اللجنة الدولية للاختبارات

دليل ترجمة وتكييف الاختبارات (الطبعة الثانية)

النسخة النهائية 2.4

الرجاء اعتماد مرجعية هذا السند على النحو التالي:

International Test Commission. (2017). The ITC Guidelines for Translating and Adapting Tests (Second edition). https://www.intestcom.org/.[1]

[1] Translation completed by Boudouda Nedjem Eddine, Bensouilah Afaf Assia, Seridi Mohamed Elmoncef , Elafri Malika (Helath & society Unit SOPHILAB ; Université 8 mai 1945-Guelma ;Algeria) ; Ben Youssef Samia (Tunis University, Tunisia) ; Gana Kamel (University of Bordeaux, France), and approved by the Maghrebian Society of Psychology.

الاعتراف بالقيمة:

يود مجلس لجنة المراجعة الدولية أن يشكر اللجنة التي عملت لعدة سنوات على تطوير إصدار الطبعة الثانية من المبادئ التوجيهية لترجمة و تكييف الاختبارات، والمكونة من الستة أشخاص الآتية أسماءهم:

David Bartram، **SHL** (المملكة المتحدة)، **Giray Berberoglu**(الجامعة التقنية للشرق الأوسط–تركيا–)

Jacques Grégoire(جامعة لوفان الكاثوليكية –بلجيكا–)، **Ronald Hambleton** (رئيس لجنة جامعة ماساتشوستس أمهيرست –الولايات المتحدة الأمريكية–) **Jose Muniz** (جامعة أوفييدو –إسبانيا–)، و **Fons van de Vijver** (جامعة تيلبورغ –هولندا–).

كما لا يفوت مجلس لجنة المراجعة الدولية أن يشكر **Chad Buckendahl** (الولايات المتحدة الأمريكية)، **Anne Herrmann** و زملائها من شركة OPP Ltd (المملكة المتحدة)، **April Zenisky** (ماساتشوستس– الولايات المتحدة الأمريكية–) لمراجعتهم بعناية للنسخة السابقة للسند.

كما تعرب **اللجنة الدولية للاختبارات** (ITC) عن امتنانها لجميع المراجعين الآخرين في جميع أنحاء العالم الذين ساهموا بشكل مباشر أو غير مباشر في الإصدار الثاني من دليل توجيهات لجنة الاختبارات الدولية لترجمة الاختبارات وتكييفها.

ملخص:

أُعدت الطبعة الثانية من توجيهات اللجنة الدولية لترجمة الاختبارات و تكييفها بين عامي 2005 و 2015 لتعديل و تحسين الإصدار الأول، استجابة للتطورات في التكنولوجيا و ممارسات الاختبار. تنتظم التوصيات الثمانية عشر في ست فئات تسهيلا لاستخدامها: الشروط السابقة(الاولية)(3) ،تطوير الاختبار (5) ،التأكيد (4) ،الإدارة (2) ،التنقيط والتفسير (2) ،والمراجع (2). لكل دليل، يتم تقديم شرح و كذلك اقتراحات للممارسة. كما يتم توفير قائمة مرجعية لتحسين تنفيذ المبادئ التوجيهية.

المحتوى

الاعتراف بالقيمة.....2

ملخص.....3

المحتوى.....4

السياق العام.....5

التوصيات..... 8

مقدمة.....8

توصيات خاصة بالشروط الأولية.....8

توصيات لإعداد الاختبارات.....11

توصيات التحقق من الصحة / التأكيد.....17

توصيات متعلقة بإدارة الاختبار.....28

توصيات لتسجيل الدرجات وتفسيرها.....29

توصيات خاصة بالتوثيق.....31

كلمة ختامية.....33

المراجع.....34

الملحق أ. قائمة مرجعية لتوصيات اللجنة الدولية للاختبارات (ITC) لترجمة وتكييف الاختبارات....40

الملحق ب. مسرد المصطلحات.....42

السياق العام

شهد مجال الترجمة ومنهجية تكييف الاختبارات تقدماً سريعًا على مدار الخمسة والعشرين سنة الماضية، يتضح ذلك من العديد من الكتب والدراسات و نماذج كثيرة عن أعمال استثنائية لتكييف الاختبارات، التي تم نشرها (يمكنكم الاطلاع على اعمال كل من van de Vijver & Leung, 1997, 2000; Hambleton, Merenda, & Spielberger, 2005 ; Grégoire & Hambleton, 2009 ; Rios & Sireci, 2014).كانت هذه التطورات ضروريةً بسبب تزايد الاهتمام في هذا المجال من طرف الدراسات في(1) علم النفس بين الثقافات،(2) دراسات مقارنة دولية واسعة النطاق حول النجاح التعليمي (على سبيل المثال، الاتجاهات في دراسة الرياضيات والعلوم TIMSS والبرنامج الدولي لرصد تحصيل التلاميذ PISA /منظمة التعاون الاقتصادي والتنمية OCDE) ،(3) اختبارات الاعتماد المستخدمة في جميع أنحاء العالم (على سبيل المثال، في مجال تكنولوجيا المعلومات من قبل شركات مثل ميكروسوفت (Microsoft و سيسكو CISCO) ، (4) الاعتبارات المتعلقة بالنزاهة في الاختبار والتي تسمح للمرشحين باختيار اللغة التي سيتم تقييمهم بها (على سبيل المثال، في التسجيلات الجامعية بإسرائيل يمكن للمترشحين إجراء العديد من الاختبارات بإحدى اللغات الست المتوفرة).

أُحرز تقدم تقني في المناهج النوعية والكمية الخاصة بقياس البنيات النفسية، الأساليب والطرق المختلفة في الاختبارات والاستبيانات المكيفة، بما في ذلك استخدام الإجراءات الإحصائية المعقدة مثل نظرية الإجابة على البند، ونموذج معادلة الهيكل ونظرية التعميم (يمكنكم الاطلاع على اعمال : Hambleton et al., 2005 ؛ Byrne، 2008) اقترحت منظمة التعاون الاقتصادي والتنمية / البرنامج الدولي لرصد تحصيل التلاميذ L'OCDE/PISA نماذج ترجمة جديدة (يمكنكم الاطلاع على أعمال، Grisay، 2003) ؛ وقد اقترحت مجموعة من المناهج لتنفيذ مشروع تكييف الاختبارات (على سبيل المثال يمكنكم الاطلاع على أعمال Hambleton & Patsula، 1999 ؛ مشاريع مثالية متوفرة لتوجيه ممارسات تكييف الاختبارات – على سبيل المثال ، مشاريع OECD / PISA و TIMSS) ؛ والعديد من التطورات الأخرى التى تم تحقيقها.

ولدت الطبعة الأولى من هذه التوصيات (يمكنكم الاطلاع على أعمال de Vijver & Hambleton, 1996 ; Hambleton, 2005) من منظور المقارنة، وذلك بغرض السماح أو تسهيل المقارنات بين مجموعات المجيبين. و بالتالي فالنموذج الأساسي الذي كانت تهدف إليه هذه التوصيات، كان يعتمد على تطوير متتالي للأدوات في سياق مقارناتي (يجب تكييف الأداة الحالية لاستخدامها في سياق ثقافي

جديد). ومع ذلك فإنه من الواضح وبشكل متزايد أن تكييف الاختبارات يغطي مجالات أوسع للتطبيق. المثال الأكثر تعبيرا هو استخدام الأدوات، سواء كانت جديدة أو موجودة مسبقا، ضمن مجموعات متعددة الثقافات، مثل استشارة العملاء من مجموعات عرقية مختلفة، وكذلك استخدام امتحانات أكاديمية عند مجموعات عرقية مختلفة تحَكُمهُمْ مختلف او متباين في لغة الاختبار، أو التوظيف الدولي لمناصب الإدارة في الشركات متعددة الجنسيات. هذا التطور في مجال التطبيقات لديه عدة عواقب فيما يتعلق بتنمية و إدارة أدوات القياس، والتحقق من صحتها وتوثيقها. من بين النتائج المحتملة الحاجة إلى تكييف بنود اختبار موجود من أجل زيادة فهمه من طرف مشاركين لغة الاختبار ليست لغتهم الأم(على سبيل المثال، عن طريق تبسيط اللغة).

الامتداد الأخر والمهم للتوصيات يخص السماح بتطوير الأدوات في وقت واحد (أي التطوير المشترك للاستبيانات باللغة المصدر واللغة المستهدفة). تتجه البحوث الدولية واسعة النطاق وبكثرة إلى التطوير المتزامن للاختبارات، هذا لتجنب ترجمة أو تكييف الأداة المطورة بلغة معينة إلى جميع لغات الدراسة. نشرت الطبعة الأولى لتوصيات اللجنة العالمية للاختبارات لترجمة وتكييف الاختبارات من طرف van de Vijver et Hambleton (1996), Hambleton (2002), Hambleton, Merenda & Spielberger (2005)شهدت هذه النسخة بعض التغييرات التحريرية البسيطة بين عامي 1996 و 2005. في غضون ذلك، تم إحراز الكثير من التقدم. أولاً، كان هناك عدد من الملاحظات المهمة بشأن توصيات اللجنة العالمية للاختبارات، من بينها مقالات كل Jeanrie & Bertrand (1999), Tanzer (2002) Hambleton ,(1999) Sim &، والتي أكدت على جودة التوصيات، مع اقتراح طرق لتحسينها. نشر Hambleton, Merend & Spielberger (2005) وقائع مؤتمر دولي عقدته اللجنة العالمية للاختبارات في عام 1999 بجامعة جورج تاون في الولايات المتحدة. الكثير منهم كانوا من مؤلفي هذه الفعاليات، نجد من بينهم Cook & Schmitt-Cascallar (2005) ، و Sireci (2005) اللذين اقترحوا نماذج ومناهج جديدة لتكييف الاختبارات. نظمت اللجنة العالمية للاختبارات عام 2006 مؤتمراً دولياً في بروكسل (بلجيكا) لمناقشة توصيات اللجنة العالمية للاختبارات الخاصة بترجمة وتكييف الاختبارات.

قام أكثر من 400 مشارك من أكثر من 40 دولة بدراسة مسألة تكييف الاختبارات، وتم تطوير العديد من التوجهات المنهجية الجديدة، واقتُرحت توصيات جديدة أيضًا كما تم تبادل أمثلة لكيفية التنفيذ الناجح لهذه العملية. كانت الأوراق المقدمة في الندوات والاجتماعات الدولية من 1996 إلى 2009 عديدة (يمكنكم

الإطلاع على سبيل المثال على أعمال، Grégoire & Hambleton, 2009 ؛ كما ندعو القراء للاطلاع على الورقة التي أعدها كل من Muniz, Elosua et Hambleton (2013) الخاصة بالنسخة الأولية باللغة الإسبانية من الطبعة الثانية من توصيات اللجنة العالمية للاختبارات). في عام 2007 ، أنشأ مجلس الهيئة العالمية للاختبارات لجنة مؤلفة من ستة أشخاص مكلفين بتحديث توصيات الهيئة العالمية للاختبارات وذلك بهدف الاستفادة من التطورات التقنية الجديدة والخبرات المكتسبة من قبل الباحثين في هذا المجال. تشمل هذه التطورات (1) ظهور نماذج معادلة الهيكل لتقييم تكافؤ معامل الاختبار عبر مجموعات لغوية مختلفة (2) مناهج واسعة النطاق لتحديد مدى الأداء التفاضلي للبنود مع مقاييس تقييم متعددة الإجابات في مجموعات لغوية مختلفة و(3) نماذج جديدة لتكييف الاختبارات تم تطويرها بواسطة مشروعات التقييم الدولية مثل OECD / PISA و TIMSS. كما قدمت اللجنة وكتبت الصيغة الأولية للتوصيات في الاجتماعات الدولية لعلماء النفس في براغ (في عام 2008) وأوسلو (في عام 2009)، واستنبطت العديد من الملاحظات المفيدة للغاية.

تم الاحتفاظ بقسم توصيات إدارة الاختبار في الإصدار الثاني. في حين تم تخفيض العدد الإجمالي للتوصيات الواردة في هذا القسم من ست توصيات إلى اثنان لتجنب الازدواجية. خصص القسم الأخير من الطبعة الأولى لـ "التوثيق / تفسير النتائج". في الإصدار الثاني، قمنا بتقسيمها إلى قسمين منفصلين يتعلق أحدهما بالنتائج وتفسيرها، والآخر بالبحث الوثائقي. بالإضافة إلى ذلك، تم تعديل توصيتين من التوصيات الأربع الأولية في هذا القسم.

بنفس الطريقة المستخدمة في الطبعة الأولى، أردنا توضيح الأمر للقراء بشأن التمييز بين ترجمة الاختبار وتكييفه. قد تكون الترجمة هي المصطلح الأكثر استخدامًا، ولكن تكييف الاختبار هو مصطلح أوسع يشير لنقل اختبار من لغة وثقافة إلى أخرى. تكييف الاختبارات يشير الى عدة نشاطات، منها : تحديد ما إذا كان اختبار بلغة وثقافة ثانية يمكنه أو لا أن يقيس نفس البناء الخاص باللغة الأصلية، اختيار المترجمين، اختيار طريقة لتقييم عمل مترجمي الاختبار (على سبيل المثال، ترجمات ثنائية الاتجاه / عكسية / عكسية ؛ ترجمة أحادية الاتجاه) . اتخاذ قرارات بشأن جميع التعديلات اللازمة؛ تغيير شكل الاختبار؛ الإشراف على مراقبة الترجمة؛ التحقق من التكافؤ اللغوي للاختبار والثقافة بالإضافة الى إجراء دراسات الصلاحية الضرورية. من ناحية أخرى، فإن ترجمة الاختبار لها معنى أكثر تقييدًا، ويقتصر على اختيار اجتياز الاختبار من لغة ومن ثقافة إلى أخرى مع الحرص على الحفاظ على المعنى اللغوي. إن

ترجمة الاختبار ليست سوى جزء من عملية التكييف، التي تعتبر، على هذا النحو، بمثابة نهج مبسط للغاية لنقل اختبار من لغة إلى أخرى دون النظر إلى المعادلات الأكاديمية أو النفسية.

التوصيات

مقدمة

تعرف التوصيات في هذا المستند كدليل / إرشادات / معايير إجرائية / إجراءات وتقييم لتكييف الاختبارات أو التطوير المتزامن للاختبارات النفسية والتعليمية بغية استخدامها مع عينات مختلفة. في النص التالي، تم تنظيم 18 توصية حول ستة أقسام رئيسية : الشروط الأولية (3)، تطوير / بناء اختبار(5)، التأكد / والتحقق من صحة [التحليلات التجريبية] (4)، إدارة الاختبار(2)، التقييم والتفسير (2) ،والوثائق الخاصة بالاختبار (2).

القسم الأول بعنوان "الشروط الأولية" يصر على أنه يجب اتخاذ قرارات مهمة قبل البدء في أي عملية ترجمة / تكييف اختبار. يركز القسم الثاني "توصيات لتطوير الاختبارات" على عملية تكييف الاختبار. يتضمن القسم الثالث "التأكيد / قواعد المطابقة" توصيات تتعلق بتجميع الأدلة التجريبية المتعلقة بالتكافؤ اللغوي والموثوقية وصلاحية الاختبار بلغات وثقافات متعددة. تتعلق الأقسام الثلاثة الأخيرة ب"إدارة الاختبار" و"تسجيل النتائج و تفسيرها" و"التوثيق حول الاختبار". حيث كان التوثيق جانباً مهملاً في مبادرات تكييف الاختبارات في علم النفس والتعليم. نود أيضًا أن نطلب من محرري المجلات العلمية ومنظمات التمويل مزيدًا من التوثيق حول عمليات تكييف الاختبارات.

قدمنا لكل توصية توضيحات واقتراحات تسمح بتنفيذها في الممارسة العملية.

توصيات خاصة بالشروط الأولية (ش أ)

ش أ -1(1) الحصول إلزاما على إذن من صاحب حقوق الملكية الفكرية الخاصة بالاختبار قبل البدء في تكييفه

تفسير. تشير حقوق الملكية الفكرية إلى مجموعة من الحقوق التي يتمتع بها الأفراد على إبداعاتهم أو اختراعاتهم أو منتجاتهم. إنها تتعلق بحماية مصالح المبدعين من خلال منحهم الحقوق المعنوية والاقتصادية على إبداعاتهم. وفقًا للمنظمة العالمية للملكية الفكرية (www.wipo.int)، "يعد قانون حقوق المؤلف جزءًا من القطاع القانوني وعلى نطاق أوسع يخص الملكية الفكرية، التي تهدف عمومًا إلى حماية المصنفات و حماية حقوق الملكية الفكرية و مصالح المبتكرين والمبدعين من خلال منحهم حقوقًا حول أعمالهم ". هناك فرعين من الملكية الفكرية، الملكية الصناعية وحقوق المؤلف. يتعلق الفرع الأول ببراءات

حماية الاختراعات والرسوم والنماذج الصناعية والعلامات التجارية للمنتجات والأسماء التجارية. فيما يتعلق الفرع الثاني بالإبداعات الفنية (بما في ذلك الأعمال القائمة على التكنولوجيا) والإبداعات الأدبية. يتمتع المبتكر بحقوق خاصة حول ابتكاراته (مثل منع بعض التشويهات عند نسخها أو تكييفها). يمكن ممارسة حقوق أخرى (مثل عمل نُسخ) من قِبل أشخاص آخرين (على سبيل المثال، ناشر) ممن حصلوا على ترخيص من المؤلف أو صاحب حقوق الطبع والنشر. بالنسبة للعديد من الاختبارات، كما هو الحال بالنسبة للمصنفات المكتوبة الأخرى، قد يتنازل المؤلف عن حقوق النشر للناشر أو الموزع. مع الأخذ بعين الاعتبار أنها إبداعات من العقل البشري ، الاختبارات التعليمية والنفسية محمية بحقوق الملكية الفكرية. لا تتعلق حقوق الطبع والنشر في الغالب بالمحتوى الخاص بالبنود. (على سبيل المثال، لا يحق لأحد الحصول على عناصر مثل "1 + 1 = ..." أو "أشعر بالحزن")، بل التنظيم الأصلي للاختبار (هيكل الاختبار، شكله، ونظامه، أدوات التصحيح وما إلى ذلك). وبالتالي، فإن تقليد اختبار أصلي، أو الحفاظ على هيكله الأصلي ونظام التصنيف الخاص به، أثناء إنشاء عناصر جديدة، يشكل انتهاكًا لحقوق الملكية الفكرية للاختبار الأصلي. عندما يؤذن للمؤلف إجراء التكييف، يجب عليه احترام خصائصه الأصلية عند تعديل الاختبار (الهيكل، الأدوات، الشكل، التصحيح...)، إلا إذا تحصل على موافقة من صاحب الملكية الفكرية تسمح له بتعديل تلك الخصائص.

اقتراحات للممارسة. يجب أن يلتزم مؤلفو تكييف الاختبار بجميع قوانين حقوق المؤلف والاتفاقيات التي تحمي الاختبار الأصلي. عليهم أن يحصلوا على اتفاقية موقعة من قبل صاحب الملكية الفكرية (أي المؤلف أو الناشر) قبل الشروع في تكييف الاختبار. الاتفاقية يجب أن تحدد التعديلات المقبولة على خصائص الاختبار الأصلي وأن تحدد صاحب حقوق الملكية الفكرية للنسخة المكيفة.

ش أ–2(2) التقرب من العينة المستهدفة لتقييم درجة اللياقة / التوافق بين التعريف ومحتوى الهيكل المقاسة بواسطة الاختبار الأصلي وأن يكون كل بند كافيا للاستخدام المقصود (أو الاستخدامات المقصودة) لنتائج الاختبار.

تفسير. تتطلب هذه التوصية فهم الموضوع الذي تم تقييمه بنفس الطريقة من جانب جميع المجموعات اللغوية والثقافية المتنافسة/الموجودة، والتي تشكل أساس لمقارنات صحيحة بين الثقافات. في هذه المرحلة من الإجراء، الاختبار أو أداة القياس ليس مكيف بعد. وعليه من المستحسن جمع أدلة تجريبية من خلال توثيق اختبارات مماثلة، وتقييم اتفاق ملائمة / بناء البند للمجموعات اللغوية المستهدفة في الدراسة. ومع ذلك، ينبغي تقييم هذه التوصية باستخدام البيانات التجريبية، وفقًا للأدلة المطلوبة في **ت–2 (10)**. ليس

الهدف من أي تحليل تحديد هيكل الاختبار، على الرغم من أهمية هذا العنصر في أي تحليل، بل التأكد من ثبات (عدم تغير) هيكل الإصدارات اللغوية المختلفة للاختبار.

اقتراحات للممارسة. يجب توظيف خبراء يعرفون كل من البناء المقاس والمجموعات الثقافية المستهدفة من أجل تقييم مدى ملاءمة / مزايا البناء عند كل من هذه المجموعات. سيحاول هؤلاء الخبراء الإجابة عن السؤال التالي: هل للبناء معنى في كل ثقافة ؟. لقد حدث، كما رأينا ذلك في العديد من المرات في الاختبارات التعليمية على سبيل المثال، أن تحكُم لجنة بأن البناء المقاس باختبار لا معنى له أو أنه فقد معناه في ثقافة مختلفة (على سبيل المثال، نوعية الحياة، والاكتئاب أو الذكاء). يمكن إذن استخدام طرق مختلفة مثل مجموعات المناقشة والمقابلات والتحقيقات للحصول على مجموعة من المعلومات حول درجة توافق البناء بين الثقافات.

ش أ-3(3) الحد بشكل كبير من تأثير الاختلافات الثقافية واللغوية الضارة / غير المرغوب فيها / غير الضرورية في استخدام قصدي للاختبار في العينات المستهدفة.

تفسير. يجب تحديد الخصائص الثقافية واللغوية التي لا تتعلق بالمتغيرات التي من المفترض أن يقيسها الاختبار من بداية المشروع. فقد تكون لها صلة بشكل البنود والمعدات المستخدمة (على سبيل المثال استخدام الكمبيوتر أو الصور أو الصور التخطيطية، وما إلى ذلك)، ومدة تمرير الاختبار...إلخ. النهج المستخدم لتحقيق ذلك هو تقييم "المسافة اللغوية والثقافية" بين اللغة المصدر واللغة المستهدفة للاختبار : قد يشمل تقييم المسافة اللغوية والثقافية اعتبارات تتعلق بالاختلافات في اللغات (اللغوية) بناء الأسرة والدين ونمط الحياة والقيم (van de Vijver & Leung, 1997) .

تعتمد هذه التوصية بشكل أساسي على مناهج نوعية وتدعو بهذا الصدد المتخصصين الذين لديهم دراية بالبحث في الاختلافات الثقافية و اللغوية الخاصة. فهي ترتكز بشكل خاص على اختيار مترجمي الاختبار و تقتضي أن يكونوا من السكان الأصليين للغة والثقافة المستهدفة، لأن مجرد معرفة اللغة المستهدفة ليس كاف لتكون قادرًا على تحديد مصادر منهج التحيز المحتملة. على سبيل المثال، لوحظت مشاكل في شكل وطول الاختبار في دراسة المقارنة الصينية الأمريكية التي أجراها Hambleton, Yu & Slater (1999) حول معرفة الرياضيات لدى طلاب الصف الرابع، بالإضافة إلى مجموعة من الخصائص الثقافية المرتبطة باختبار الرياضيات المصممة لطلاب الصف الرابع.

اقتراحات للممارسة. من الصعب تلبية هذه التوصية في أي وقت كان بمساعدة البيانات التجريبية. و هذا يعتبر صحيح وبشكل خاص في المراحل الأولى من تكييف الاختبار. بالرغم من ذلك، يمكن غالباً جمع أدلة نوعية :

- سواء كان ذلك من خلال الملاحظة أو المقابلة أو مجموعة التعبير أو التحقيق، فهنا الأمر يتعلق بتحديد مستوى تحفيز المشاركين وفهمهم للتعليمات وتجربتهم في الاختبارات النفسية ، وسرعة تمرير الاختبار ، والألفة بمقاييس الاستجابة والاختلافات الثقافية (مهما ادت هذه المقارنات الى ظهور مشكلات بسبب الاختلافات الثقافية في فهم المتغيرات نفسها). عندما يمثل جمع هذه البيانات من المشاركين مشكلة ، احصل على أكبر قدر ممكن من المعلومات من المترجمين. جزء من هذه العملية يمكن القيام به قبل التقدم في تكييف الاختبار.

يبقى من الممكن أخذ هذه "المتغيرات الطفيلية" بعين الاعتبار في أي تحليل تجريبي لاحق وذلك بمجرد تكييف الاختبار و يصبح جاهزا لتقديمه إلى دراسات قواعد المطابقة. يتم ذلك عن طريق عملية تحليل التغاير أو التحليلات الأخرى التي تمكن ايضا ، من التحكم في مستوى التحفيز أو الإلمام بمقياس استجابة معين مع المشاركين من لغات وثقافات مختلفة (يمكنكم الاطلاع على أعمال Johnson, 2003 ; Javaras & Ripley, 2007).

توصيات لإعداد الاختبارات (إ ا)

إ ا-1(4) التأكد من أن إجراءات الترجمة والتكييف تأخذ بعين الاعتبار الاختلافات اللغوية والنفسية والثقافية للعينات المستهدفة من خلال اختيار الخبراء ذوي الخبرة اللازمة.

تفسير. هذه واحدة من التوصيات التي كان لها أكبر تأثير على مر السنين ، حيث توجد أدلة كثيرة تشير إلى أنها كانت ذات تأثير كبير على الأبحاث التي أجرتها وكالات التقييم و المترجمين المؤهلين الذين تتجاوز قدرتهم مجرد معرفة اللغتين للمشاركين في تكييف الاختبار (على سبيل المثال ، يمكنكم الاطلاع على أعمال ، Grisay, 2003). لقد أصبحت معرفة الثقافات والمعارف العامة للبناء المعني وبناء الاختبارات جزءًا من معايير اختيار المترجمين. بالإضافة إلى ذلك، يبدو أن هذه التوصية لعبت دورًا مهمًا في تشجيع منظمات الترجمة و تكييف الاختبارات على استخدام مترجِمَيْنِ على الأقل وفقًا للنموذج الذي تم اختياره (على سبيل المثال، نموذج الترجمة ثنائي الاتجاه / العكسي). الممارسة القديمة المتمثلة في الاعتماد على مترجم واحد يتخذ جميع القرارات ، بغض النظر عن مؤهلاته ، قد اختفت من قائمة الممارسات المقبولة اليوم.

المعرفة / الخبرة في الثقافة المستهدفة تتطلب استخدام مترجمين اللغة المستهدفة هي لغتهم الأم، ويفضل أن يكونوا يعيشون في المنطقة المستهدفة. لن ينتج الشخص القاطن بالمنطقة ترجمة دقيقة فحسب، بل سينتج أيضًا ترجمة سهلة القراءة تتوافق و السياق المحلي. العيش في المنطقة المستهدفة يسمح بالاطلاع المستمر على تحديث اللغة و استخدامها في الوقت الحالي. لذلك، فإن تعريفنا لـ "الخبير" هو شخص أو فريق لديه معرفة مشتركة كافية (1) باللغات المعنية، (2) بالثقافات المعنية، (3) بمحتوى الاختبار المعني، و (4) بالمبادئ العامة للاختبار، والهدف من ذلك هو إنتاج ترجمة / تكييف ذا جودة عالية. من المستحسن في الممارسة العملية، دعوة فرق ذات مؤهلات مختلفة (على سبيل المثال، مترجمون ذوي خبرة أو دونها في المجال المعين للاختبارات، وخبير في الاختبارات، وما إلى ذلك) وذلك من أجل تحديد الجوانب المحتمل إهمالها من قبل احدهم. في جميع الحالات ، معرفة المبادئ العامة للاختبارات، بالإضافة إلى معرفة محتوى الاختبارات، يجب أن يكون جزءًا لا يتجزأ من تدريب المترجمين.

اقتراحات للممارسة. نقترح ما يلي:

- يُنصح باختيار المترجمين المحليين الذين لغتهم الأم هي اللغة المستهدفة و لديهم معرفة شاملة بالثقافة التي تم تكييف الاختبار إليها. الخطأ الشائع هو اعتبار الأشخاص الذين يعرفون اللغة كمترجمين، في حين لا يعرفون جيدًا الثقافة المحلية، حيث أن المعرفة الدقيقة للثقافة غالبًا ما تكون ضرورية لضمان التكافؤ الثقافي لإصدارات الاختبار. امتلاك هذه المعارف الثقافية يساعد في تحديد المراجع الثقافية (على سبيل المثال ، لعبة الكريكيت ، برج إيفل ، الرئيس لينكولن ، الكنغر، وما إلى ذلك) و التي قد تكون غير مألوفة عند المشاركين الذين يستهدفهم التكييف.
- إذا أمكن اختر مترجمين ، لديهم خبرة بمحتوى الاختبار و يعرفون مبادئ الاختبار (على سبيل المثال ما يتعلق ببنود الاختيارات المتعددة، لا يجب أن تكون الإجابة الصحيحة أطول أو أقصر من الاختيارات الأخرى. لا ينبغي أن تشير الدلائل النحوية إلى الإجابة الصحيحة ؛ بالنسبة للبنود صحيح / خاطئ ، لا يجب أن تكون النصوص "الصحيحة" أطول بشكل أو بأخر عن النصوص "الخاطئة").
- في الممارسة العملية، من المستحيل إيجاد مترجمين لديهم معرفة بمبادئ بناء / تطوير الاختبار. لذلك من الضروري تدريب المترجمين على المبادئ الأساسية وكتابة البنود وشكل البنود التي سيتعاملون معها. في حالة غياب التكوين، فإن المترجمين الذين لديهم حس زائد يصبحون في هذا السياق مصدر خطأ ما يهدد صحة الاختبار المترجم. على سبيل المثال ، يمكن للمترجم أن يضيف أحيانًا ملاحظة توضيحية توحي الوصول إلى الإجابة الصحيحة. على هذا الأساس يمكن للمترجم أن يجعل السؤال أسهل

من ما كان عليه ، أو يجعل ايضا الإجابة الصائبة للاستبيان متعدد الخيارات QCM أطول موفرا بذلك مؤشرا للمرشحين الذين يجتازون الاختبار.

إ ا-2(5) استخدم تصميمات وإجراءات ترجمة مناسبة لزيادة ملائمة تكييف الاختبار مع المجموعات المستهدفة.

تفسير. تفرض هذه التوصية أن القرارات التي يتخذها المترجمون أو مجموعة المترجمين تزيد إلى حد أقصى الملائمة و النسخة التي تم تكييفها للعينات المستهدفة. هذا يعني أن الكلمات المستخدمة يجب أن تكون طبيعية ومقبولة ، وهذا بالآخذ بعين الاعتبار معادلة التكافؤ اللغوي الوظيفي بدلاً من التكافؤ الحرفي. نماذج الترجمة الأكثر شعبية لتحقيق هذه الأهداف هي الترجمات أحادية الاتجاه والترجمات ثنائية الاتجاه / الاتجاه المعاكس. يقدم كل من(1999) Hambleton & Patsula ,(1986) Brislin تحليلًا شاملاً للنموذجين، بما في ذلك تعريفهما و نقاط قوتهما وضعفهما. و مع ذلك، تجدر الإشارة إلى أن كلا النموذجين لهما نقائص، ونادراً ما يوفران أدلة للتحقق من صحة الاختبار المترجم والمكيف. العيب الرئيسي للترجمة ثنائية الاتجاه، إذا تم إجراؤها في أدق أشكالها، هو أنها تستبعد مباشرة أي فحص / تحقق من النسخة المترجمة. هذا الإجراء يولد في الكثير من الأحيان نسخة مترجمة من الاختبار ، مما يزيد من سهولة الترجمة العكسية / العكسية ، ما يؤدي في بعض الأحيان إلى ترجمات خرقاء.

يهدف الإجراء المزدوج للترجمة إلى التحقق و معالجة أوجه القصور والمخاطر الكامنة جراء الترجمات الفردية. في هذا النهج ، يقوم مترجم ثالث مستقل أو فريق من الخبراء بتحديد وحل جميع التناقضات بين الترجمات البديلة و يوحدهم في نسخة واحدة. في برامج التقييم عبر الثقافات واسعة النطاق مثل PISA ، يمكن استخدام نسختين مختلفتين من اللغات (على سبيل المثال ، الإنجليزية والفرنسية) كمصادر منفصلة لإنشاء ترجمتين ، يتم دمجهما بعد ذلك في نسخة لغوية واحدة مستهدفة(Grisay, 2003). يوفر هذا النهج مزايا مهمة مثل تحديد التناقضات المحتملة في اللغة المستهدفة وفحصها مباشرة. بالإضافة إلى ذلك ، يساعد استخدام أكثر من لغة مصدر في التقليل من تأثير الخصائص الثقافية.

الاختلافات في الهياكل اللغوية يمكن أن تسبب مشاكل في ترجمة الاختبارات. على سبيل المثال ، مقياس مشهور تم تطويره باللغة الإنجليزية من قِبل(1950) ,Rotter & Rafferty البنود عبارة عن جمل يجب إكمالها : "احب ..."؛ "أتأسف..." ؛ "لا استطيع...". ومع ذلك، فقد تبين أن نفس الشكل غير مناسب باللغة التركية، حيث يجب أن يسبق موضوع الجملة الفعل والفاعل. فاستخدام الجمل لإكمالها ، كما هو

الحال في النسخة الإنجليزية ، من شأنه أن يغير تمامًا سلوك الاجابة ، حيث أن الطلاب الأتراك يجب عليهم أن ينظروا أولاً إلى نهاية البيان قبل البدء في ملء البداية.

اقتراحات للممارسة. جمع أحكام الخبراء يبدو مفيدًا بشكل خاص للتحقق / و ضمان احترام هذه التوصية: استخدم مقاييس التقييم المقترحة من قبل Brislin (1986), Jeanrie & Bertrand (1999), Hambleton & Zenisky (2010). Hambleton يوفر و Zenisky قائمة معتمدة من الناحية التجريبية تضم 25 خاصية للاختبار المترجم والتي يجب التحقق منها أثناء عملية التكييف. فيما يلي بعض الأمثلة لأسئلة من هذه القائمة : "هل كلمات البنود المترجمة تحمل نفس الصعوبات و نقاط تتشابه مقارنةً بكلمات البنود في النسخة الأصلية ؟" و "هل ادخلت الترجمة تغييرات في النص ؟ (الإغفالات أو البدائل أو الإضافات) التي باستطاعتها أن تؤثر على صعوبة بند الاختبار في كلتا النسختين اللغويتين ؟"

- إذا أمكن ، استخدم عدة نماذج للترجمة. على سبيل المثال ، يمكن استخدام خطة ترجمة عكسية لإعادة التحقق من النسخة التي تم إنشاؤها من قبل لجنة من الخبراء بعد ترجمة مزدوجة وملخص للاثنين. إذا كان الاختبار مخصصًا للاستخدام عبر الثقافات، يجب من البداية النظر لتطوير متزامن لنسخ متعددة اللغات، ما يسمح بتجنب المشاكل المحتملة للترجمة / تكييف النسخة المصدر. على سبيل المثال نجد في أعمال Solano-Flores, Trumbull & Nelson-Barber (2002) مزيد من التفاصيل حول تطوير الاختبارات المتزامنة. يستحسن على الاقل تطوير نسخة مصدر لتسهيل الترجمات الممكنة وذلك لتجنب بقدر الامكان المشاكل المحتملة. و على وجه الخصوص ، ينبغي تجنب المراجع الثقافية ،العناصر المميزة / الحوارية / الفردية وأشكال الاجابة غير اللائقة ، إلخ.

- بالنظر إلى الاختلافات في بناء الجمل بين اللغات، يجب تجنب استخدام الأشكال التي ترتكز على هيكل صلب من الجمل؛ في التقييمات الدولية الواسعة النطاق و ربما أيضًا في الاختبارات النفسية، نظرًا لمشاكل الترجمة التي تواجهها هذه الأشكال.

إ ا-3(6) تقديم ادلة بأن تعليمات الاختبار ومحتوى البنود لها نفس المعنى لجميع المجموعات المستهدفة.

تفسير. يمكن جمع الأدلة المطلوبة لهذه التوصية من خلال مجموعة متنوعة من الاستراتيجيات (يمكنكم على سبيل المثال الاطلاع على أعمال van de Vijver et Tanzer, 1997) والتي تشمل (1) استخدام خبراء من السكان الاصليين ؛ (2) استخدام عينات من المجيبين تتحكم في لغتين ؛ (3) استخدام تحقيقات محلية لتقييم الاختبار ؛ و (4) استخدام تطبيقات الاختبار غير الموحدة من أجل زيادة مقبوليته

وصلاحيته. إجراء دراسة تمهيدية / تجريبية للنسخة المعدلة للاختبار و هي فكرة مستحسنة. لا تتضمن هذه الدراسة فقط تطبيق الاختبار وتحليل البيانات، ولكن أيضًا، والأهم من ذلك، إجراء مقابلات مع كل من يقوم بتطبيق الاختبار و المشاركين في الاختبار بغرض جمع أحكامهم بشأن الاختبار. هناك طرق أخرى ممكنة مثل دعوة مختصين من لغات مختلفة أو مختصين في لغتين لتقييم المحتوى. على سبيل المثال ، يمكن أن يطلب من المختصين في لغتين تقييم التشابه و الصعوبة التي يتسبب فيها شكل البنود ومحتوى كل نسخة. المقابلة المعرفية هي طريقة واعدة أخرى(Levin et coll., 2009) .

اقتراحات للممارسة. قدمت عدة اقتراحات أعلاه لتلبية هذه التوصية. فيما يلي بعض الأمثلة:

- استخدم خبراء محليين من الثقافة واللغة المحلية لتقييم ترجمة / تكييف الاختبار.
- استخدم عينات من المجيبين يتحكمون في اللغتين لتجميع اقتراحات حول معادلة النسختين من الاختبار، سواء فيما يخص التعليمات أو بنود الاختبار.
- إجراء تحقيقات محلية لتقييم الاختبار.هذه الدراسات الأولية يمكن أن تكون مفيدة للغاية. تأكدوا من إجراء مقابلات مع كل من طبق الاختبار والمجيبين بعد إجراء الاختبار، حيث إن تعليقاتهم غالبًا ما تكون جد قيمة من ردود الفعل البسيطة للمشاركين حول بنود الاختبار.
- تكييف إدارة الاختبار لزيادة مقبوليته وصلاحيته.
- الامتثال لتعليمات مطابقة لا معنى له ، إذا أسيء فهمها من قبل المجيبين من اللغة الثانية/ المجموعة الثقافية.

إ ا–4(7) اثبات ان اشكال البنود مقاييس الإجابة التصنيف وفئات التصنيف واتفاقيات الاختبار وطرق ادارته وأي إجراء آخر مناسبة لجميع الفئات المستهدفة.

تفسير. قد تكون أشكال البنود مثل مقاييس الإجابة المكونة من خمس نقاط أو أشكال البنود الجديدة مثل "اسحب–ضع" أو "الإجابة على كل ما هو صحيح" أو حتى "اختيار إجابة واحدة" مصدر ارتباك للمجيبين الذين لم يروا قط هذا الشكل من البنود. حتى تصميم البنود أو استخدام الرسومات أو الظهور السريع لشكل بنود إلكترونية يمكن أن يكون مربكًا للمترشحين. هناك العديد من الأمثلة حول هذا النوع من الأخطاء في الولايات المتحدة التي بادرت إلى حوسبة الكثير من الاختبارات الموحدة للأطفال. بفضل التمارين التجريبية ، يمكن التغلب على هذه المشاكل عند معظم الأطفال. يجب أن تكون اشكال البنود الجديدة مألوفة لدى المجيبين ، لأنها تستطيع ادخال مصدر تحيز يمكن أن يشوه نتائج الاختبار.

اقتراحات للممارسة. الأدلة التي تستند على البيانات النوعية والكمية تلعب دورا في الأخذ بعين الاعتبار هذه التوصية. هناك العديد من خصائص الاختبار المكيف التي تتطلب التحقق منها:

- التحقق أن التمارين / و بنود التدريب كافية لجلب المجيبين إلى المستوى المطلوب ، ما يسمح لهم بتقديم إجابات صادقة و / أو إجابات تعكس مستوى إتقانهم لأدوات الاختبار.
- التأكد من أن المجيبين على دراية بشكل البنود الجديدة أو الطرق الجديدة لإدارة الاختبار (مثل الإدارة بمساعدة الكمبيوتر / الإدارة المحوسبة) التي أصبحت الآن جزءًا من إجراءات الاختبار.
- التحقق من أن جميع اتفاقات الاختبار (على سبيل المثال مكان الرسوم التوضيحية أو تدوين العلامات في ورقة الإجابة) واضحة للمجيبين.

هنا أيضًا، يمكن استخدام شبكة التقييم المقدمة من طرف ,(1999) Jeanrie & Bertrand (2010) Hambleton & Zenisky. على سبيل المثال ، Hambleton و Zenisky ادخلوا أسئلة مثل "هل شكل البنود ، بما في ذلك العرض التقديمي الفعلي ، هو نفسه في كلا النسختين و باللغتين ؟ و "إذا تم استخدام شكل من أشكال الكلمات أو الجمل (غامق أو مائل أو مسطر أو غير ذلك) في البند المصدر ، يجب التأكد ما اذا تمت مراعاة هذا الشكل في البند المترجم؟

إ ا-5(8) جمع البيانات الخاصة بالدراسات التجريبية للاختبار المكيف من أجل إجراء تحليل البنود وتقييم وثوق وصلاحية القيمة التي تسمح بالمراجعات اللازمة.

تفسير. من المهم الحصول على أدلة مسبقة تؤكد الجودة السيكومترية للاختبار المكيف. قبل إجراء دراسات واسعة النطاق حول وثوق وصلاحية درجات الاختبار / أو الدراسات المرجعية التي تستغرق وقتًا طويلاً و تكون باهظة الثمن. يمكن إجراء العديد من التحليلات السيكومترية للحصول على أدلة أولية حول وثوق وصلاحية النتائج. على سبيل المثال ، في مرحلة تطوير الاختبار ، يمكن أن يوفر تحليل البنود باستخدام عينة صغيرة (ع = 100) بيانات مفيدة للغاية عن أداء بعض بنود الاختبار. يمكن مراجعة البنود التي تكون سهلة للغاية أو صعبة للغاية مقارنة بالعناصر الأخرى، أو التي تظهر احتمال تمييز منخفض أو سلبي. في حالة بنود الاختيار المتعددة، من المستحسن دراسة فاعلية البنود المسلية. هذا ما يسمح برصد المشاكل و إجراء التغييرات المناسبة. بالإضافة إلى ذلك ، باستخدام نفس البيانات التي تم جمعها لتحليل البنود ، يوفر حساب معامل ألفا أو معامل أوميغا (McDonald, 1999) معلومات قيمة يمكن استخدامها لدعم القرارات المتعلقة بالطول المناسب للنسخ اللغوية المصدر و المستهدفة للاختبار.

في بعض الحالات، هناك تساؤلات قد تبقى موجودة حول بعض جوانب التكييف : هل تعليمات الاختبار سيتم فهمها ؟ هل يجب أن تكون التعليمات مختلفة لتوجيه فعال للمترشحين المنحدرين من اللغة الجديدة والثقافة الجديدة الموجه لهم الاختبار المكيف ؟ هل ادارة الاختبار باستخدام الكمبيوتر ستؤثر سلبا على بعض المجيبين ؟ (على سبيل المثال ، المجيبين ذوي وضع اجتماعي واقتصادي منخفض) من السكان المستهدفين للاختبار المكيف ؟ هل هناك أسئلة كثيرة جدًا بالنظر إلى مدة الاختبار ؟ يمكن الإجابة على كل هذه الأسئلة و غيرها من خلال دراسات الصلاحية الأولية. الهدف من هذا هو تجميع بيانات كافية تسمح باتخاذ قرار حول ما إذا كان يجب الاستمرار في إجراء تكييف الاختبار.إذا تم اتخاذ قرار مواصلة الإجراء، يمكن تخطيط سلسلة من الدراسات واسعة النطاق وتنفيذها (على سبيل المثال ، دراسات لفحص مدى سير مستوى تفارق البنود ، وهيكل عامل الاختبار)

اقتراحات للممارسة. يمكن إجراء عدد من التحليلات الأساسية:

- إجراء دراسة كلاسيكية لتحليل البنود و الحصول على معلومات حول المتوسطات و مؤشرات تمييز البنود، وأيضًا إجراء تحليل الانتباه للبنود ذات الاختيارات المتعددة.
- إجراء تحليل الموثوقية (على سبيل المثال ، KR–20 مع بنود ثنائية التقسيم ، أو معامل ألفا أو معامل أوميغا مع بنود متوافقة/مركبة).
- إذا لزم الأمر، قم بإجراء واحد أو اثنين من الدراسات الأولية (التجريبية) لفهم افضل لصحة الاختبار المكيف. لنفترض، على سبيل المثال، أن الاختبار المكيف يدار على جهاز كمبيوتر.قد يكون من المستحسن إجراء دراسة لتقييم كيفية إدارة الاختبار (مثلا شكل قلم–ورقة مقابل الاختبار على الكمبيوتر). لنفترض أن تعليمات الاختبار تدعو المشاركين إلى الإجابة على جميع البنود. قد يكون من الضروري القيام ببحوث لتحديد أفضل الطرق لصياغة مثل هذه التعليمات تسمح بتحقيق هذا الهدف. فقد وجد الباحثون أنه من الصعب جلب بعض المجيبين للرد على جميع البنود عندما نشجعهم على تخمين الإجابات.

توصيات التحقق من الصحة / التأكيد

ت –1(9) اختيار عينة تكون خصائصها ذات صلة بالاستخدام المقصود للاختبار و حجمها وأهميتها كافية للتحليل التجريبي.

تفسير. اعداد جمع البيانات يشير إلى كيفية جمع البيانات لوضع معايير (إذا لزم الأمر) التكافؤ بين الإصدارات اللغوية للاختبار، ولإجراء دراسات الصلاحية والموثوقية ، و دراسات العملية التفاضلية للبنود.

الشرط الأول فيما يتعلق بجمع البيانات هو أن تكون العينات كبيرة بما فيه الكفاية للسماح بتوفر معلومات إحصائية مستقرة. على الرغم من أن هذا الشرط ينطبق على أي نوع من البحوث ، إلا أنه ذو أهمية خاصة في سياق دراسة مطابقة تكييف الاختبار، لأن التقنيات الإحصائية اللازمة لإثبات معادلة أو مطابقة الاختبار و البنود (مثل تحليل عامل التأكيد ، أنماط الاجابة على البنود لتحديد تلك التي يحتمل أن تكون متحيزة) هي عمليات لا يمكن تطبيقها على نحو فعال سوى على عينات كبيرة بما فيه الكفاية ما يسمح بتقدير موثوقية معلمات النموذج (يعتمد الحجم الموصى به لعينة ما على مدى تعقيد البيانات و وطبيعتها).

بالإضافة إلى ذلك ، يجب أن تكون عينة دراسة المطابقة واسعة النطاق و ممثلة للمجتمع المستهدف من خلال الاختبار. نلفت الانتباه إلى الوثيقة المهمة التي أعدها van de Vijver and Tanzer (1997) و المساهمات المنهجية لكل من van de Vijver and Leung (1997) و Hambleton ، Merenda Spielberger (2005) و ، Byrne (2008) ، Byrne and van de Vijver (2014) ، لتوجيه اختيار الخطط والتحليلات الإحصائية المناسبة. ناقش Sireci (1997) المشاكل والقضايا المتعلقة بإنشاء رابط بين الاختبارات متعددة اللغات ومقياس مشترك (اختبار).

قد يحدث في بعض الأحيان ، في إطار الممارسة الميدانية أن يحصل المجتمع المعني بنسخة الاختبار المعدل على نتائج أقل أو أعلى كثيرًا، و/أو تكون أكثر أو أقل تجانسًا من تلك التي يتحصل عليها باللغة المصدر أو الأصلية. هذا ما يخلق مشاكل كبيرة لبعض طرق التحليل، مثل دراسات الموثوقية والصلاحية. أحد الحلول المقترحة هو اختيار عينة فرعية من المجتمع المصدر تتطابق و المجتمع المستهدف. فباستخدام العينات المتطابقة ، يمكن التخلص من أي اختلاف في الشكل والتوزيعات بين المجموعتين (انظر Sireci و Wells، 2010). على سبيل المثال ، تحتوي مقارنات هياكل الاختبار عمومًا على تغايرات ، تختلف باختلاف توزيع النتائج. باستخدام العينات المتطابقة ، يمكننا استبعاد دور توزيع الدرجات على النتائج في تفسير الفروق .

يمكن أن يساعد مثال آخر في توضيح مشكلة التوزيع المختلف للدرجات في المجموعات اللغوية المصدر والمجموعات المستهدفة. افترض أن موثوقية نتيجة الاختبار هي .80 في مجموعة اللغة المصدر ، ولكن فقط .60 في المجموعة اللغوية المستهدفة. قد يبدو الفرق مقلقًا ويطرح تساؤلات حول أهمية نسخة اختبار اللغة الهدف. ومع ذلك ، غالبًا ما ننسى أن الموثوقية هي سمة مشتركة بين الاختبار والمجتمع (McDonald، 1999) لأنها تعتمد على كل من التباين الحقيقي للنتائج (خاصية المجتمع) وتباين

الأخطاء (سمة الاختبار). لذلك ، يمكن أن يؤدي التباين نفسه في الخطأ إلى زيادة الموثوقية ببساطة بسبب التباين الأكبر في النتيجة الفعلية في مجموعة اللغة المصدر. يوضح McDonald (1999) أن الخطأ المعياري للقياس (وهو الجذر التربيعي لتباين الخطأ) هو في الواقع كمية أكثر ملاءمة لمقارنة العينات وليس الموثوقية. البديل الآخر ، وهو استخدام معاملات الموثوقية ، أي أخذ عينة متطابقة من المرشحين من مجموعة اللغة المصدر وإعادة حساب موثوقية نتائج الاختبار.

تسمح الأساليب الحديثة لاختبار الثبات المتري (عدم التغير) و الهيكلي باستخدام تحليل عوامل المطابقة متعدد المجموعات (CFA) بتقييم العينات ذات التوزيعات المختلفة للسمات الكامنة. في مثل هذه النماذج ، على الرغم من افتراض أن معلمات القياس مثل أحمال عامل البند واعتراضاته متساوية بين المجموعات ، فقد تختلف المتوسطات الحسابية و التباينات المترابطة بين الصفات الكامنة. يسمح ذلك باستخدام عينات كاملة مع مراعاة السيناريو الأكثر واقعية لتوزيعات مختلفة من السمات المقاسة في مجتمعات مختلفة.

اقتراحات (عملية) للممارسة. في جميع الأبحاث تقريبًا ، تم تقديم اقتراحين لوصف العينة (العينات):

- خذ عينة كبيرة قدر الإمكان ، لأن دراسات تحديد البنود التي يحتمل أن تكون متحيزة تتطلب ما لا يقل عن 200 شخص لكل نسخة من الاختبار (Mazor، Clauser & Hambleton، 1992 ؛ Subok، 2017). مطلوب أيضا عينة من 500 شخص على الأقل لإجراء تحليلات نظرية الاستجابة للبنود ودراسات ملائمة النموذج ، (Hulin، Lissak و Drasgow، 1982 ؛ Hambleton، Swaminathan and Rogers، 1991) ، بينما تتطلب الدراسات التي أجريت على هيكل عامل الاختبار عينة كبيرة إلى حد ما ، ما يتراوح بين 300 مستجوب أو أكثر (Wolf، Harrington، Clark and Miller، 2013). من الواضح أن التحليلات باستخدام عينات أصغر ممكنة أيضًا، ولكن القاعدة الأولى هي إنشاء عينات كبيرة كلما كان ذلك ممكنًا.
- حيثما أمكن ، اختر عينات تمثيلية من المستجيبين. فتعميم النتائج المستخلصة من عينات غير ممثلة للمستجيبين تكون محدودة. من المستحسن غالبًا أخذ عينة من مجموعة اللغة المصدر لمطابقة مجموعة اللغة المستهدفة. هذا يسمح بإزالة الفروقات في النتائج التي تتسبب فيها العوامل المنهجية مثل الاختلافات في توزيع الدرجات ، حيث قد تكون مقارنات الأخطاء المعيارية في القياس أكثر ملاءمة.

ت–2 (10) تقديم بيانات إحصائية مقبولة /موثوقة تخص معادلات البناء المناهج و البنود عبر المجتمعات المستهدفة

تفسير. يعتبر من المهم تحديد التكافؤ المفاهيمي للإصدارات اللغوية للاختبار أي بين اللغة المصدر واللغة المستهدفة ، لكن هذا لا يعتبر التحليل التجريبي الوحيد المهم الذي يجب إجراؤه. بالإضافة إلى ذلك ، تمت مناقشة طرق تكافؤ البناء **(ش أ–2)** وتكافؤ الطريقة **(ش أ–3)** بصورة وجيزة في التوصيات الرئيسية.

يحتاج الباحثون أيضًا الاهتمام بالتكافؤ على مستوى البنود عبر المجموعات اللغوية المختلفة. تتم دراسة معادلة البنود تحت عنوان "تحليل الأداء التفاضلي للبنود (DIF)". يعرض البند و عامل التفاضل عندما يجتاز شخصين من مجموعتين مختلفتين (ثقافية ولغوية) الاختبار وعلى الرغم من تساوي مستوى السمة و المهارة المقاسة إلا أن هناك احتمال مختلف للاستجابة / النجاح في هذا البند لديهم. فمن الممكن أن يحدث عامة اختلاف في أداء الاختبار بين المجموعات ، لكن هذا لا يعتبر مشكلة في حد ذاته. فعندما تتم مطابقة أفراد المجتمع في البناء المقاس بواسطة الاختبار (عمومًا درجة اختبار كاملة أو درجة اختبار كاملة مطرحًا منها درجة البند المدروس) ، وعندما توجد فروق في الأداء على البند بين المجموعات فهذا البند يعرض عملية تفاضلية. يتم إجراء هذا النوع من التحليل لكل بند من بنود الاختبار. بعد ذلك ، يتم إجراء محاولة لفهم أسباب وجود IPIS العملية التفاضلية في بعض البنود، ووفقًا لهذا الفحص التقديري ، يمكن تحديد بنود معينة على أنها معيبة وتعديلها أو سحبها بالكامل من الاختبار.

تعد مشكلات الترجمة والاختلافات الثقافية من أهم المصادر المحتملة ل" الأداء التفاضلي للبند (DIF)" التي يجب أن تخضع للتقييم. وبشكل أكثر تحديدًا ، قد يكون الأداء التفاضلي للبند (DIF) ناتجا عن (1) عدم تكافؤ الترجمة بين اللغة المصدر واللغة المستهدفة للاختبار ، مثل الإلمام بالمفردات المستخدمة ، وتغيير في صعوبة البند ، وتغيير 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؛ Simon، and Ercikan & Oliveri، 2013).

أثناء الترجمة ، من الممكن استخدام مفردات أقل شيوعًا في اللغة المستهدفة. قد تكون المعاني متشابهة في الإصدارات المترجمة ، ولكن قد تكون كلمة أكثر شيوعًا من الأخرى في ثقافة واحدة. من الممكن

أيضًا تغيير مستوى صعوبة البند بسبب طول الجمل وتعقيدها واستخدام مفردات سهلة أو صعبة. يمكن أن يتغير المعنى أيضًا في اللغة المستهدفة بحذف أجزاء معينة من الجمل وترجمات غير دقيقة لديها تعدد في المفردات المستخدمة في اللغة الهدف وانطباعات حول معنى بعض الكلمات من ثقافة إلى أخرى الخ. يمكن أن تعمل البنود بشكل مختلف من لغة إلى أخرى بسبب الاختلافات الثقافية. على سبيل المثال ، قد لا يتم فهم كلمات مثل "همبرغر" أو "الصندوق النقدي" أو يكون لها معان مختلفة في ثقافتين.

هناك أربع مجموعات على الأقل من التحليلات للتحقق مما إذا كانت البنود تعمل بشكل مختلف من مجموعة لغوية و / أو ثقافية إلى أخرى. Ellis ، 1989 ؛ Thissen، Steinberg and Wainer، 1988؛ 1993، Ellis and Kimmel، (1992، (b) (إجراء Mantel–Haenszel MH) وتوسعاته (انظر ، على سبيل المثال ، Dorans and Holland، 1993؛ Hambleton، Clauser، Mazor and Jones، 1993؛ Holland and Wainer، 1993؛ Holland and Wainer، 1993؛ Sireci and Allalouf، 2003)، (c) إجراءات الانحدار اللوجستي (Swaminathan and) (LR Rogers، 1990؛ Rogers and Swaminathan، (d) and (1993) تحليل العوامل المقيدة (Oort and Berberoğlu، 1992).

في المقاربات المعتمدة على نظرية الرد على البند ، تتم مطابقة المرشحين المترشحين للاختبار في اللغتين وفقا للدرجات ذات الصفات الكامنة. في منهجيات البعد العالي HD و الانحدار المنطقي LR ، يتم استخدام الدرجة المرصودة أو المقدرة للاختبار كمعيار مطابق قبل مقارنة أداء المستجيبين من المجموعتين. على الرغم من أن الدرجة الإجمالية الملاحظة هي معيار المطابقة الأكثر شيوعًا في هذه الإجراءات ، إلا أنه يمكن استخدام الدرجات التقديرية الأخرى أيضًا ، على سبيل المثال انطلاقا من تحليل العوامل. يتم "تنقية" هذه النقاط بشكل تكراري أيضًا عن طريق إزالة البنود المشكوك فيها. يجب أن يكون معيار المطابقة صحيحًا وموثوقًا فيه بدرجة كافية لتسمح بالتقييم الصحيح للأداء التفاضلي للبنود. في عملية تحليل العوامل المقيدة ،يعتمد كل بند على المتغير الكامن مشيرا لعضويته للمجموعة (على سبيل المثال المتغير الذي يحتمل أن يعطل تناسق البنود) و كذلك السمات الكامنة. يشكل كل عامل تشبع وحدة ضبط تقدر في النموذج. يتم تقييم ملائمة هذه الأخيرة مقارنة بالنموذج الفارغ أين المتغير الكامن المنتمي للمجموعة لا يشبع البند المعني إذا كان النموذج يوفر ملاءمة أفضل بكثير ، فسيتم وضع في هذا البند علامة (DIF أي أن البند متحيز).

عندما يكون الاختبار معقدًا على مستوى الأبعاد ، يصعب العثور على معيار مطابقة مناسب (Clauser، Nungester، Mazor & Ripkey، 1996). يمكن أن يؤدي استخدام معايير المطابقة متعددة المتغيرات ، مثل درجات العامل المختلفة التي تم الحصول عليها بعد تحليل العوامل ، إلى تغيير تفسير الأداء التفاضلي للبنود. لذلك، تشير هذه التوصيات إلى أنه في حال الاختبار متعدد الأبعاد، يمكن للباحثين استخدام معايير مختلفة لاكتشاف وتقييم البنود التي تعرض أداء تفاضلي. مطابقة المتغيرات المتعددة قد يقلل من عدد البنود التي تعرض أداء تفاضلي FID عبر المجموعات اللغوية والثقافية.

تدعو هذه التوصية الباحثين لتحديد مصادر الانحياز المحتملة في الاختبار المكيف. تشمل مصادر التحيز المنهجي على (1) مستويات التحفيز المختلفة للمشاركين في الاختبار ، (2) الفرق في ألفة المجيبين مع الاختبارات النفسية ، (3) اجتياز الاختبار بشكل أسرع في مجموعة لغوية واحدة مقارنة بالمجموعة الأخرى. (4) معرفة مختلفة لتنسيق الاستجابة بين المجموعات اللغوية ، (5) عدم تجانس أسلوب الاستجابة ، إلخ. كانت على سبيل المثال تحيزات الاستجابة ، مصدر قلق كبير في تفسير نتائج تحقيق PISA و لقد أصبحت موضوع اهتمام عدد من البحوث .

أخيرًا ، وهذه نقطة مهمة ، في هذه التوصية التي ستفرض على الباحث الاهتمام بمسألة معادلة البناء. حيث أن هناك أربعة مناهج إحصائية على الأقل لتقييم معادلة البناء في إصدارات اللغة المصدر واللغة الهدف للاختبار و هي : تحليل العوامل الاستكشافية (EFA) ، وتحليل عوامل المطابقة (CFA) ، والتحجيم متعدد الأبعاد (MDS) ، ومقارنة الشبكات الاسمية (Sireci، Patsula، Hambleton &، 2005).

وفقًا van de Vijver و Poortinga (1991) ، فإن تحليل العوامل (FTE و CFA) هو الأسلوب الإحصائي الأكثر استخدامًا لتقييم معادلة البناء عبر اللغات والثقافات المختلفة. لا يزال تأكيد عمل 1991 هذا ساري الفعل حتى اليوم ، على الرغم من التقدم الكبير في أساليب النمذجة الإحصائية (انظر على سبيل المثال ، Hambleton & Lee، 2013، Byrne، 2008). نظرًا لصعوبةAFE تحليل العامل الاستكشافي مقارنة بهياكل العوامل المنفصلة و على الرغم من أنه لا توجد قواعد مشتركة للحكم على تكافؤ هذه الهياكل ، فالمقاربات الإحصائية مثل التحليل القانوني للارتباط ACC (انظر على سبيل المثال أعمال ، Byrne ، 2001 ، 2003 ، 2006 ، 2008) و WMDS (القياس متعدد الأبعاد المرجح) تعتبر مستحبة لأنها تسمح بتقييم عدة مجموعات في وقت واحد (Sireci، Harter، Yangو Bhola ، 2003).

في العديد من الدراسات، تم استخدام التحليل القانوني للارتباط ACC لتقييم ما إذا كان هيكل عامل الإصدار الأصلي للاختبار متناسقًا في جميع إصداراته المعدلة (على سبيل المثال ، Byrne and van de Vijver، 2014). يعتبر التحليل القانوني للارتباط ACC عملية جديرة بالاهتمام لأنها تسمح من جهة بتقييم التكافؤ الهيكلي للاختبارات المعدلة مع معالجة عدة مجموعات في وقت واحد ، و من جهة اخرى فهي توفر اختبارات إحصائية ومؤشرات وصفية لمدى ملائمة النموذج للبيانات (Sireci، Patsula ، و Hambleton، 2005). فإمكانية معالجة عدة مجموعات في وقت واحد أمر جد مهم حيث قد أصبح من الشائع تكييف الاختبارات بالعديد من اللغات و في وقت واحد(على سبيل المثال ، يتم الآن ترجمة / تكييف بعض اختبارات الذكاء في أكثر من 100 لغة ، وفي دراسات TIMSS الاتجاهات في الدراسات الدولية في الرياضية والعلوم و البرنامج الدولي لتقييم التلاميذ OECD / PISA المنظمة الدولية للتعاون والتنمية الاقتصادية ، يتم تكييف الاختبارات بأكثر من 30 لغة). ومع ذلك ، نظرًا لأن المطلب الصارم المتمثل في عدم وجود الشحنات المتقاطعة في ACC التحليل القانوني للارتباط لا يتوافق غالبًا مع البيانات المتعلقة بالأدوات المعقدة متعددة الأبعاد ، فإن النمذجة الاستكشافية للمعادلات الهيكلية (ESEM) تزداد شعبية ، لا سيما بالنسبة لبيانات الشخصية أو المتغيرات الأكثر تعقيدًا وترابطًا (Asparouhov & Muthén، 2009).

يعتبر التحجيم متعدد الأبعاد الموزونWMDS طريقة أخرى ملفتة للانتباه لأنها تسمح بتقييم التكافؤ المفاهيمي بين الإصدارات اللغوية المختلفة للاختبار. كما هو الحال مع معدل النشاط بمكافئات الدوام الكامل EPT ، لا يتطلب تحليل التحجيم متعدد الأبعاد الموزون WMDS تحديدًا مسبقًا لهيكل الاختبارات ، كما هو الحال مع ACC التحليل القانوني للارتباط ، فإنه يسمح بتحليل عدة مجموعات (مثل Sireci et al، 2003).

اقترح Van de Vijver and Tanzer (1997) أنه ينبغي للباحثين عبر الثقافات أن يدرسوا مدى موثوقية كل نسخة ثقافية من الاختبار المعني والبحث عن أدلة صحة متقاربة ومميزة في كل مجموعة ثقافية. غالبًا ما تكون هذه الدراسات عملية أكثر من الدراسات التي أجريت على بنية عامل الاختبار والتي تتطلب عينات كبيرة جدًا.

ومع ذلك ، يجب الاعتراف بأن مقارنة أداء المرشحين للاختبار بين نسختين من لغتي الاختبار ليست دائمًا هدفا لترجمة أو تكييف الاختبار. على سبيل المثال قد يتمثل الهدف ببساطة في التمكن من تقييم المرشحين من مجموعة لغوية مختلفة عن تلك التي تم تطوير الاختبار من أجلها. في هذه الحالة ، من

الضروري أن تدرس بعناية صحة الاختبار في مجموعة اللغة الثانية ، لكن البحث عن دليل الأداء في النسختين ليس ضروري. أهمية هذا المبدأ التوجيهي تعتمد على هدف (أهداف) اختبار اللغة الثانية (أي مجموعة اللغة المستهدفة). تتطلب الاختبارات مثل تلك المستخدمة في PISA أو TIMSS وجود دليل تشابه كبير في المحتوى من إصدار إلى آخر لأن النتائج المتحصل عليها تُستخدم لمقارنة تحصيل الطلاب في العديد من البلدان. إن استخدام مقياس الاكتئاب مترجم من الإنجليزية إلى الصينية للسماح للباحثين بدراسة الاكتئاب أو المعالجين/المهنيين بتقييم اكتئاب عملائهم لن يتطلب تشابهًا كبيرًا في المحتوى بين النسختين. بدلاً من ذلك ، استخدام مقياس الاكتئاب هذا في الصين يتطلب في بداية الأمر بدليل صحته.

يمكن أيضًا تناول هذه التوصية بطرق إحصائية بمجرد تكييف الاختبار. على سبيل المثال ، إذا كنا نعتقد أن المجموعات الثقافية تختلف في متغيرات مهمة لا علاقة لها بالمفهوم (بالبنية) الذي يتم قياسه ، فيمكن استخدام الخطط والتحليلات الإحصائية للسيطرة على هذه المتغيرات "الضارة الطفيلية". يمكن استخدام تحليل التغاير، وتصميمات الكتل العشوائية ، والتقنيات الإحصائية الأخرى (تحليل الانحدار ، الارتباط الجزئي ، وما إلى ذلك) للتحكم في تأثيرات مصادر التباين الغير مرغوب فيها بين المجموعات.

اقتراحات (عملية) للممارسة. هذا المبدأ التوجيهي مهم للغاية لأنه يدعو إلى إجراء العديد من التحليلات. فيما يتعلق بتحليل التكافؤ، نقدم الاقتراحات التالية للممارسة:

إذا كان حجم العينة كافيًا ، فقم بإجراء دراسة مقارنة للتكافؤ المفاهيمي لإصدارات المصدر واللغة المستهدفة للاختبار. هناك العديد من حزم البرامج لتسهيل هذه التحليلات (انظر Byrne، 2006).

لتحديد معادلة هيكل الاختبار عبر المجموعات اللغوية و / أو ثقافية قم بإجراء التحليل الاستكشافي (يفضل أن يكون ذلك بالتناوب على بنية الهدف – وهذا ما يسمى "التناوب الهدف") أو تحليل عامل التأكيد ، و / أو تحليل مرجح متعدد الأبعاد. الحاجة إلى عينات كبيرة إلى حد ما (10 مشاركين لكل متغير) يجعل هذه الدراسات صعبة التنفيذ في العديد من الدراسات بين الثقافات. هناك توضيح ممتاز لهذا النوع من الدراسة متاح في (Byrne و van de Vijver 2014).

–ابحث عن دليل صحة التقارب والتمييز (بشكل أساسي ، ابحث عن أدلة ارتباطية بين مجموعة من المفاهيم وتحقق من استقرار هذه العلاقات بين المجموعات اللغوية و / أو الثقافية) (انظر van de Vijver & Tanzer، 1997).

فيما يتعلق بالأداء التفاضلي للبنود ، يتم تقديم بعض الاقتراحات أدناه. بالنسبة للنُهج الأكثر تطوراً ، يتم تشجيع الباحثين على قراءة الأدبيات المهنية حول الأداء التفاضلي للبنود:

قم بإجراء تحليل الأداء التفاضلي للبنود باستخدام أحد الإجراءات القياسية (إذا كانت البنود ثنائية التفرع ، ، فإن إجراء Mantel–Haenszel قد يكون أبسط ؛ وإذا تم تسجيل البنود بطريقة متعددة ، فإن إجراء Mantel–Haenszel المعمم هو خيار ممكن).

هناك حلول أخرى أكثر تعقيدا بما فيها النهج القائم على نظرية الاستجابة للبند. إذا كان حجم العينة أصغر ، يمكن أن يكشف "مخطط دلتا" عن البنود التي يحتمل أن تكون منحازة. تعد المقارنات الشرطية إمكانية أخرى (فيما يخص طرق المقارنة في حال وجود عينات صغيرة ، انظر ، على سبيل المثال ، Muñiz، Hambleton، Xing &، 2001).

ت–3(11) تقديم أدلة لدعم المعايير والموثوقية والصلاحية الخاصة بالنسخة المعدلة للاختبار في المجموعات المستهدفة.

تفسير. لا تنطبق معايير وإثباتات الصحة وإثبات موثوقية الاختبار في صيغته بلغة المصدر تلقائيًا على التعديلات المحتملة الأخرى للاختبار في ثقافات ولغات مختلفة. لذلك ، يجب أيضًا تقديم الصلاحية التجريبية وإثبات موثوقية أي نسخة مطورة جديدة. يجب إدراج جميع أنواع الأدلة التجريبية لدعم الاستدلالات المبنية على (المستمدة من) الاختبار في دليل الاختبار. يجب إيلاء اهتمام خاص للمصادر الخمسة لإثبات الصلاحية بناءً على: محتوى الاختبار ، وعمليات الاستجابة ، والهيكل الداخلي ، والعلاقات مع المتغيرات الأخرى ونتائج الاختبار (AERA، APA، NCME، 2014) . يعد تحليل العوامل الاستكشافية والتأكيدية ، ونمذجة المعادلة الهيكلية ، وتحليل الطرق متعددة المسارات ، من الأساليب الإحصائية التي يمكن استخدامها للحصول على البيانات وتحليلها بشأن صحة الأدلة المستندة إلى البنية الداخلية.

اقتراحات (عملية) للممارسة. الاقتراحات هي نفسها تلك المطلوبة لأي اختبار يتم التفكير في استخدامه:

- إذا كان يُقترح استخدام المعايير التي تم تطويرها للإصدار الأصلي للاختبار مع الإصدار المعدل ، فيجب تقديم دليل على أن هذا الاستخدام مناسب من الناحية الإحصائية وعادل. في حالة عدم تقديم دليل على هذا الاستخدام للمعايير الأصلية ، يجب تطوير معايير محددة للإصدار المعدل وفقًا للمعايير الحالية و الخاصة بتطوير المعايير.

- قم بتجميع أدلة كافية على الموثوقية لتبرير استخدام إصدار الاختبار للغة الهدف. يمكن أن تتضمن الأدلة عادةً تقدير الاتساق الداخلي (على سبيل المثال ، معاملات KR–20 أو معاملات ألفا أو أوميغا).
- قم بتجميع أكبر قدر من إثباتات الصلاحية حسب الضرورة لتحديد ما إذا كان يجب استخدام إصدار الاختبار للغة الهدف. يعتمد نوع الأدلة التي يتم تجميعها على الاستخدام المقصود للنتائج (مثل صلاحية المحتوى لاختبارات التحصيل ، الصلاحية التنبؤية لاختبارات الكفاءة ، إلخ)

ت–4(12) استخدم إجراء التكافؤ و إجراءات تصميم وتحليل البيانات المناسبة لمطابقة النتائج المتحصل عليها في الإصدارات اللغوية المختلفة للاختبار.

تفسير. عندما يتم ربط نسختين بلغتين من الاختبار بمقياس تقرير واحد ، تكون هناك عدة خيارات ممكنة. إذا تم استخدام مجموعة مشتركة من البنود ، فيجب تقييم أداء هذه البنود المشتركة في المجموعتين اللغويتين ، وإذا لوحظ الأداء التفاضلي ، فيجب النظر في إزالتها من البيانات المستخدمة لإنشاء الرابط. تخدم الرسوم البيانية دلتا (Angoff و Modu، 1973) هذا الغرض جيدًا ، وقد أوضح (Cook and Schmitt–Cascallar،2005) كيفية استخدامها لتحديد البنود التي لها معاني مختلفة لمجموعتي الاشخاص المختبرين. ليس لكل أنواع البنود نفس الإمكانية للربط بين الاصدارات اللغوية. يمكن تخطيط تقديرات صعوبة البند والتمييز للمعلمات المشتقة في إطار نظرية الاجابة على البند على رسم تخطيطي للمساعدة في تحديد البنود الشائعة التي تؤدي أداءً سيئًا (انظر Hambleton، Swaminathan، Rogers &، 1991).

لكن إنشاء روابط (أي "التكافؤ") بين نتائج الإصدارين اللغويين للاختبار يعتبر دائمًا مشكلة بسبب وجود افتراضات قوية حول البيانات. في بعض الأحيان يفترض بجرأة أن الإصدارات المختلفة من الاختبار متكافئة، وبالتالي يمكن استخدام درجات النسختين من الاختبار بالتبادل. يمكن أن يكون هذا الافتراض صحيحًا في حالة اختبارات الرياضيات، لأن الترجمة / التكييف بسيطة عمومًا. لذلك يمكن انصافه إذا تم إنشاء الإصدارين من الاختبار بعناية ونستطيع ان نفترض أن إصدار اللغة المصدر للاختبار يعمل مع سكان اللغة المصدر بطريقة تعادل تلك الخاصة بإصدار اللغة الهدف للاختبار في مجتمع اللغة الهدف. يمكن تبرير هذه الفرضية إذا كانت جميع الدلائل الأخرى متاحة وتشير إلى أن النسختين من لغتي الاختبار متكافئتين وأنه لا توجد تحيزات منهجية تؤثر على النتائج في إصدار اللغة الهدف من الاختبار.

يوجد حلان آخران ، لكن لا يوجد حل مثالي. أولاً ، إقران النسختين باستخدام نموذج فرعي للبنود التي تعتبر متعادلة أساسًا في نسختي الاختبار اللغوي. على سبيل المثال ، قد تكون البنود هي تلك التي كانت سهلة الترجمة أو التكييف. من حيث المبدأ، يمكن أن يعمل الحل، لكن هذا يتطلب أن تكون البنود المتكافئة و البنود الأخرى في الاختبار تسمح بقياس نفس البنية. يتمثل الحل الثاني في إنشاء روابط تكافؤ من خلال عينة من المرشحين يتقنون اللغتين. مع أخذ هذه العينة للإجابة على كلا إصداري الاختبار، سيكون من الممكن إنشاء جدول تحويل النتائج. لا يجب أن تكون العينة صغيرة جدًا، ويجب أن يكون ترتيب عرض أشكال الاختبار متوازناً في التصميم. الافتراض العام لهذا النهج هو أن المرشحين يتحدثون لغتين بالفعل، وبالتالي، بالإضافة إلى الصعوبات النسبية للنماذج، ينبغي للمرشحين أن يحسنوا الأداء في كلا النموذجين. يتم استخدام أي اختلاف لضبط الدرجات عند تحويل النتائج من إصدار اختبار إلى آخر.

اقتراحات للممارسة. سيكون إنشاء روابط بين نتائج النسخ المكيفة من الاختبار مشكلة في أفضل الحالات ، لأن جميع نماذج المعادلات بها على الأقل فجوة رئيسية واحدة. قد تكون أفضل استراتيجية هي اتباع جميع الخطوات لتحديد درجة معادلة الصف بالكامل. إذا كانت الأدلة المتعلقة بالأسئلة الثلاثة أدناه مثنة ، يمكن معالجة نتائج الإصدارين من الاختبار بالتبادل:

- هل هناك دليل على أن نفس بناء المفهوم يقاس فعليًا في إصدارات اللغة المصدر و اللغة الهدف من الاختبار؟ هل بناء المفهوم له نفس العلاقة مع المتغيرات الخارجية الأخرى في الثقافة الجديدة؟
- هل هناك أدلة قوية على أن مصادر التحيز المنهجي قد تم القضاء عليها (على سبيل المثال ، لا توجد مشاكل في الوقت ، كما أن أشكال الاختبار مألوفة أيضًا للمرشحين ، ولا تشوش في التعليمات ، ولا تحريفات منهجية في مجموعة أو أخرى ، الإرشادات الموحدة ، عدم وجود أنماط استجابة (تقييمات متطرفة ، دوافع مختلفة ...)؟
- هل الاختبار خالي من بنود محتملة التحيز؟ هنا ، يمكن استخدام تخطيط القيم p (أي نسبة الإجابات الصحيحة على البنود) أو من الأفضل قيم دلتا ، من بنود إصداري الاختبار. يجب دراسة النقاط التي لا تقع على طول خط المعادلة الخطية لتحديد ما إذا كانت البنود المرتبطة مناسبة في كلتا اللغتين. توفر تحليلات الوظيفة التفاضلية للبنودFDI دليلاً أقوى على تكافؤ البنود بين المجموعات اللغوية والثقافية.

إذا اجريت محاولات لربط النتائج المستأصلة من إصدارات مختلفة من الاختبار ، فيجب اختيار نموذج الربط المناسب وتطبيقه. يجب تقديم أدلة عن صحة هذه العملية ...

توصيات متعلقة بإدارة الاختبار

إ-1(13) **تحضير الأدوات وتعليمات إدارة الاختبار بغرض التقليل من أي مشكلة لها علاقة بالثقافة واللغة تتسبب فيها إجراءات إدارة الاختبار وأنماط الاستجابة التي قد تؤثر على صلاحية التفسيرات المستمدة من النتائج.**

تفسير. يجب أن يبدأ تنفيذ توصيات إدارة **الاختبار** بتحليل جميع العوامل التي قد تهدد صحة نتائج الاختبار في سياق ثقافي ولغوي محدد. يمكن أن تكون الخبرة في إدارة **الاختبار** في سياق أحادي اللغة أو أحادي الثقافة مفيدة بالفعل في توقع المشكلات التي يمكن انتظارها في سياق متعدد اللغات أو متعدد الثقافات. على سبيل المثال، غالبًا ما يعرف الأشخاص ذوو الخبرة في إدارة الاختبارات ما هي جوانب التعليمات التي قد تكون صعبة على المستجيبين. يمكن أن تظل هذه الجوانب صعبة بعد الترجمة أو التكيف. إن تطبيق الأدوات في سياق لغوي أو ثقافي جديد يمكن أن يطرح أيضًا مشاكل لم تكن موجودة من قبل في التطبيقات أحادية الثقافة.

اقتراحات للممارسة. من المهم ، كجزء من هذا المبدأ التوجيهي ، توقع العوامل المحتملة التي يمكن أن تخلق مشاكل في إدارة الاختبارات. فيما يلي بعض العوامل التي يجب مراعاتها لضمان النزاهة في إدارة الاختبارات:

وضوح تعليمات الاختبار (بما في ذلك نسختها المترجمة) ، وآلية الاستجابة (على سبيل المثال ، ورقة الإجابات) ، و المدة المخصصة (مصدر الخطأ الشائع يخص عدم إعطاء الوقت الكافي للمرشحين لإكمال الاختبار) ، وتحفيز المرشحين لإكمال الاختبار ، ومعرفة الغرض من الاختبار وكيف سيتم تسجيله.

إ-2(14) **حدد شروط الاختبار التي يجب أن تكون متطابقة لدى أفراد المجتمع محل الاهتمام.**

تفسير. الغرض من هذه التوصيات هو تشجيع مصممي الاختبار على وضع تعليمات الاختبار والإجراءات ذات الصلة (مثل ظروف الاختبار والأوقات وما إلى ذلك) التي يمكن متابعتها عن كثب في جميع فئات السكان محل الاهتمام. الغرض الرئيسي من هذا المبدأ التوجيهي هو تشجيع أولئك الذين يديرون الاختبارات على الالتزام بتعليمات موحدة. في نفس الوقت، يمكن توضيح بعض التعديلات للاستجابة لاحتياجات مجموعات فرعية معينة من الأفراد داخل كل مجتمع ،مثل الوقت الإضافي؛ المستند

المطبوع بحجم أكبر ، و ظروف إدارة اختبار أكثر هدوءًا ، إلخ. في مجال الاختبارات ، هذه التدابير معروفة اليوم تحت اسم " تكيفات إجراء الاختبار". ليس الغرض من هذه التسهيلات تضخيم نتائج المرشحين ، بل تهدف لتهيئة بيئة امتحان محددة لهؤلاء المرشحين تسمح لهم بالتعبير عن ما يشعرون به ، أو ما يعرفونه وما يمكنهم القيام به.

ينبغي ملاحظة الاختلافات في ظروف الاختبار القياسية بحيث يمكن في وقت لاحق من هذه العملية مراعاة هذه الاختلافات وتأثيرها على التعميمات والتفسيرات.

اقتراحات للممارسة. قد يتداخل هذا المبدأ التوجيهي جزئيًا مع المبدأ التوجيهي إ-1(13) ، ولكن تم إعادة صياغتها هنا للتأكيد على أهمية أن يتقدم المرشحون للامتحان في ظروف مماثلة قدر الإمكان. يعد هذا ضروريًا إذا كانت نتائج الإصدارين باللغتين تُستخدم بالتبادل. إليك بعض الاقتراحات:

- ينبغي تكييف تعليمات الامتحانات والإجراءات ذات الصلة وإعادة كتابتها بطريقة موحدة ، التي تتناسب مع اللغة والثقافة الجديدة.
- إذا كانت تعليمات الاختبار والإجراءات ذات الصلة مناسبة للثقافات الجديدة ، فيجب أن يتلقى المسؤولون عن إدارة الاختبار تدريباً على الإجراءات الجديدة ؛ و يجب إبلاغهم بالالتزام و احترام هذه الإجراءات و ليس الإجراءات الأصلية.

توصيات لتسجيل الدرجات وتفسيرها (ت ل ت)

ت ل ت -1(15). تفسير كل اختلاف في الدرجات(العلامات/القيم) بين المجموعات مع الاخذ بعين الاعتبار كل المعلومات المتاحة ذات الصلة بالموضوع.

تفسير. حتى إذا تم تكييف الاختبار باستخدام إجراءات سليمة تقنيًا ، وتم بالفعل إثبات صحة الدرجات إلى حد ما، يجب تذكر أن معنى الاختلافات بين المجموعات يمكن تفسيره بعدة طرق بما في ذلك: بسبب الاختلافات الثقافية أو غيرها بين البلدان و / أو الثقافات المعنية. في مقال له ، استعرض Sireci (2005)طريقة تقييم تكافؤ نسختين لغويتين مختلفتين للاختبار ، من خلال عرض النسخ اللغوية المختلفة للاختبار على مجموعة من المستجيبين يجيدون اللغتين (ثنائي اللغة) ومن نفس المجموعة الثقافية أو اللغوية. و في ان واحد قام بوصف بعض خيارات بروتوكول البحث لدراسات التكافؤ باستخدام مستجيبين ثنائي اللغة ، ووضع قائمة من المتغيرات الطفيلية للتحكم فيها ، وقدم اقتراحات قيمة لتفسير النتائج.

اقتراحات للممارسة. يوجد أدناه اقتراح لتحسين الممارسة.

حسب موضوع البحث (أو السياق الذي يتم من أجله إجراء مقارنات بين المجموعات) ، يجب النظر في عدة تفسيرات محتملة ، قبل اختيار أحدها على وجه الخصوص. على سبيل المثال ، من المهم مراعاة الاختلافات في الدافعية للنجاح في الاختبار قبل استنتاج أن أداء مجموعة أفضل من الأخرى. قد تكون هناك أيضًا تأثيرات ذات معنى للسياق على أداء الاختبار. على سبيل المثال، قد تكون مجموعة من الأشخاص ببساطة جزءًا من نظام تعليمي أقل فعالية، مما قد يكون له تأثير ذا معنى على أداء الاختبار.

ت ل ت–2(16)لا يتم مقارنة الدرجات بين المجموعات إلا عندما يكون مستوى الثبات(عدم التغير) قد تم تحديده على المقياس الذي يتم تسجيل الدرجات عليه.

تفسير. عندما تكون دراسات المقارنة بين المجموعات اللغوية والثقافية في صميم مبادرة الترجمة والتكييف ، يجب وضع نسخ متعددة اللغات من الاختبار على مقياس مشترك ، ويتم ذلك من خلال عملية تسمى "التوأمة ". "أو" المطابقة ". ويتطلب ذلك عينات كبيرة الحجم وأدلة تبين أن النسخة المعدلة من الاختبار لا تحتوي على تحيز في البناء ، أو تحيز في الأسلوب ، أو تحيز في البنود .

حدد Van de Vijver and Poortinga (2005) عدة مستويات من تكافؤ الاختبارات بين المجموعات اللغوية والثقافية ، وكان عملهما مفيدًا بشكل خاص لفهم هذا المفهوم الذي يشتركان في تأليفه. على سبيل المثال ، أشاروا إلى أن تكافؤ وحدات القياس / التقييم يتطلب أن تكون مقاييس التنقيط لكل مجموعة بنفس النظام المتري ، مما يضمن ان الاختلافات بين الأفراد داخل المجموعات لها نفس المعنى. (على سبيل المثال، يمكن مقارنة الفروق بين الرجال والنساء في عينة صينية بتلك الموجودة في عينة فرنسية). ومع ذلك، لا يمكن إجراء مقارنات مباشرة صحيحة للدرجات إلا عندما يكون للنتائج مستوى عالي من التكافؤ، يسمى التكافؤ العددي أو التكافؤ الكامل للنتائج ، مما يتطلب أن يكون لمقاييس كل مجموعة نفس وحدة القياس/ التقييم ونفس الأصل من مجموعة إلى أخرى.

تم اقتراح العديد من الطرق (سواء في إطار النظرية الكلاسيكية للاختبارات أو تلك الخاصة بنظرية الاستجابة للبند) لمطابقة(تكافؤ) أو توأمة درجات مجموعتين (أو الإصدارات اللغوية للاختبار). يمكن للقراء المهتمين الرجوع إلى Angoff (1984) و Kolen and Brennan (2004) لفهم هذا الموضوع بشكل أفضل. وقد اقترح Cook and Schmitt–Cascallar (2005) قاعدة لفهم الأساليب الإحصائية المتوفرة حاليًا لمعادلة و توسيع استخدام الاختبارات التربوية والنفسية . وصف و انتقد المؤلفون إجراءات مطابقة المقاييس المستخدمة في دراسات تكييف الاختبارات. كما وضحو بعض الإجراءات والمشكلات المتعلقة بمطابقة المقاييس من خلال وصف وانتقاد ثلاث دراسات تم إجراؤها على مدار العشرين عامًا

الماضية من أجل معادلة درجات اختبار التقييم المدرسي في نسخته الإسبانية ، Prueba te Aptitude Academica.

قتراحات للممارسة. النقطة الحاسمة هنا هي أنه لا ينبغي المبالغة في تفسير درجات الاختبار:

- تفسير الدرجات وفقًا لمستوى الصدق(صحيح) المتاح. على سبيل المثال ، لا تقم بعمل نتائج مقارنة حول مستويات أداء المستجيبين في مجموعتين لغويتين ، ما لم يتم التحقق من الثبات المتري(قياس الاختلاف) للدرجات في الاختبارات المقارنة ...

توصيات خاصة بالتوثيق

وثيقة–1 (17). تقديم الوثائق التقنية لأي تغييرات ، بما في ذلك الأدلة التي تم الوصول إليها للدفاع عن التكافؤ، عندما يتم تكييف اختبار للاستخدام في مجتمع آخر.

تفسير. تم وضع هذه التوصية والتأكيد علي اهميتها من قبل العديد من الباحثين (انظر ، على سبيل المثال ، Grisay، .(2003 ونجحت للغاية كل من دراسة TIMSS ودراسة PISA في احترام هذه التوصية من خلال توثيق التغييرات بعناية خلال أعمال التكييف. بالرجوع إلى هذه المعلومات، يمكن تقييم مدى ملاءمة التغييرات التي تم إجراؤها.

كما يجب أن تحتوي الوثائق التقنية على تفاصيل كافية عن المنهجية حتى يتمكن الباحثون المستقبليون من تكرار الإجراءات المستخدمة على نفس المجتمع أو على مجتمعات اخرى. يجب أن تحتوي على معلومات كافية عن أدلة تكافؤ البناء وتكافؤ العينات التحجيم(إذا تحقق) لتبرير استخدام الأداة في المجتمع الجديد.

عند اجراء مقارنات بين المجتمعات، يجب أن يتم توثيق الأدلة التي تم الاستناد إليها لتحديد تكافؤ(معادلة) الدرجات بين المجتمعات.

أحيانا يطرح السؤال حول إلى من يوجه التوثيق التقني. يجب كتابة الوثائق لفائدة الخبراء ولأولئك الذين سيقومون بتقييم فائدة استخدام الاختبار في المجتمع الجديد أو في مجتمعات أخرى. (يمكن إضافة وثيقة مبسطة لفائدة غير الخبراء)

اقتراحات للممارسة

- يجب أن تكون الاختبارات المعدلة مصحوبة بدليل تقني يوثق جميع الأدلة الكيفية والكمية المحيطة بعملية التكييف. من المفيد بشكل خاص توثيق أي تغييرات تم إجراؤها لملاءمة الاختبار مع لغة وثقافة أخرى. بشكل أساسي ، يرغب الخبراء ومحررو المجلات في الحصول على مراجع

حول الاجراءات التي تم اتباعها لإعداد اصدار الاختبار في اللغة المستهدفة والتحقق من صحته. كما يرغبون أيضًا في رؤية نتائج أي تحليلات تم إجراؤها. فيما يلي أنواع الأسئلة التي يتم طرحها:

- ما هي الأدلة المتاحة لدعم فائدة الاختبار المكيف للمجتمع الجديد؟
- ما هي البيانات التي تم جمعها عن البنود ومن أي عينات؟
- ما هي البيانات الأخرى التي تم الحصول عليها لتقييم صحة المحتوى والمعايير والبناء؟
- كيف تم تحليل البيانات المختلفة؟
- ما هي النتائج التي تم التوصل إليها؟

وثيقة-2 (18). تقديم وثائق لمستخدمي الاختبار من شأنها أن تعزز التطبيق الجيد للاختبار المكيف مع افراد المجتمع الجديد.

تفسير. يجب كتابة الوثائق لأولئك الذين سيستخدمون الاختبار في اطار تقييم تطبيقي. يجب أن يتوافق مع الممارسات الجيدة المحددة في التوصيات الخاصة باستخدام الاختبارات الصادرة عن لجنة الاختبارات الدولية انظر (www.InTestCom.org).

اقتراحات للممارسة.

يجب على مؤلف الاختبار تقديم معلومات محددة حول كيفية تأثير السياقات الاجتماعية والثقافية والبيئية للسكان على أداء الاختبار. يجب أن يحتوى دليل المستخدم على ما يلي:

- وصف البنية (التركيبات) التي يتم قياسها بواسطة الاختبار بالإضافة إلى إجراء التكييف.
- ملخص بالأدلة الداعمة للتكييف ، بما في ذلك الأدلة على المزايا الثقافية، محتوى البنود ، ومدى ملاءمة تعليمات الاختبار ، وطريقة الاجابة ، إلخ.
- تحديد أساس استخدام الاختبار مع مجموعات فرعية مختلفة من المجتمع وأي قيود أخرى في الاستخدام.
- شرح المشاكل / الرهانات التي يجب أن تؤخذ بعين الاعتبار فيما يتعلق بالممارسة الجيدة في تطبيق الاختبارات.
- تحديد إذا أمكن ذلك إجراء مقارنات بين المجتمعات، عند الضرورة، شرح كيفية القيام بذلك.

- توفير المعلومات المطلوبة للتسجيل والمعايرة (على سبيل المثال ، جداول البحث للمعايير ذات الصلة) أو وصف كيف يمكن للمستخدمين الوصول إلى إجراءات التصنيف (على سبيل المثال ، عندما تكون محوسبة).
- تقديم توصيات لتفسير الدرجاتz ، بما في ذلك معلومات عن بيانات الثبات(الثقة) والصدق(الصلاحية) عن التفسيرات التي يمكن استخلاصها من درجات الاختبارات.

كلمة ختامية

لقد بذلنا قصارى جهدنا للتوصل إلى مجموعة من التوصيات لمساعدة المؤلفين / المبدعين والمستخدمين للاختبارات في عملهم. ومع ذلك ، لكي تكون هذه الجهود التي تهدف الى تغيير الممارسات السيئة فعالة ، يجب وضع آليات جيدة لنشرها. أظهرت دراسة حديثة أجراها Rios و Sireci (2014) أن غالبية مشاريع تكييف الاختبارات في الأدبيات المنشورة لم تتبع توصيات اللجنة الدولية للاختبارات (ITC) المتاحة منذ حوالي 20 عامًا حتى الآن. لذلك نشجع القراء على بذل كل ما في وسعهم لتحسيس زملائهم حول هذه الطبعة الثانية كمصدر أساسي لأفضل الممارسات التي ساهم فيها العديد من المختصين من جميع أنحاء العالم.

في نفس الوقت ، نحن ندرك أنه كما تم استبدال الإصدار الأول من هذه التوصيات ، سيتم استبدال الإصدار الثاني أيضًا. المعايير المعروفة للتقييم في التربية وعلم النفس من AERA و APA و NCME هي الآن في طبعتها السادسة (AERA، APA & NCME، 2014). نتوقع أن تخضع توصيات اللجنة الدولية للاختبارات (ITC) الخاصة بتكييف الاختبارات لمزيد من التنقيحات في السنوات القادمة. إذا كنتم تعرفون أي دراسات جديدة يجب الاستشهاد بها ، أو قد تؤثر على الإصدار الثالث ، أو إذا كنتم ترغبون في اقتراح توصيات أو تنقيحات جديدة للتوصيات الثمانية عشر المعروضة هنا ، يرجى إخبار اللجنة الدولية للاختبارات (ITC). يمكنكم الاتصال بالرئيس الحالي للجنة البحوث والتوصيات التي أنتجت الإصدار الثاني و / أو سكرتير مركز التجارة الدولية على عنوان البريد الإلكتروني التالي 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 مرجعية لتوصيات اللجنة الدولية للاختبارات (ITC) لترجمة وتكييف الاختبارات.

فيما يلي قائمة مرجعية للتذكير بالتوصيات الثمانية عشر للجنة الدولية للاختبارات (ITC). ندعوكم للتحقق من تلك التي تعاملتم معها بشكل مُرضٍ في مشروعكم لترجمة / تكييف الاختبارات ، ثم التعامل مع تلك التي لا تزال دون إجابة.

توصيات خاصة بالشروط الأولية (ش أ)

ش أ -1(1) الحصول إلزاما على إذن من صاحب حقوق الملكية الفكرية الخاصة بالاختبار قبل البدء في تكييفه

ش أ-2(2) التقرب من العينة المستهدفة لتقييم درجة اللياقة / التوافق بين التعريف ومحتوى الهيكل المقاسة بواسطة الاختبار الأصلي وأن يكون كل بند كافيا للاستخدام المقصود (أو الاستخدامات المقصودة) لنتائج الاختبار.

ش أ-3(3) الحد بشكل كبير من تأثير الاختلافات الثقافية واللغوية الضارة / غير المرغوب فيها / غير الضرورية في استخدام قصدي للاختبار في العينات المستهدفة.

توصيات لإعداد الاختبارات (إ)

إ ا-1(4) التأكد من أن إجراءات الترجمة والتكييف تأخذ بعين الاعتبار الاختلافات اللغوية والنفسية والثقافية للعينات المستهدفة من خلال اختيار الخبراء ذوي الخبرة اللازمة.

إ ا-2(5) استخدم تصميمات وإجراءات ترجمة مناسبة لزيادة ملائمة تكييف الاختبار مع المجموعات المستهدفة.

إ ا-3(6) تقديم ادلة بأن تعليمات الاختبار ومحتوى البنود لها نفس المعنى لجميع المجموعات المستهدفة.

إ ا-4(7) اثبات ان اشكال البنود، مقاييس الإجابة، التصنيف، وفئات التصنيف، واتفاقيات الاختبار وطرق ادارته، وأي إجراء آخر مناسبة لجميع الفئات المستهدفة.

إ ا-5(8) جمع البيانات الخاصة بالدراسات التجريبية للاختبار المكيف من أجل إجراء تحليل البنود وتقييم وثوق وصلاحية القيمة التي تسمح بالمراجعات اللازمة.

توصيات التحقق من الصحة / التأكد

ت –1(9) اختيار عينة تكون خصائصها ذات صلة بالاستخدام المقصود للاختبار و حجمها وأهميتها كافية للتحليل التجريبي.

ت–2 (10) تقديم بيانات إحصائية مقبولة /موثوقة تخص معادلات البناء، المناهج، و البنود عبر المجتمعات المستهدفة

ت–3(11) تقديم أدلة لدعم المعايير والموثوقية والصلاحية الخاصة بالنسخة المعدلة للاختبار في المجموعات المستهدفة.

ت–4(12) استخدم إجراء التكافؤ و إجراءات تصميم وتحليل البيانات المناسبة لمطابقة النتائج المتحصل عليها في الإصدارات اللغوية المختلفة للاختبار.

التوصيات المتعلقة بإدارة الاختبار

إ–1(13) تحضير الأدوات وتعليمات إدارة الاختبار بغرض التقليل من أي مشكلة لها علاقة بالثقافة واللغة تتسبب فيها إجراءات إدارة الاختبار وأنماط الاستجابة التي قد تؤثر على صلاحية التفسيرات المستمدة من النتائج.

إ–2(14) حدد شروط الاختبار التي يجب أن تكون متطابقة لدى جميع أفراد المجتمع محل الاهتمام.

توصيات لتسجيل الدرجات وتفسيرها

ت ل ت –1(15). تفسير كل اختلاف في الدرجات(العلامات/القيم) بين المجموعات مع الاخذ بعين الاعتبار كل المعلومات المتاحة ذات الصلة بالموضوع.

ت ل ت–2(16)لا يتم مقارنة الدرجات بين المجموعات إلا عندما يكون مستوى الثبات(لزوم) قد تم تحديده على المقياس الذي يتم تسجيل الدرجات عليه.

توصيات خاصة بالتوثيق

وثيقة–1 (17). تقديم الوثائق التقنية لأي تغييرات، بما في ذلك الأدلة التي تم الوصول إليها للدفاع عن التكافؤ، عندما يتم تكييف اختبار للاستخدام في مجتمع آخر.

وثيقة–2 (18). تقديم وثائق لمستخدمي الاختبار من شأنها أن تعزز التطبيق الجيد للاختبار المكيف مع افراد المجتمع الجديد.

الملحق ب. مسرد المصطلحات

ألفا (أو تسمى أحيانًا "معامل ألفا" أو "ألفا كرو نباخ").

معامل الموثوقية للاختبار الذي من المفترض أن تقوم عناصره بقياس صفة واحدة مشتركة ولديها نفس الإمكانيات للتمييز (وبالتالي فهي حالة خاصة لأوميغا–انظر أدناه). ويكون هذا هو الحد الأدنى لمتطلبات الموثوقية في الحالات الأكثر عمومية.

أوميغا (أو تسمى أحيانًا "معامل أوميغا" أو "ماكدونالدز أو ميغا")

معامل الثبات للاختبار الذي من المفترض أن تقيس بنوده صفة واحدة مشتركة (تتلاءم مع نموذج العامل العام). هو أكثر قابلية للتطبيق من معامل ألفا.

اختبار الأبعاد

هو عدد الأبعاد أو العوامل التي يقيسها الاختبار. غالبًا ما يتم هذا التحليل إحصائيًا باستخدام واحد من العديد من الإجراءات، بما في ذلك مخططات القيم الذاتية أو نمذجة المعادلة الهيكلية.

إصدار اللغة المصدر

اللغة التي يكتب بها الاختبار في الأصل.

إصدار اللغة الهدف

اللغة التي يتم بها ترجمة / تكييف الاختبار. على سبيل المثال، إذا تمت ترجمة اختبار من الإنجليزية إلى الإسبانية، فغالبًا ما تسمى النسخة الإنجليزية "إصدار اللغة المصدر" وتسمى النسخة الإسبانية "إصدار اللغة الهدف."

«بيسا"(PISA)

اختصار لـ " البرنامج الدولي لتقييم التلاميذ "وهو تقييم دولي يرتكز على ترتيب الدول حسب الأداء المدرسي. هذا البرنامج ترعاه المنظمة الدولية للتعاون والتنمية الاقتصادية (OCDE) مع مشاركة أكثر من 40 دولة.

تحليل عوامل المطابقة (CFA)

يتم مسبقًا صياغة فرضية حول بنية الاختبار، ثم يتم إجراء تحليلات لتقييم هذا الهيكل من مصفوفة الارتباط لعناصر الاختبار. فيتم إجراء اختبار إحصائي لمعرفة ما إذا كان الهيكل الافتراضي والهيكل

المقدر للاختبار متقاربان بما فيه الكفاية بحيث لا يمكن رفض الفرضية الفارغة التي بموجبها يتساوى الهيكلان.

ترجمة أحادية الاتجاه / إلى الأمام / مباشرة / استباقية

مع هذا البروتوكول، يتم تكييف الاختبار في اللغة الهدف ذات الاهتمام من قبل مترجم، أو في كثير من الأحيان، مجموعة من المترجمين، ثم يحكم مترجم أو مجموعة أخرى من المترجمين على معادلة إصدارات الاختبار في اللغة المصدر واللغة الهدف.

الترجمة المزدوجة والتدقيق المتقاطع

في بروتوكول الترجمة هذا، يقوم مترجم مستقل أو مجموعة من الخبراء بتحديد وحل التناقضات بين مختلف الترجمات (منها أحادية الاتجاه / الأمامية / الاستباقية / المباشرة)، والتوفيق / التداخل بينها لتصميم الإصدار الواحد.

تصميم ترجمات ثنائية الاتجاه

باستخدام هذا التصميم، يترجم الاختبار من إصدار اللغة المصدر إلى إصدار اللغة الهدف بواسطة مجموعة من المترجمين، ومن ثم يتم إرجاع إصدار اللغة الهدف مرة أخرى إلى اللغة المصدر، وذلك بواسطة مترجم ثان أو مجموعة من المترجمين. تتم مقارنة الإصدارات أصلية المصدر والإصدارات المعاد ترجمتها، ويتم إصدار قرار بشأن مدى تلاؤم إصدار لغة المصدر للاختبار. إذا كانت نسختا اللغة المصدر في منتهى التقارب، فمن المفترض أن يكون إصدار اللغة الهدف من الاختبار مقبولا.

صيغة Kuder–Richardson 20 (في بعض الأحيان تسمى "KR–20.")

معامل الثبات للاختبار المتكون من بنود ثنائية، التي من المفترض أن تقيس صفة مشتركة ولها تمييز على نفس قدم المساواة.

قيم/ مؤشرات دلتا

قيم / مؤشرات دلتا هي ببساطة قيم p التي تم تحويلها بطريقة غير خطية وتم تطبيقها على البنود ذات التصنيف الثنائي. قيمة دلتا للبند هي الانحراف المعتدل المقابل للمنطقة الواقعة تحت التوزيع المعتدل (متوسط = 0.0، الانحراف المعياري=1.0) حيث تساوي المساحة الواقعة تحت التوزيع المعتدل نسبة المرشحين الذين يجيبون بشكل صحيح على البند. و بالتالي إذا كانت p = 0.84 فقيمة دلتا للبند ستكون 1.0. ويتم اعتماد هذا التحويل في ظل افتراض أن قيم / مؤشرات دلتا ستكون على الأرجح على مقياس فاصل مساوي لقيم p .

معادلة نتائج/ علامات/ قيم الاختبار

إجراء إحصائي بصدده تزدوج نتائج اختبارين إثنين يقيسان نفس البنية دون ان يكون هذان الاختباران متوازيان تمامًا.

نظرية الرد على البند(IRT)

فئة من النماذج الإحصائية لربط استجابات البنود بصفة ما أو بمجموعة من الصفات التي يتم قياسها بواسطة بنود الاختبار. بإمكان نماذج محددة ل IRT معالجة بيانات الاستجابة الثنائية ومتعددة الحلقات (polytomous). قد تكون البيانات الثنائية نتيجة تسجيل بنود متعددة الخيارات أو نتيجة البنود الحقيقية الخاطئة لمقياس الشخصية. أما البيانات المتعددة الإجابة (polytomous) فقد تكون نتيجة تسجيل النتائج من خلال مهام الأداء أو من خلال المهام العملية لاختبار الأداء أو من خلال مقاييس التقييم مثل "Likert"

إجراء Mantel–Haenszel (MH) لتحديد الوظيفة التفاضلية للعناصر(DIF)

إجراء إحصائي لمقارنة أداء مجموعتان من المستطلعين على بند اختبار. يتم إجراء مقارنات للمستطلعين من كل مجموعة والذين يتم مطابقتهم على الصفة أو البنية المقاسة بواسطة الاختبار.

الانحدار المنطقي (LR) لتحديد الوظيفة التفاضلية للبنود.(DIF)

هذا الإجراء الإحصائي هو طريقة إضافية لإجراء تحليلات DIF. يتم تعديل وضبط المنحنى المنطقي مع معطيات وبيانات الأداء الخاصة بكل مجموعة، ثم تتم مقارنة إحصائية لاثنين من منحنى منطقي، واحد لكل مجموعة لغوية.

التحليل العاملي الاستكشافي (AFE)

تحليل العوامل هو إجراء إحصائي يتم تطبيقه، على سبيل المثال، مع مصفوفة الارتباط المنبثقة من الارتباط البيني بين مجموعة من البنود متواجدة في الاختبار (أو مجموعة من الاختبارات). الهدف من ذلك هو محاولة شرح الارتباط البيني بين بنود الاختبار (أو الاختبارات) اعتمادا على عدد صغير من العوامل التي يُعتقد أنها تقاس بواسطة الاختبار (أو الاختبارات). على سبيل المثال، في حالة اختبار في الرياضيات، يمكن أن يحدد تحليل العوامل تواجد ثلاث مجموعات/ فئات من العناصر في الاختبار ذاته: مجموعات من بنود الحساب ومجموعات من بنود المفاهيم ومجموعات من بنود حل المشاكل. يمكن إذن القول إن اختبار الرياضيات يقيس ثلاثة عوامل وهي عامل الحساب وعامل مفاهيم الرياضيات، وعامل حل المشاكل في مجال الرياضيات.

التطور المتزامن للاختبارات

تطوير في آن واحد لاستبيانات باللغة المصدر وباللغة الهدف، وذلك باستخدام إجراءات موحدة لمراقبة جودة الترجمة. يعتمد التطوير المتزامن بشكل متزايد في المشاريع الدولية الواسعة النطاق، وذلك لتفادي استحالة أو صعوبة ترجمة أو تكييف النسخة المطورة بلغة واحدة إلى جميع لغات الدراسة.

المستطلعين

مصطلح يستخدم بالتبادل في مجال الاختبارات مع "المختبرين" و "المرشحين" و "المجيبين" و "الطلبة" (إذا كان مجال الاختبارات هو النجاح الأكاديمي).

الموقع

مصطلح شائع في مجال الاختبار. يستخدم لوصف عملية إجراء اختبار معد بلغة وثقافة قابل للاستخدام في لغة أخرى. المصطلحات المطابقة هي الترجمة /أو التكييف.

النمذجة بالمعادلات الهيكلية

مجموعة من النماذج الإحصائية المعقدة التي يتم استخدامها لتحديد البنية الأساسية للاختبار أو مجموعة من الاختبارات. غالبًا ما تستخدم هذه النماذج لدراسة الاستدلالات السببية حول العلاقات بين مجموعة من المتغيرات.

الوظيفة التفاضلية للبنود (DIF)

هي مجموعة من الإجراءات الإحصائية التي يمكنها تحديد ما إذا كان البند يتماشى بشكل أو بآخر بنفس الطريقة مع مجموعتين مختلفتين. تتم مقارنات الأداء أولا بمطابقة المستطلعين على الصفة التي يقيسها الاختبار. عند ملاحظة فروقات، يُقال إن البند قد يكون متحيزًا. ينبغي في هذه الحالة شرح الاختلافات الشرطية المنبثقة من أداء مستطلعي المجموعتين المطابقتين للصفة المقاسة حسب البند.

TIMSS

اختصار ل "الاتجاهات في الدراسات الدولية في الرياضيات والعلوم"، وهو تقييم دولي للتلاميذ في مجال الرياضيات والعلوم في الصفوف 4 و8 و12 (أي ما يعادل المستوى الابتدائي والإعدادي والثانوي) في مختلف البلدان، وبرعاية الرابطة الدولية لتقييم التحصيل التربوي.(IEA).

WDMS

اختصار ل «التحجيم متعدد الأبعاد الموزون". إنه إجراء إحصائي آخر لمعالجة أبعاد الاختبار.