات المتاحة). تشمل الحلول الأخرى والأكثر تعقيدًا المناهج القائمة على إجراءات نظرية الاستجابة المفردة. إذا كانت أحجام العينة

صغيرة، يمكن أن تكشف "أشكال دلتا delta plot" عن البنود التي يحتمل أن تكون معيبة. تعد المقارنات المشروطة conditional comparisons بديل متاح أخر. ينصح بالاطلاع على المقارنة المفيدة للأساليب الاحصائية التي تعتمد على عينات صغيرة (انظر ، Muñiz, Hambleton, & Xing, 2001).

(11) C-3 تقديم أدلة تدعم المعايير، والثبات، وصدق النسخة المقننة من الاختبار في المجتمع المستهدف.

<u>الشرح:</u> لا تنطبق المعايير وأدلة الصدق والثبات للاختبار في نسخته الأصلية تلقائيًا على النسخ الأخرى للاختبار في الثقافات واللغات المختلفة. يجب أيضًا تقديم أدلة امبريقية/تجريبية على صدق وثبات أي اصدار جديد تم تطويره من الاختبار. وينبغي إدراج جميع أنواع الأدلة التجريبية التي تدعم الاستدلالات المستخلصة من الاختبار في دليل الاختبار. كما ينبغي الاهتمام بشكل خاص بالمصادر الخمسة لأدلة الصدق والتي تستند إلى: محتوى الاختبار، وعمليات الاستجابة، والبناء الداخلي، والعلاقات مع المتغيرات الأخرى، وما يترتب على العملية الاختبارية (AERA, APA, NCME, 2014). إن التحليل العاملي الاستكشافي والتوكيدي، ونمذجة المعادلات البنائية، والتحليلات متعددة السمات والطرق multitrait–multimethod analyses، هي بعض من الأساليب الإحصائية التي يمكن استخدامها للحصول على بيانات وتحليلها لتقديم الأدلة على الصدق وتحليلها استنادا إلى البناء الداخلي للاختبار.

<u>اقتراحات للتّطبيق</u>. لا تختلف الاقتراحات هنا عن الاقتراحات المطلوبة لأي اختبار يتم جمع الأدلة عليه صدقه وثباته قبل استخدامه:

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

- إذا تم اقترح استخدام المعايير التي وضعت في الإصدار الأصلي من الاختبار مع الإصدار المنقول للغة أو الثقافة الجديدة، ينبغي تقديم أدلة مناسبة وعادلة على أن هذا الاستخدام صحيح من الناحية الإحصائية. وإذا لم يكن بالإمكان تقديم أي دليل على استخدام المعايير الأصلية، ينبغي وضع معايير محددة للصيغة المنقول إليها الاختبار وفقاً لأسس اشتقاق المعايير.
- تجميع أدلة ثبات كافية لتبرير الاستخدامات المختلفة للغة أو الثقافة المنقول الاختبار إليها. وقد تتضمن الأدلة عادة تقدير للاتساق الداخلي (على سبيل المثال، صيغة كيودر ريتشاردسون–20، أو معاملات ألفا، أو أوميجا).
- تجميع أدلة الصدق بقدر ما هو مطلوب للتحقق من إمكانية استخدام النسخة الجديدة من الاختبار وفقاً للاستخدام المقصود. يتوقف نوع الأدلة التي يتم تجميعها على الاستخدام المقصود للدرجات (مثل صدق المحتوى للاختبارات التحصيلية، والتحقق من الصدق التنبؤي لاختبارات الاستعداد، وما إلى ذلك).

(12) C–4 استخدم تصميم المعادلة المناسب وإجراءات تحليل البيانات عند ربط مقاييس الدرجات من إصدارات الاختبار باللغات المختلفة.

<u>الشرح:</u> عند ربط نسختين من الاختبار بلغتين مختلفتين في مقياس واحد، فهناك عدة خيارات ممكنة. إذا تم استخدام مجموعة مشتركة من البنود، فيجب تقييم أداء هذه البنود المشتركة عبر مجموعتين لغويتين، وإذا لوحظ الأداء التفاضلي/التمييزي لبعض البنود، فيجب النظر في حذف تلك البنود قبل استخدامها في الربط. وتخدم أشكال دلتا (Angoff & Modu, 1973) هذا الغرض بشكل جيد، وقدم "كوك وشميت – كاسكالار" (Cook & Schmitt–Cascaller, 2005) مثالًا جيدًا عن كيفية استخدام أشكال دلتا لتحديد البنود التي لها معنى مختلف لمجموعتي المفحوصين.

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

ليس لدى كافة أنواع البنود نفس إمكانيات الربط عبر إصدارات اللغة. يمكن رسم تقديرات معاملات التمييز، وصعوبات البنود المستمدة من اطار نظرية الاستجابة للبند للمساعدة في تحديد البنود المشتركة ذات الأداء غير المناسب (انظر، Hambleton, Swaminathan, & Rogers, 1991).

لكن ربط (أي "معادلة") الدرجات عبر نسختين من للاختبار بلغتين مختلفتين سيكون دائمًا مشكلة، نظرًا لضرورة وضع افتراضات قوية حول البيانات. وفي بعض الأحيان، يتم وضع افتراض صعب للغاية بأن النسخ اللغوية المختلفة للاختبار تكون متكافئة، ومن ثم يمكن استخدام الدرجات من نسختي من الاختبار بشكل متبادل. قد يكون لهذا الافتراض ميزة مع اختبارات الرياضيات لأن الترجمة/النقل عادة ما تكون مباشرة. وبالإضافة إلى ذلك، إذا تم إنشاء نسختين من الاختبار بعناية، وبالتالي يمكن افتراض أن نسخة الاختبار باللغة الأصلية تعمل مع المجتمع المخصصة له بنفس الطريقة التي تعمل بها نسخة الاختبار باللغة المستهدفة مع المجتمع المستهدف الجديد. قد يكون لهذا الافتراض ميزة إذا كانت جميع الأدلة الأخرى المتاحة تشير إلى أن النسختين اللغويتين للاختبار متكافئتين، ولا توجد تحيزات في الأسلوب تؤثر على الدرجات في نسخة الاختبار باللغة المستهدفة.

يوجد حلان آخران، ولكن لا يعتبر أي منهما مثالي. الحل الأول أن يعتمد إجراء الربط باستخدام عينة فرعية من البنود التي تعتبر متكافئة بشكل أساسي عبر نسختي الاختبار. على سبيل المثال، قد نختار عينة البنود التي تعد الأسهل من حيث الترجمة أو النقل إلى ثقافة أخرى. من حيث المبدأ، يمكن أن يعمل هذا الحل بكفاءة، ولكنه يتطلب أن تكون بنود الربط وبقية بنود الاختبار تقيس نفس الشيء (البنية). الحل الثاني يعتمد على الربط من خلال عينة من المتقدمين للاختبار الذين يتحدثون لغتين. عند تطبيق كلا النسختين من الاختبار على هذه العينة من

المستجيبين، سيكون من الممكن إنشاء جدول تحويل الدرجات بين النسختين. لاحظ أن هذه العينة من المستجيبين يحب أن لا تكون صغيرة الحجم جدًا، كما أنه عند تصميم الدراسة يجب أن يتم ترتيب عرض نسختي الاختبار بشكل متوازن (البعض يستجيب على النسخة الأصلية أولاً، والبعض الأخر يستجيب على النسخة المنقول لها الاختبار أولاً). والافتراض الأساسي في هذه الطريقة هو أن المستجيبين يتحدثون اللغتين الأصلية والمستهدفة فعلياً بطلاقة، وبالتالي، بصرف النظر عن الصعوبات النسبية لنسختي الاختبار، ينبغي للمستجيبين أن يستجيبوا بنفس الطريقة على كلا النسختين بصرف النظر عن لغة كل نسخة (لا يواجهوا صعوبات أكبر نتيجة لغة نسخة الاختبار). في هذه الحالة، يتم استخدام الفروق التي تظهر في الأداء لضبط الدرجات عند تحويلها من نسخة ما من الاختبار إلى النسخة الأخرى.

اقتراحات للتّطبيق. يعد ربط النتائج عبر نسخ الاختبار المختلفة مشكلة في أغلب الأوقات لأن جميع تصميمات المعادلة لها عيب رئيسي واحد على الأقل. ربما تكون أفضل استراتيجية هي معالجة جميع الخطوات بشكل كامل لإنشاء تكافؤ الدرجات. إذا كان الأدلة التي تتناولها الأسئلة الثلاثة أدناه قوية، فيمكن حينها التعامل مع الدرجات من اصدارين مختلفين من الاختبار بشكل متبادل:

- هل هناك أدلة على أن نفس البناء يتم قياسه في النسخ الأصلية للاختبار، والنسخ الأخرى باللغات المستهدفة؟ هل للبناء المقاس نفس العلاقة مع المتغيرات الخارجية الأخرى في الثقافة الجديدة المنقول لها الاختبار؟
- هل هناك أدلة قوية على أن مصادر التحيز المنهجية قد تم التخلص منها (على سبيل المثال، لا توجد مشكلات في زمن تطبيق الاختبار بين النسختين، والصيغ المستخدمة في

الاختبار مألوفة أيضًا للمستجيبين، ولا يوجد ارتباك حول التعليمات، ولا يوجد تحريف منتظم في إدارة الاختبار من مجموعة إلى أخرى، التعليمات المعيارية، عدم وجود أنماط استجابة خاصة (الاستجابات المتطرفة، الاختلاف الدافعية...)؟

- هل الاختبار خالٍ من البنود التي يحتمل أن تكون متحيزة؟ قد يكون من المفيد جداً هنا، استعراض مخططات قيم p (معامل صعوبة البند) أو قيم دلتا لبنود الاختبار في نسختي الاختبار. يجب دراسة النقاط التي لا تقع على طول خط المعادلة الخطي لتحديد ما إذا كانت البنود المرتبطة مناسبة بشكل متساوٍ في كلتا اللغتين. يوفر تحليل الأداء التفاضلي/التمييزي للبند أدلة أقوى على تكافؤ البنود عبر المجموعات اللغوية والثقافية.
- إذا تمت محاولة ربط الدرجات، فيجب اختيار وتنفيذ تصميم الربط المناسب. يجب تقديم أدلة على صدق التصميم.

المبادئ الإرشادية لتطبيق وإدارة الاختبار

(13) A–1 إعداد مواد إدارة الاختبار وتعليمات تطبيقه للتقليل من أي مشاكل متعلقة بالثقافة واللغة التي قد تنتج عن إجراءات تطبيق الاختبار وأنماط الاستجابة التي يمكن أن تؤثر على صدق الاستنتاجات المستمدة من الدرجات.

الشرح: يجب أن يبدأ تنفيذ المبادئ الإرشادية لتطبيق الاختبار وإدارته من تحليل جميع العوامل التي يمكن أن تهدد صدق درجات الاختبار في سياق ثقافي ولغوي محدد. قد تكون الخبرة في تطبيق أداة ما في سياق أحادي اللغة، أو أحادي الثقافة مفيدة بالفعل في توقع المشكلات التي يمكن توقعها في سياق متعدد اللغات، أو متعدد الثقافات. على سبيل المثال، غالبًا ما يعرف مسؤولو الاختبار المتمرسون جوانب التعليمات التي قد تكون صعبة على المستجيبين. قد تظل

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

هذه الجوانب صعبة بعد الترجمة أو النقل. يمكن أن تجد تطبيقات الأدوات في سياق لغوي أو ثقافي جديد مشكلات لم تكن موجودة من قبل في التطبيقات أحادية الثقافة.

اقتراحات للتطبيق. من المهم مع هذا المبدأ الارشادي توقع العوامل المحتملة التي قد تخلق مشكلات في تطبيق وإدارة الاختبار. فيما يلي بعض تلك العوامل التي يجب دراستها لضمان العدالة في تطبيق إدارة الاختبار:

- وضوح تعليمات الاختبار (بما في ذلك ترجمة تلك التعليمات)، وآلية الإجابة (على سبيل المثال، ورقة الإجابة)، والوقت المسموح به (حيث أن أحد المصادر الشائعة للخطأ هو عدم إتاحة الوقت الكافي للمستجيبين على الاختبار)، ودافعية المستجيبين لإكمال الاختبار، بالإضافة إلى معرفة الغرض من الاختبار وكيف سيتم وضع الدرجات عليه.

(14) A–2 تحديد شروط الاختبار التي يجب اتباعها عن كثب عند تطبيق الاختبار على كل المجتمعات المستهدفة.

الشرح: الهدف من هذا المبدأ الارشادي هو تشجيع مطوري الاختبارات على وضع تعليمات الاختبار والإجراءات ذات الصلة (على سبيل المثال، شروط الاختبار، الحدود الزمنية، إلخ) التي يمكن اتباعها عن كثب عند تطبيق الاختبار على كل المجتمعات المستهدفة. يسعى هذا المبدأ الارشادي في المقام الأول إلى تشجيع مسؤولي الاختبار على الالتزام بالإرشادات المعيارية. في الوقت نفسه، قد يتم تحديد الإجراءات اللازمة للتعامل مع المجموعات الفرعية الخاصة من الأفراد ممن قد يحتاجون إلى بعض الاستثناءات مثل الوقت الإضافي، والطباعة بخطوط أكبر حجماً، وأيضا ظروف تطبيق وإدارة الاختبار بسلاسة، إلخ. في مجال الاختبارات اليوم، تُعرف هذه الإجراءات باسم "تسهيلات الاختبار test accommodations". الهدف من هذه التسهيلات ليس

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

تضخيم درجات المتقدمين للاختبار، بل إنشاء بيئة اختبار للمستجيبين مناسبة تمكنهم من إظهار ما قد يشعرون به، أو ما يعرفون، أو ما يمكنهم فعله.

يجب ملاحظة اختلافات شروط تطبيق وإدارة الاختبار عن الشروط المعيارية المقننة لتطبيق الاختبار، بحيث يمكن لاحقًا مراعاة هذه الاختلافات وتأثيرها على التعميمات والتفسيرات.

اقتراحات للتّطبيق. قد يتداخل هذا المبدأ الارشادي جزئيًا مع المبدأ A–1 (13)، ولكن نعيد التذكير هنا بأهمية إدارة وتطبيق الاختبار في جميع المجموعات في ظل ظروف مماثلة قدر الإمكان. هذا الإجراء ضروري إذا كانت الدرجات المستخرجة من نسختين لغويتين مختلفتين للاختبار ستستخدم بالتبادل. نورد هنا بعض الاقتراحات:

- يجب نقل تعليمات الاختبار والإجراءات ذات الصلة وإعادة كتابتها بطريقة موحدة ومعيارية، بحيث تتناسب مع اللغة والثقافة الجديدة.
- إذا تم تغيير تعليمات الاختبار والإجراءات ذات الصلة في الثقافة الجديدة المنقول إليها الاختبار، فيجب تدريب المسؤولين عن إدارة وتطبيق الاختبار على الإجراءات الجديدة؛ ويجب إبلاغهم بالالتزام بهذه الإجراءات الجديدة، وليس الإجراءات الأصلية.

المبادئ الإرشادية لرصد الدرجات وتفسيرها

SSI–1 (15) تفسير أي اختلافات في درجات المجموعة بالإشارة إلى جميع المعلومات المتاحة ذات الصلة.

الشرح: حتى إذا تم نقل الاختبار إلى لغة/ثقافة جديدة باستخدام إجراءات سليمة من الناحية الفنية، وتم بالفعل التحقق من صدق درجات الاختبار إلى حد ما، يجب أن يوضع في الاعتبار أن

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

معنى الاختلافات بين المجموعات يمكن تفسيره بعدة طرق بسبب اختلاف الثقافة أو الاختلافات الأخرى عبر البلدان و/أو الثقافات المشاركة.

استعرض Sireci (2005) منهج تقييم تكافؤ اصداريين لغويين مختلفتين للاختبار من خلال إدارة الإصدارات اللغوية المنفصلة للاختبار لمجموعة من المستجيبين على الاختبار الذين يتقنون لغتين (ثنائي اللغة) والذين ينحدرون من نفس المجموعة الثقافية أو اللغوية. وقد حَددَ Sireci بعض خيارات تصميم البحوث لدراسات التكافؤ باستخدام مستجيبين ثنائي اللغة، وأَدرجَ المتغيرات المربكة المحتملة التي تحتاج إلى التحكم فيها، وقدمَ بعض الاقتراحات القيمة لتفسير النتائج.

<u>اقتراحات للتطبيق</u>. فيما يلي نورد اقتراح واحد لتحسين تنفيذ هذا المبدأ الإرشادي:

- اعتماداً على السؤال البحثي (أو السياق الذي يتم من خلاله إجراء المقارنات بين المجموعات)، يمكن النظر في عدد من التفسيرات المحتملة، قبل تحديد أحدها في النهاية. على سبيل المثال: من المهم استبعاد الدوافع التفاضلية differential motivation للأداء الجيد في الاختبار قبل استنتاج أن أداء مجموعة واحدة في الاختبار أفضل من أداء مجموعة أخرى. قد تكون هناك أيضاً تأثيرات للسياق، والتي كان لها مردود أيضاً على الأداء في الاختبار بشكل كبير. على سبيل المثال: قد تكون مجموعة واحدة من مجموعات المقارنة جزءًا من نظام تعليمي أقل فعالية أو كفاءة؛ الأمر الذي قد يكون له تأثير كبير على الأداء في الاختبار.

(16) SSI-2 مقارنة الدرجات عبر المجتمعات فقط بعد تحديد مستوى التكافؤ على المقياس الذي يتم تسجيل الدرجات فيه.

<u>الشرح:</u> عندما تكون الدراسات المقارنة عبر اللغات والمجموعات الثقافية هي المحور الرئيسي لعملية الترجمة والنقل للاختبار، فيجب وضع الإصدارات متعددة اللغات للاختبار على

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

مقياس مشترك، ويتم تنفيذ ذلك من خلال عملية تسمى "الربط" أو "المعادلة". وهذا يتطلب أحجام عينات كبيرة، ودليل على عدم وجود تحيز في البناء الفرضي المقاس، والمنهجية، والبند في الإصدار الجديد المنقول له الاختبار.

حدد (Van de vivjer & Poortinga, 2005) عدة مستويات من تكافؤ الاختبار عبر اللغة والمجموعات الثقافية؛ ويعد عملهما مفيداً بشكل خاص في فهم هذا المفهوم. وفي الواقع، تم تقديم المفهوم الأصلي من قبل هؤلاء المؤلفين، على سبيل المثال: هم من أشاروا إلى أن تكافؤ وحدة القياس يتطلب أن يكون لتدريج المقاييس في كل مجموعة نفس النظام، وبالتالي يمكن ضمان أن تكون للاختلافات بين الأشخاص داخل المجموعات نفس المعنى. (على سبيل المثال، يمكن مقارنة الاختلافات بين الذكور والإناث في عينة الصينية بعينة فرنسية). ومع ذلك، لا يمكن إجراء إجراء مقارنة صحيحة بين الدرجات مباشرة؛ إلا عندما تُظهر الدرجات أعلى مستوى من التكافؤ، والذي يسمى التكافؤ القياسي scalar equivalence أو معادلة الكاملة لدرجات المقياس، الأمر الذي يتطلب أن يكون للمقاييس في كل مجموعة نفس وحدة القياس ونفس نقطة الأصل عبر المجموعات.

تم طرح العديد من الطرق (في إطار النظرية الاختبارية الكلاسيكية ونظرية الاستجابة للبند) لربط أو معادلة الدرجات من مجموعتين (أو الإصدارات اللغوية للاختبار). يمكن للقراء المهتمين الرجوع إلى (Angoff, 1984; Kolen & Brennan, 2004) لاكتساب فهم أعمق لهذا الموضوع. يقترح (Cook & Schmitt-Cascallar, 2005) أساسًا لفهم الأساليب الإحصائية المتوفرة حاليًا لمعادلة وتقييس scaling الاختبارات التربوية والنفسية. يصف المؤلفون وينتقدون إجراءات ربط المقياس المحددة المستخدمة في دراسات نقل الاختبارات ثقافياً، ويوضحون

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

إجراءات وقضايا ربط مختارة من خلال وصف وانتقاد ثلاث دراسات تم إجراؤها على مدار العشرين عامًا الماضية لربط الدرجات من اختبار التقييم الدراسي.

اقتراحات للتطبيق. النقطة الأساسية هنا هي أنه لا ينبغي المبالغة في تفسير درجات الاختبار:

- فسر النتائج في ضوء مستوى أدلة الصدق المتاحة. على سبيل المثال، لا تتوصل إلى استنتاجات تتعلق بالمقارنات بين مستويات أداء المستجيب في مجموعتين لغويتين مختلفتين، ما لم يتم التحقق من تكافؤ القياس لدرجات الاختبار التي تتم مقارنتها في تلك المجموعتين اللغويتين.

القائمة الإرشادية للتقرير والتوثيق

(17) Doc-1 تقديم تقريراً فنيًا لأي تغييرات، بما في ذلك سرد للأدلة التي تم الحصول عليها لدعم التكافؤ، عندما يتم نقل الاختبار للاستخدام في مجتمع آخر.

الشرح: تم إدراك أهمية هذا المبدأ الإرشادي وأكده العديد من الباحثين (انظر، على سبيل المثال، Grisay, 2003). نجح كل من TIMSS وPISA في تنفيذ هذا المبدأ الإرشادي من خلال رصد عملية نقل الاختبار للغة/الثقافة الجديدة وتقريرها بعناية. باستخدام هذه المعلومات، يمكن التركيز على مدى ملاءمة هذه التغييرات التي تم إجراؤها في الاختبار ليناسب اللغة أو الثقافة الجديدة.

يجب أن تحتوي الوثائق الفنية أيضًا على تفاصيل كافية عن المنهجية المستخدمة في النقل، مما يتيح للباحثين تكرار الإجراءات المستخدمة على نفس المجتمعات أو غيرها مستقبلاً. يجب أن تحتوي هذه الوثائق على معلومات كافية من الأدلة على التكافؤ البنائي تكافؤ درجات الاختبار (إذا تم تنفيذه)، لدعم استخدام الأداة في المجتمع الجديد. عند إجراء مقارنات بين

المجتمعات، يجب أن توثق التقارير الفنية والعلمية الأدلة المستخدمة لتحديد معادلة الدرجات بين المجتمعات موضع المقارنة.

في بعض الأحيان، يطرح السؤال حول الجمهور المستهدف للتقرير الفني لعملية نقل الاختبار. يجب كتابة الوثائق للخبير الفني وكذلك للأشخاص الذين سيُطلب منهم تقييم فائدة الاختبار للاستخدام في المجتمعات الجديدة أو غيرها. (يمكن إضافة تقرير تكميلي موجز يُكتب لغير الخبراء الفنيين).

اقتراحات للتطبيق. يجب أن تحتوي وثائق الاختبارات المنقولة إلى لغة/ثقافة جديدة على دليل فني يوثق جميع الأدلة النوعية والكمية المرتبطة بعملية النقل. من المفيد بشكل خاص تقرير أي تغييرات تم إجراؤها لكي يناسب الاختبار اللغة وثقافة الجديدة. بشكل أساسي، سيحتاج الأشخاص الفنيون ومحررو المجلات إلى وثائق حول العملية التي تمت لإنتاج والتحقق من صدق الإصدار اللغوي الجديد للاختبار. وبالطبع أيضًا، سيرغبون في رؤية نتائج جميع التحليلات. فيما يلي بعض أنواع الأسئلة التي يجب الإجابة عليها في هذا السياق:

- ما هي الأدلة المتاحة لدعم فائدة البناء الفرضي موضع القياس والاختبار المنقول للمجتمع الجديد؟
- ما هي البيانات التي جمعت على مستوى البنود ومن أي عينات؟
- ما هي البيانات الأخرى التي تم الحصول عليها لتقييم صدق المحتوى، والصدق المحكي، وصدق البناء (صدق التكوين الفرضي)؟
- كيف تم تحليل مجموعات البيانات المختلفة؟
- ما هي النتائج؟

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

(18) Doc-2 توفير الوثائق لمستخدمي الاختبار التي ستدعم الممارسة الصحيحة لاستخدام اختبار منقول مع الأشخاص في سياق المجتمع الجديد.

الشرح: يجب كتابة الوثائق للأشخاص الذين سيستخدمون الاختبار في مواقف التقييم العملية. يجب أن يكون متسقًا مع الممارسات الجيدة المحددة في الدليل للجنة الاختبار الدولية بشأن استخدام الاختبار Test Use، انظر (www.InTestCom.org)

اقتراحات للتطبيق. يجب على مطور الاختبار تقديم معلومات محددة حول الطرق التي قد تؤثر بها السياقات الاجتماعية والثقافية والبيئية للمجتمع على الأداء في الاختبار. يجب أن يقوم دليل المستخدم بما يلي:

- بوصف الأبنية التي يتم قياسها بواسطة الاختبار، وبتلخيص المعلومات، وبوصِف عملية نقل الاختبار للغة/الثقافة الجديدة.
- بتلخيص الأدلة الداعمة لعملية النقل للغة/الثقافة الجديدة، بما في ذلك الدليل على الملاءمة الثقافية لمحتوى البند، وتعليمات الاختبار، وطريقة الاستجابة، وغيرها.
- بتحديد مدى ملاءمة استخدام الاختبار مع مجموعات فرعية مختلفة ضمن المجتمع وأي قيود أخرى على الاستخدام.
- بشرح أي قضايا يجب أن يتم أخذها في الاعتبار فيما يتعلق بالممارسة الجيدة في تطبيق وإدارة الاختبار.
- بشرح ما إذا كان من الممكن إجراء المقارنات بين المجتمعات، وكيف يمكن ذلك.

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

- بتوفير المعلومات اللازمة لتسجيل الدرجات والمعايير (على سبيل المثال، جداول المعايير ذات الصلة)، أو بوصف كيف يمكن للمستخدمين الوصول إلى إجراءات الحصول على الدرجات (مثال، عندما تكون الاختبار مدار بواسطة الحاسوب).
- بتوفير الإرشادات اللازمة لتفسير النتائج، بما في ذلك المعلومات عن الآثار المترتبة المتعلقة ببيانات الصدق والثبات على الاستنتاجات التي يمكن استخلاصها من درجات الاختبار.

كلمة ختاميّة:

لقد بذلنا قصارى جهدنا لتقديم مجموعة من الإرشادات لمساعدة مطوّري الاختبارات ومستخدميهم في عملهم. ومع ذلك، لكي يكون هناك تأثير فعلي للمبادئ الإرشادية والجهود الأخرى في تحسين الممارسات الضّعيفة؛ يجب أن تكون هناك آليّات نشر جيدّة وأن تكون في مكانها الصّحيح. أظهرت دراسة ممنهجة حديثة أجراها "ريوس وسيرسي" (Rios & Sireci, 2014) أنّ غالبيّة مشاريع نقل الاختبارات لثقافة جديدة في الأدبيات المنشورة لم تتبع المبادئ الإرشادية التي حددتها لجنة الاختبارات الدولية ITC، والتي كانت متاحة منذ عشرين سنة قبل نشر هذا البحث. ولذلك فإننا نشجّع الباحثين على بذل كل الجهود الممكنة لزيادة الوعي بين زملائهم في الإصدار الثّاني هذا، كمصدر أساسي لأفضل الممارسات التّي ساهم فيها العديد من المتخصصين حول العالم.

في الوقت نفسه، نعلم أنّه مثلما تم الآن استبدال الإصدار الأول من هذه الإرشادات بالإصدار الثاني الحالي؛ فإنه سوف يتم استبدال إرشادات الإصدار الثّاني بطبيعة الحال بإصدارات أحدث يوماً ما. إن معايير الاختبارات ذائعة الصيت والتي حددتها AERA وAPA وNCME ، وصلت الآن لإصدارها السّادس (AERA, APA, & NCME, 2004). لذلك نتوقّع أن تخضع إرشادات لجنة الاختبارات الدّوليّة الخاصة بنقل الاختبار إلى ثقافة جديدة لمراجعة أخرى أيضا في السّنوات القادمة. إذا كنت تعرف دراسات جديدة يجب الاستشهاد بها، أو قد تؤثّر على الإصدار الثّالث، أو إذا كنت ترغب في تقديم إرشادات أو مراجعات جديدة للإرشادات الـ18 المعروضة هنا، فيرجى إخبار لجنة الاختبارات الدّوليّة. يمكنك الاتصال بالرئيس الحالي للجنة البحوث والمبادئ

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

الإرشادية، والتّي أنتجت هذا الإصدار الثّاني، أو الاتصال بأمين لجنة الاختبارات الدّوليّة على عنوان البريد الإلكتروني الموجود على موقعنا على الانترنت: www.InTestCom.org.

المراجع

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement, 36*(3), 185-198.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service.

Angoff, W. H., & Modu, C. C. (1973). Equating the scales of the Prueba de Apititud Academica and the Scholastic Aptitude Test (Research Rep No. 3). New York: College Entrance Examination Board.

Asparouhov, T. & Muthén, B. (2009). Exploratory structural modeling. *Structural Equation Modeling, 16*, 397-438.

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 Publications.

Byrne, B. (2001). Structural equation modeling with AMOS, EQS, and LISREL: Comparative approaches to testing for the factorial validity of a measuring instrument. *International Journal of Testing, 1*, 55-86.

Byrne, B. (2003). Measuring self-concept measurement across culture: Issues, caveats, and application. In H. W. Marsh, R. Craven, & D. M. McInerney (Eds.), *International advances in self research*. Greenwich, CT: Information Age Publishing.

Byrne, B. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Publishers.

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema*, *20*, 872-882.

Byrne, B. M., & van de Vijver, F.J.R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing, 10,* 107-132.

Byrne, B. M., & van de Vijver, F.J.R. (2014). Factorial structure of the Family Values Scale from a multilevel-multicultural perspective. *International Journal of Testing, 14*, 168-192.

Clauser, B. E., Nungester, R. J., Mazor, K., & Ripley, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. *Journal of Educational Measurement, 33*(2), 202-214.

Cook, L. L., & Schmitt-Cascallar, A. P. (2005). Establishing score comparability for tests given in different languages. In R. K. Hambleton, P. F. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 139-170).

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and Practice* (pp. 137-166).

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

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. (1998). Translation effects in international assessments. *International Journal of Educational Research, 29*(6), 543-533.

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

Ercikan, K., Gierl, J. J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education, 17*(3), 301-321.

Ercikan, K., Simon, M., & Oliveri, M. E. (2013). Score comparability of multiple language versions of assessments within jurisdictions. In M. Simon, K. Ercikan, & M. Rousseau (Eds.), *An international handbook for large-scale assessments* (pp. 110-124). New York:

Grégoire, J., & Hambleton, R. K. (Eds.). (2009). Advances in test adaptation research [Special Issue]. *International Journal of Testing, 9*(2), 73-166.

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

Hambleton, R. K. (2002). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment, 17*(3), 164-172.

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3-38). Mahwah, NJ: Lawrence Erlbaum Publishers.

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., & 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., 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., & Lee, M. (2013). Methods of translating and adapting tests to increase cross-language validity. In D. Saklofske, C. Reynolds, & V. Schwean (Eds.), *The Oxford handbook of child assessment* (pp. 172-181). New York: Oxford University Press.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. (Eds.). (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum Publishers.

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

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

Hambleton, R. K., & Zenisky, A. (2010). Translating and adapting tests for cross-cultural assessment. In D. Matsumoto & F. van de Vijver (Eds.), *Cross-cultural research methods* (pp. 46-74). New York, NY; Cambridge University Press.

Harkness, J. (Ed.). (1998). *Cross-cultural survey equivalence*.

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

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*, 249-260.

Javaras, K. N., & Ripley, B. D. (2007). An 'unfolding' latent variable model for Likert attitude data: Drawing inferences adjusted for response style. *Journal of the American Statistical Association, 102*, 454-463.

Jeanrie, C., & Bertrand, R. (1999). Translating tests with the International Test Commission Guidelines: Keeping validity in mind. *European Journal of Psychological Assessment, 15*(3), 277-283.

Johnson, T. R. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika, 68*, 563-583.

Kolen, M. J., & Brennan, R. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Levin, K., Willis, G. B., Forsyth, B. H., Norberg, A., Stapleton Kudela, M., Stark, D., & Thompson, F. E. (2009). Using cognitive interviews to evaluate the Spanish-language translation of a dietary questionnaire. *Survey Research Methods, 3*(1), 13-25.

Li, Y., Cohen, A. S., & Ibarra, R. A. (2004). Characteristics of mathematics items associated with gender DIF. *International Journal of Testing, 4*(2), 115-135.

Mazor, K.H., Clauser, B.E., & Hambleton, R.K. (1992). The effect of simple size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52*(2), 443-451.

Muniz, J., Elosua, P., & Hambleton, R. K. (2013). Directrices para la traduccion y adaptacion de los tests: segunda edicion. *Psicothema, 25(2),* 149-155.

Muñiz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing, 1*(2), 115-135.

Oort, F. J., & Berberoğlu, G. (1992). Using restricted factor analysis with binary data for item bias detection and item analysis. In T. J. Plomp, J. M. Pieters, & A. Feteris (Eds.), *European Conference on Educational Research: Book of Summaries* (pp. 708-710). Twente, the Netherlands: University of Twente, Department of Education.

Park, H., Pearson, P. D., & Reckase, M. D. (2005). Assessing the effect of cohort, gender, and race on DIF in an adaptive test designed for multi-age groups. *Reading Psychology, 26*, 81-101.

Rios, J., & Sireci, S. (2014). Guidelines versus practices in cross-lingual assessment: A disconcerting disconnect. *International Journal of Testing, 14*(4), 289-312.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*(2), 105-116.

Rotter, J.B. & Rafferty, J.E. (1950). *Manual: The Rotter Incomplete Sentences Blank: College Form*. New York: Psychological Corporation.

Scheuneman, J. D., & Grima, A. (1997). Characteristics of quantitative word items associated with differential performance for female and Black examinees. *Applied Measurement in Education, 10*(4), 299-319.

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

Sireci, S. G. (2005). Using bilinguals to evaluate the comparability of different language versions of a test. In R. K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 117-138). Mahwah, NJ: Lawrence Erlbaum Publishers.

Sireci, S. G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing, 20*(2), 148-166.

Sireci, S. G., & Berberoğlu, G. (2000). Using bilingual respondents to evaluate translated-adapted items. *Applied Measurement in Education, 13*(3), 229-248.

Sireci, S. G., Patsula, L., & Hambleton, R. K. (2005). Statistical methods for identifying flaws in the test adaptation process. In R. K. Hambleton, P. Merenda, & C. Spielberger, C. (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93-116). Mahwah, NJ: Lawrence Erlbaum Publishers.

Sireci, S. G., Harter, J., Yang, Y., & Bhola, D. (2003). Evaluating the equivalence of an employee attitude survey across languages, cultures, and administration formats. *International Journal of Testing, 3*(2), 129-150.

Sireci, S. G., & Wells. C. S. (2010). Evaluating the comparability of English and Spanish video accommodations for English language learners. In P. Winter (Ed.), *Evaluating the comparability of scores from achievement test variations* (pp. 33-68). Washington, DC: Council of Chief State School Officers.

Solano-Flores, G., Trumbull, E., & Nelson-Barber, S. (2002). Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities. *International Journal of Testing, 2*(2), 107-129.

Subok, L. (2017). Detecting differential item functioning using the logistic regression procedure in small samples. *Applied Psychological Measurement, 41(1)*, 30-43.

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

Tanzer, N. K., & Sim, C. O. E. (1999). Adapting instruments for use in multiple languages and cultures: A review of the ITC Guidelines for Test Adaptation. *European Journal of Psychological Assessment, 15*, 258-269.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147-169). Mahwah, NJ: Lawrence Erlbaum Publishers.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 67-113). Mahwah, NJ: Lawrence Erlbaum Publishers.

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 Publications.

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). Dordrecht, the Netherlands: Kluwer Academic Publishers.

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., & Poortinga, Y. H. (2005). Conceptual and methodical issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. Spielberger (Eds.), *Adapting educational and*

psychological tests for cross-cultural assessment (pp. 39-64). Mahwah, NJ: Lawrence Erlbaum Publishers.

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*(4), 263-279.

Wolf, E.J., Harrington, K.M., Clark, S.L., & Miller, M.W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement, 73(6),* 913–934.

الملحق أ: قائمة مراجعة المبادئ الإرشادية للجنة الاختبارات الدولية لترجمة ونقل الاختبارات لثقافة جديدة

فيما يلي قائمة مرجعية لتذكيرك بالمبادئ الإرشادية (عددها 18 مبدأ إرشادي) للجنة الاختبارات الدولية. نوصيك بوضع علامة عند تلك المبادئ التي تشعر أنك تعاملت معها بشكل مُرضٍ في مشروعك لترجمة أو نقل الاختبار الخاص بك لثقافة جديدة، ومن ثم الاهتمام بتلك التي لم يتم التعامل معها بشكل مرضي.

القائمة الإرشادية للشروط المسبقة

(1)PC–1 الحصول على الإذْن اللازم من صاحب الحقوق الملكيّة الفكريّة المتعلّقة بالاختبار قبل إجراء أي نقل للاختبار للثقافة الجديدة. [ ]

(2) PC–2 القيام بالتأكد من أن مقدار التّداخل في التّعريف ومحتوى التكوين الفرضي construct المقاس بواسطة الاختبار، ومحتوى البنود في المجتمعات التي سينقل منها وإليها الاختبار، كافية وفقاً للغرض المقصود من استخدام درجات الاختبار. [ ]

(3) PC–3 التقليل من تأثير أي اختلافات ثقافية ولغويّة لا علاقة لها بالاستخدامات المقصودة للاختبار في المجتمع المنقول له. [ ]

القائمة الإرشادية لتطوير الاختبار

(4) TD–1 التأكد من أن عمليات الترجمة ونقل الاختبار لثقافة جديدة تأخذ في الاعتبار الاختلافات اللغوية والنفسية والثقافية في المجتمعات المستهدفة من خلال اختيار الخبراء ذوي الخبرة المناسبة. [ ]

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

(5) TD–2 استخدام تصاميم وإجراءات الترجمة المناسبة لتحقيق أقصى قدر من ملاءمة عملية نقل الاختبار للثقافة جديدة في المجتمعات المستهدفة. [ ]

(6) TD–3 تقديم أدلة على أن تعليمات الاختبار ومحتوى البنود لها معنى مماثل لجميع المجتمعات المستهدفة. [ ]

(7) TD–4 تقديم أدلة على أن صيغ البنود، ومقاييس التقدير، وفئات التسجيل وتعليمات الاختبار، وطرق التطبيق، وغيرها من الإجراءات مناسبة لجميع المجتمعات المستهدفة. [ ]

(8) TD–5 جمع البيانات الاستطلاعية من النسخة الاختبارية الجديدة حتى يمكن تحليل البنود، وتقييم الثبات ودراسة الصدق (على نطاق صغير) بحيث يمكن إجراء أية مراجعات ضرورية للنسخة الاختبارية الجديدة. [ ]

القائمة الإرشادية للتحقق

(9) C–1 اختيار عينة ذات خصائص متعلقة بالاستخدام المقصود للاختبار وبحجم كافي للتحليلات الأمبريقية التي ستقوم بها. [ ]

(10) C–2 تقديم أدلة إحصائية ذات صلة حول التكافؤ البنائي، والتكافؤ المنهجي، والتكافؤ على مستوى البند لجميع المجتمعات المستهدفة. [ ]

(11) C–3 تقديم أدلة تدعم المعايير، والثبات، وصدق النسخة المقننة من الاختبار في المجتمع المستهدف. [ ]

(12) C–4) استخدم تصميم المعادلة المناسب وإجراءات تحليل البيانات عند ربط مقاييس الدرجات من إصدارات الاختبار باللغات المختلفة. [ ]

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

القائمة الإرشادية لتطبيق وإدارة الاختبار

(13) A–1 إعداد مواد إدارة الاختبار وتعليمات تطبيقه للتقليل من أي مشاكل متعلقة بالثقافة واللغة التي قد تنتج عن إجراءات تطبيق الاختبار وأنماط الاستجابة التي يمكن أن تؤثر على صدق الاستنتاجات المستمدة من الدرجات. [ ]

(14) A–2 تحديد شروط الاختبار التي يجب اتباعها عن كثب عند تطبيق الاختبار على كل المجتمعات المستهدفة. [ ]

القائمة الإرشادية لرصد الدرجات وتفسيرها

(15) SSI–1 تفسير أي اختلافات في درجات المجموعة بالإشارة إلى جميع المعلومات المتاحة ذات الصلة. [ ]

(16) SSI–2 مقارنة الدرجات عبر المجتمعات فقط بعد تحديد مستوى التكافؤ على المقياس الذي يتم تسجيل الدرجات فيه. [ ]

القائمة الإرشادية للتقرير والتوثيق

(17) Doc–1 تقديم تقريراً فنيًا لأي تغييرات، بما في ذلك سرد للأدلة التي تم الحصول عليها لدعم التكافؤ، عندما يتم نقل الاختبار للاستخدام في مجتمع آخر. [ ]

(18) Doc–2 توفير الوثائق لمستخدمي الاختبار التي ستدعم الممارسة الصحيحة لاستخدام اختبار منقول مع الأشخاص في سياق المجتمع الجديد. [ ]

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

الملحق ب. قائمة المصطلحات

ألفا Alpha (أو تسمى أحيانًا "معامل ألفا" أو "ألفا كرونباك"): أحد معاملات ثبات الاختبار الذي يُفترض أن بنوده تقيس سمة واحدة مشتركة ولديها معاملات تمييز متساوية (لذلك فهي حالة خاصة من أوميجا – انظر أدناه)؛ يمكن اعتباره الحد الأدنى للثبات.

تصميم الترجمة العكسية Backward Translation Design: في هذا التصميم، تتم ترجمة الاختبار من إصدار بلغة ما (تسمى اللغة الأصلية للاختبار) إلى إصدار أخر بلغة أخرى (تسمي اللغة المستهدفة) بواسطة مترجم واحد أو مجموعة واحدة من المترجمين، ثم تتم إعادة ترجمة إصدار الاختبار باللغة المستهدفة إلى اللغة الاصلية مرة أخرى؛ بواسطة مترجم آخر مستقل أو مجموعة أخرى من المترجمين. تتم مقارنة الاختبار في لإصداره الأصلي بالاختبار في المصدر المعاد ترجمته للغة الأصلية، ومن ثم إصدار حكم حول مدى ملاءمة إصدار عن مدى ملاءمة الإصدار الأصلي من الاختبار. إذا كان الإصداران اللذان يتم مقارنتهما قريبين من بعض بدرجة كافية، يتم قبول إصدار الاختبار باللغة المستهدفة.

التحليل العاملي التوكيدي Confirmatory Factor Analysis (CFA): بعد البدء بافتراض بنية محددة للاختبار موضع الاهتمام، يتم إجراء تحليل للحصول على بنية الاختبار من مصفوفة معاملات الارتباط بين البنود المكونة للاختبار. يتم إجراء اختبار إحصائي لمدى التقارب بين بنية الاختبار المفترضة والبنية التي تم الحصول عليها من مصفوفة معاملات الارتباط. يقوم الاختبار الاحصائي بالتحقق من صحة الفرض الصفري الذي ينص على تساوي البنيتين موضع المقارنة. إذا كانت البنيتين قريبتين بدرجة كافية وفقاً للاختبار الاحصائي، فإنه لا يمكننا رفض الفرض الصفري، وبالتالي التحقق من البناء العاملي للاختبار.

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

قيم دلتا Delta Values: قيم دلتا هي ببساطة تحويل خطي لقيم *p* (معامل الصعوبة للبند)، ويتم تطبيقها على البنود ذات تقسيم الدرجات الثنائي binary-scored items. قيمة دلتا لبند ما هي الانحراف الطبيعي المقابل للمنطقة الواقعة تحت التوزيع الطبيعي (التوزيع الطبيعي له متوسط = 0، وانحراف معياري = 1)، حيث تكون المنطقة الواقعة تحت التوزيع الطبيعي مساوية لنسبة الأفراد الذين يجيبون على البند بشكل صحيح. لذلك، إذا كانت 84. = *p*، فإن قيمة دلتا للبند ستكون (−1). يتم إجراء هذا التحويل وفقاً للافتراض بأن قيم دلتا من المرجح أن تكون على مقياس فترى بشكل أفضل من قيم *p*.

الأداء التفاضلي/التمييزي للبند Differential Item Functioning (DIF): هناك فئة من الإجراءات الإحصائية يمكنها تحديد ما إذا كان البند يعمل بشكل متماثل إلى حد ما في مجموعتين مختلفتين أم لا. يتم إجراء مقارنات الأداء عن طريق مطابقة المستجيبين أولاً على السمة التي يتم قياسها بواسطة الاختبار. عندما يتم ملاحظة فروق في الأداء، يُقال أنه من المحتمل أن يكون البند متحيزًا. يحاول الباحث بعد التوصل لبند من المحتمل تحيزه؛ تفسير سبب الاختلافات في الأداء المستجيبين على هذا البند في المجموعتين المتطابقين من حيث السمة المقاسة بواسطة هذا البند.

الترجمة المزدوجة والتوافق Double-Translation and Reconciliation: أحد تصميمات الترجمة، حيث يقوم مترجم مستقل أو لجنة خبراء بتحديد وحل أي تناقضات قد تظهر بين عدد من الترجمات البديلة لنفس البنود، والوصول لإصدار واحد يوفق بينها.

الممتحنون/المستجيبون Examinees: تُستخدم بالتبادل في مجال الاختبارات مع "المتقدمين للاختبار"، و"المرشحين للاستجابة على اختبار ما"، و"المستجيبين"، و"الطلاب" (يستخدم عادة مصطلح الطلاب في حالة الاختبارات التحصيلية).

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

التحليل العاملي الاستكشافي Exploratory Factor Analysis (EFA): هو عبارة عن إجراء إحصائي يتم تطبيقه، على سبيل المثال مع مصفوفة معاملات الارتباط التّي تنتجها العلاقات المتبادلة بين مجموعة من البنود في اختبار ما (أو مجموعة من الاختبارات). الهدف من هذا الإجراء الاحصائي هو محاولة تفسير العلاقات المتبادلة بين بنود الاختبار (أو الاختبارات) من خلال الوصول لعدد صغير من العوامل التّي يُعتقد أنّ الاختبار (أو الاختبارات) يقيسها. على سبيل المثال مع اختبار للرّياضيّات، قد يحدّد التحليل العاملي حقيقّة أن البنود تقع في ثلاث مجموعات هي: الحساب، والمفاهيم، وحل المسائل. يمكننا القول في هذه الحالة أن اختبار الرياضيات يقيس ثلاثة عوامل: الحساب، ومفاهيم الرّياضيات، وحل المسائل الحسابيّة.

تصميم التّرجمة الأمامية Forward Translation Design: أحد تصميمات الترجمة، حيث يتم نقل الاختبار للغة ما مستهدفة من قِبل مترجم، أو في كثير من الأحيان، مجموعة من المترجمين. بعد ذلك يقوم مترجم مستقل أو مجموعة من المترجمين المستقلين بالحكم على تكافؤ إصدارات الاختبار بلغة الأصلية واللغة المستهدفة.

نظرية استجابة على البند Item Response Theory (IRT): فئة من النماذج الاحصائية لربط الاستجابات على البنود بسمة أو مجموعة من السّمات التّي يتم قياسها بواسطة بنود الاختبار الذي ينتمي إليه البند. يمكن لبعض نماذج IRT التّعامل مع بيانات كل من الاستجابة الثّنائية أو المتعدّدة. قد تأتي بيانات ثّنائية الاستجابة من الإجابة على بنود الاختيار من متعدد أو الصواب والخطأ في أحد مقاييس الشّخصيّة، وقد تأتي بيانات متعدّدة الاستجابة من الإجابة على مهام الأداء أو الأسئلة المقالية في أحد الاختبارات التحصيلية، أو من مقاييس التقدير والتي من أبرز أمثلتها مقياس "ليكرت Likert".

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

معادلة كيودر–ريتشاردسون 20 (أو أحيانًا تسمّى ببساطة KR–20): معادلة لتقدير معامل ثبات الاختبار الذي يتكون من بنود ثنائية الاستجابة، والذّي يفترض أنّها تقيس سمة واحدة مشتركة ولديها معاملات تمييز متساوية.

التّكييف المحلّي Localization: وهذا المصطلح الشائع في مجال الاختبار يُستخدم لوصف العمليّة التي يتم من خلالها جعل اختبار ما صُمم بلغة معينة لثقافة محددة، مقبولاً وصالحاً للاستخدام في ثقافة أخرى بلغة أخرى. المصطلح المكافئ هو ترجمة أو نقل الاختبار لثقافة جديدة.

تحليل الانحدار اللوجستي لتحديد أداء البند التّفاضلي Logistic Regression Procedure for Identifying Differential Item Functioning: هو إجراء إحصائي هو أحد الطرق البديلة لإجراء تحليلات الأداء التفاضلي للبند DIF. يتم فيه التحقق من ملاءمة المنحنى اللوجستي مع بيانات الأداء لكل مجموعة؛ ومن ثم تتم المقارنة الإحصائية بين المنحنيين اللوجستيين (منحنى لكل مجموعة لغوية).

إجراء مانتل–هانزل لتحديد أداء البند التّفاضلي Mantel–Haenszel Procedure for Identifying Differential Item Functioning: هو إجراء إحصائي لمقارنة أداء مجموعتين من المستجيبين على أحد بنود الاختبار، حيث يتم إجراء مقارنات بين المستجيبين الذّين يتطابقون في السّمة أو البنية المقاسة في الاختبار في كل مجموعة.

أوميجا Omega (أو تسمّى أحيانًا "معامل أوميجا" أو "ماكدونالدز أوميجا"): أحد معاملات ثبات الاختبار الذي يُفترض أن بنوده تقيس سمة واحدة مشتركة (والذي عادة ما يناسب نموذج العامل العام الواحد). بشكل عام يمكن اعتبار هذا المعامل أكثر عملية من معامل ألفا من حيث الافتراضات التي يفترضها في البيانات.

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

بيزا PISA: اختصار لعبارة "برنامج تحصيل الطّلاب الدّوليين Programme for International Student Achievement". هذا التقييم الدّولي للتحصيل ترعاه منظّمة التّعاون الاقتصادي والتّنمية (OECD) مع أكثر من 40 دولة مشاركة.

تطوير الاختبار المتزامن Simultaneous Test Development: تطوير وبناء أدوات القياس باللغة الاصلية واللغة المستهدفة في نفس الوقت، باستخدام إجراءات موحدة ومعيارية لمراقبة جودة الترجمة. تستخدم المشاريع الدولية واسعة النطاق التطوير المتزامن بشكل متزايد من أجل تجنب المشكلة المتمثلة في أن النسخة المطورة بلغة واحدة لا يمكن ترجمتها /نقلها إلى جميع لغات الدراسة.

نسخة/إصدار الاختبار باللغة الأصلية Source Language Version: اللغة التي كُتب بها الاختبار في الأصل.

نمذجة المعادلة البنائية (Structural Equation Modeling (SEM: مجموعة من النماذج الإحصائية المعقدة، التي تُستخدم لتحديد البنية الأساسية لاختبار ما أو مجموعة من الاختبارات. غالبًا ما تُستخدم هذه النماذج للتحقيق في الاستدلالت السببية حول العلاقات بين مجموعة من المتغيرات.

نسخة/إصدار اللغة المستهدفة Target Language Version: اللغة التي يتم ترجمة/نقل الاختبار إليها. لذلك على سبيل المثال، إذا تمت ترجمة اختبار من الإنجليزية إلى الإسبانية ، فغالبًا ما تسمى النسخة الإنجليزية " إصدار اللغة الأصلية" وتسمى النسخة الإسبانية "إصدار اللغة المستهدفة".

ترجمة ونقل الاختبارات لثقافة جديدة (الإصدار الثاني) النسخة النهائيّة 2.4

أبعاد الاختبار Test Dimensionality: يشير إلى عدد الأبعاد أو العوامل التي يقيسها الاختبار. في كثير من الأحيان، يتم إجراء هذا التحليل إحصائيًا باستخدام أحد الإجراءات العديدة بما في ذلك، مخططات الجذر الكامن eigenvalue plots أو نمذجة المعادلة البنائية SEM.

معادلة درجات الاختبار Test Score Equating: إجراء إحصائي لربط الدرجات في اختباران يقيسان نفس البناء الفرضي بدون أن يفترض أن الاختباران متوازية تمامًا.

تيمز TIMSS: اختصار لعبارة "التوجهات في الدراسات الدولية للرياضيات والعلوم Trends in International Mathematics and Science Studies"، وهو تقييم دولي للصفوف الرابع و الثامن و والثاني عشر في عدد من الدول في مجالات الرياضيات والعلوم وترعاه الرابطة الدولية لتقييم التحصيل التربوي International Association for the Evaluation of Educational Achievement (IEA).

القياس متعدد الأبعاد الموزون WDMS. وهو أحد الإجراء إحصائية البديلة للتحقق من أبعاد الاختبار.