INTERNATIONAL TEST COMMISSION

Diretrizes do ITC para Tradução e Adaptação de Testes

(Segunda Edição)

Versão 2.4

Por favor, use a seguinte referência para este documento:

International Test Commission. (2017). *The ITC Guidelines for Translating and Adapting Testes (Second edition).* https://www.intestcom.org/. Translation authorized by Instituto Brasileiro de Avaliação Psicológica (IBAP).

Tradução: Solange Muglia Wechsler, Larissa A. Alexandrino de Azevedo Porto, Jéssica Particelli Gobbo, Renan de Morais Afonso, Cristina Maciel Massens. Curso de pós-graduação em Psicologia. Pontifícia Universidade Católica de Campinas. São Paulo, Brasil.

AGRADECIMENTOS

O Conselho da Comissão Internacional de Testes deseja agradecer ao comitê de seis pessoas que trabalhou durante vários anos para produzir a segunda edição das Diretrizes para Traduzir e Adaptar Testes: David Bartram, SHL, Reino Unido; Giray Berberoglu, Universidade Técnica do Oriente Médio, Turquia; Jacques Grégoire, Universidade Católica de Lovaina, Bélgica; Ronald Hambleton, Presidente do Comitê, Universidade de Massachusetts Amherst, EUA; José Muniz, Universidade de Oviedo, Espanha; e Fons van de Vijver, Universidade de Tilburg, Holanda.

Além disso, a Comissão Internacional de Testes deseja agradecer a Chad Buckendahl (EUA); Anne Herrmann e seus colegas da OPP Ltd. (Reino Unido); e April Zenisky, da Universidade de Massachusetts (EUA), por sua cuidadosa revisão de um rascunho anterior do documento. O ITC também agradece a todos os outros revisores de todo o mundo que, direta ou indiretamente, contribuíram para a segunda edição das Diretrizes do ITC para tradução e adaptação de testes.

RESUMO

A segunda edição das Diretrizes ITC para tradução e adaptação de testes foi preparada entre 2005 e 2015 para melhorar a primeira edição e responder aos avanços em testes de tecnologia e práticas. As 18 diretrizes estão organizadas em seis categorias para facilitar seu uso: Pré-condição (3), desenvolvimento de teste (5), confirmação (4), administração (2), pontuação e interpretação (2) e documentação (2). Para cada orientação, uma explicação é fornecida junto com sugestões de prática. Uma lista de verificação é fornecida para melhorar a implementação das diretrizes.

CONTEÚDO

HISTÓRICO

O campo da metodologia de tradução e adaptação de testes avançou rapidamente nos últimos 25 anos, com a publicação de vários livros e muitos estudos novos e exemplos de excelentes trabalhos de adaptação de testes (ver, por exemplo, van de Vijver & Leung, 1997, 2000; Hambleton, Merenda e Spielberger, 2005; Grégoire & Hambleton, 2009; Rios e Sireci, 2014). Estes avanços têm sido necessários devido ao crescente interesse em (1) psicologia intercultural, (2) estudos comparativos internacionais de larga escala de desempenho educacional (por exemplo, TIMSS e OCDE / PISA), (3) exames de credenciamento sendo usados em todo o mundo (por exemplo, no campo da tecnologia da informação por empresas como Microsoft e Cisco), e (4) justiça na consideração dos testes, permitindo que os candidatos escolham o idioma em que as avaliações serão administradas (por exemplo, admissões em universidades em Israel nas quais os candidatos podem fazer muitos de seus testes em um dos seis idiomas).

Avanços técnicos têm sido feitos nas áreas qualitativas e quantitativas para a avaliação de viés de construto, método e item em testes e questionários adaptados, incluindo o uso de procedimentos estatísticos complexos, como Teoria de Resposta ao Item, modelagem de equações estruturais e teoria de generalização (ver Hambleton et al., 2005; Byrne, 2008). Delineamentos de tradução tiveram avanços pela OCDE / PISA (ver Grisay, 2003); foram dados passos para concluir projetos de adaptação de testes (ver, por exemplo, Hambleton & Patsula, 1999; projetos exemplares estão disponíveis para orientar as práticas de adaptação de testes - por exemplo, projetos OECD / PISA e TIMSS); e muitos outros avanços foram feitos.

A primeira edição das Diretrizes (ver van de Vijver & Hambleton, 1996; Hambleton, 2005) partiu de uma perspectiva comparativa, e teve por finalidade da adaptação do teste para permitir ou facilitar comparações entre grupos de participantes. O modelo implícito para o qual as diretrizes foram planejadas utilizava um desenvolvimento de instrumento sucessivo em um contexto comparativo (o instrumento existente deveria ser adaptado para uso em um novo contexto cultural). Está ficando cada vez mais claro, no entanto, que as adaptações de teste têm um domínio de aplicação mais amplo. O exemplo mais importante é o uso de um instrumento novo ou existente em um grupo multicultural, por exemplo, clientes em aconselhamento que vêm de grupos étnicos diferentes, avaliação educacional em grupos etnicamente diversos com um domínio diferencial da linguagem de teste e recrutamento orientado internacionalmente para funções de gestão em empresas multinacionais. Essas mudanças no domínio de aplicabilidade têm implicações para o desenvolvimento, administração, validação e documentação. Por exemplo, consequências possíveis podem resultar em itens de um teste existente, que deve ser adaptado para aumentar sua compreensão para falantes não nativos (isto é, simplificando o idioma). Outra extensão importante das diretrizes seria o desenvolvimento simultâneo (ou seja, o desenvolvimento combinado de testes de idioma de origem e de destino). Projetos internacionais de larga escala usam cada vez mais o desenvolvimento simultâneo para evitar o problema de que a versão desenvolvida em uma língua não possa ser traduzida / adaptada para todos os idiomas do estudo.

A primeira edição das Diretrizes ITC para tradução e adaptação de testes foi publicada por van de Vijver e Hambleton (1996) e por Hambleton (2002) e Hambleton, Merenda e Spielberger (2005). Apenas pequenas mudanças editoriais foram realizadas na publicação das diretrizes entre 1996 e 2005. Entretanto, muitos avanços ocorreram desde 1996. Primeiro, há uma série de revisões úteis das Diretrizes ITC, que incluem

trabalhos de Jeanrie e Bertrand (1999), Tanzer e Sim (1999) e Hambleton (2002). Todos os autores destacaram o valor das diretrizes, mas ofereceram uma série de sugestões para melhorá-las. Hambleton, Merenda e Spielberger (2005) publicaram os principais anais de uma conferência internacional da ITC realizada em 1999 na Georgetown University, nos EUA. Vários autores do capítulo criaram novos paradigmas para adaptações de teste e ofereceram nova metodologia, incluindo Cook e Schmitt-Cascallar (2005), e Sireci (2005). Em 2006, o ITC realizou uma conferência internacional em Bruxelas, na Bélgica, para enfocar as Diretrizes ITC para tradução e adaptação de testes. Mais de 400 pessoas de mais de 40 países concentraram-se no tema da adaptação de testes e muitas novas ideias metodológicas foram avançadas, novas diretrizes foram sugeridas e exemplos de implementações bem-sucedidas foram compartilhados. Foram abundantes os trabalhos apresentados em simpósios em reuniões internacionais de 1996 a 2009 (ver, por exemplo, Grégoire & Hambleton, 2009) e ver Muniz, Elosua e Hambleton (2013) para uma versão inicial da segunda edição das Diretrizes ITC em espanhol.

Em 2007, o Conselho do ITC formou um comitê de seis pessoas e lhes atribuiu a tarefa de atualizar as Diretrizes do ITC para enfatizar o novo conhecimento que estava sendo desenvolvido e as muitas experiências que estavam sendo obtidas pelos pesquisadores no campo. Esses avanços incluem (1) o desenvolvimento de modelagem de equações estruturais para identificar a equivalência fatorial de um teste entre grupos de idiomas, (2) abordagens expandidas para identificar o funcionamento diferencial de itens com escalas de avaliação de respostas politômicas entre grupos de idiomas e (3) novos projetos de adaptação pioneira em projetos de avaliação internacional como OCDE / PISA e TIMSS. O comitê também forneceu apresentações e rascunhos das novas diretrizes em reuniões internacionais de psicólogos em Praga (em 2008) e Oslo (em 2009) e recebeu *feedback* substancial sobre eles.

A seção de Diretrizes Administrativas foi mantida na segunda edição, mas as diretrizes sobrepostas foram combinadas e o número total foi reduzido de seis para dois; "Documentação/ interpretação de escore" que foi a seção final da primeira edição. Na segunda edição, dividimos isso em duas seções: uma focada em escalas e interpretações de pontuação e a outra focada na documentação. Além disso, duas das quatro diretrizes originais desta seção foram substancialmente revisadas.

Como na primeira edição, queremos que fique claro para os leitores nossa distinção entre tradução de teste e adaptação de teste. Provavelmente, a tradução do teste é o termo mais comum, mas a adaptação é o termo mais amplo e refere-se à mudança de um teste de um idioma e de uma cultura para outra. A adaptação do teste refere-se a todas as atividades, incluindo: decidir se um teste em um segundo idioma ou cultura poderia medir o mesmo construto na primeira língua; selecionando tradutores; escolher um desenho para avaliar o trabalho dos tradutores de teste (por exemplo, traduções e traduções reversas) escolher quaisquer acomodações necessárias; modificando o formato do teste; conduzir a tradução; verificar a equivalência do teste no segundo idioma ou cultura e realizar outros estudos de validade necessários. A tradução de teste, por outro lado, tem um significado mais limitado, restrito à escolha real da língua para transpor o teste de um idioma ou cultura para outro, a fim de preservar o significado linguístico. A tradução do teste é apenas uma parte do processo de adaptação, mas pode ser, por si só, uma abordagem muito simplista para transportar um teste de um idioma para outro, sem considerar a sua equivalência educacional ou psicológica do mesmo.

AS DIRETRIZES

INTRODUÇÃO

Um guia é definido em nosso trabalho como uma prática que é importante para a condução e avaliação de adaptação ou desenvolvimento simultâneo de um teste psicológico ou educacional para uso em diferentes populações. No texto a seguir, 18 diretrizes são organizadas em torno de seis tópicos abrangentes: Pré-Condição (3), Desenvolvimento de Testes (5), Confirmação [Análise Empírica] (4), Administração (2), Pontuação das Escalas e Interpretação (2), e Documentação (2).
A primeira sessão chamada "Pré-Condição" destaca o fato que decisões precisam ser tomadas antes do processo de tradução/adaptação começar. A segunda sessão "Diretrizes de Desenvolvimento de Testes" é focada no processo atual de adaptação de teste. A terceira sessão, "Confirmação", inclui as diretrizes associadas à compilação de evidências empíricas que abordam a equivalência, confiabilidade e validade de um teste em múltiplas linguagens e culturas. As últimas três sessões destacam a "Administração", a "Pontuação das Escalas e Interpretação", e a "Documentação". A Documentação tem sido um tópico negligenciado em iniciativas de adaptação de testes na psicologia e educação, e nós gostaríamos de ver editores de periódicos e agências de financiamento exigirem mais quando se trata de documentação do processo de adaptação de testes.
Para cada diretriz, nós oferecemos uma explicação e sugestão para a sua implementação na prática.

Diretrizes de Pré-Condição

PC-1 (1) Obter a permissão necessária para possuir os direitos de propriedade intelectual relacionados ao teste antes de conduzir qualquer adaptação.

Explicação. Direitos de propriedade intelectual referem-se a um conjunto de direitos que as pessoas têm sobre suas criações, invenções ou produtos. Isto porque protege o interesse dos criadores por lhes dar direitos morais e econômicos por suas próprias criações. De acordo com a *World Intellectual Property Organization* (www.wipo.int), *"propriedade intelectual refere-se aos itens de informação ou conhecimento, os quais podem ser incorporados em objetos tangíveis, e, ao mesmo tempo em um número ilimitado de cópias em diferentes locais em qualquer país."*

Há dois tipos de propriedade intelectual: Propriedade industrial e direitos autorais. O primeiro refere-se às patentes que protegem invenções, desenhos industriais, marcas comerciais e nomes comerciais. Os direitos autorais referem-se às criações artísticas ou baseadas em tecnologias. O criador (o autor) tem direitos específicos sobre sua criação (por exemplo, prevenção de algumas distorções quando esta for copiada ou adaptada). Outros direitos (por exemplo, fazer fotocópias) podem ser exercidos por outras pessoas (por exemplo, o editor) o qual possui a licença do autor ou do detentor dos direitos autorais. Para muitos testes, como em outros trabalhos escritos, os direitos autorais são atribuídos pelo autor ao editor ou distribuidor.

Como os testes educacionais ou psicológicos são claramente criações da mente humana, eles estão sob os direitos de propriedade intelectual. Na maioria das vezes, o direito autoral não se refere especificamente ao conteúdo dos itens (por exemplo, ninguém tem direito sobre itens como *"1+1 = ..."* ou *"Eu me sinto triste"),*

mas sim sobre a organização original do teste (estrutura do modo de correção, organização do material, etc.). Consequentemente, imitar um teste existente, por exemplo, mantendo a estrutura original do teste e seu sistema de pontuação, mesmo criando novos itens, é uma violação nos direitos de propriedade intelectual. Quando autorizado a conduzir uma adaptação, o desenvolvedor do teste deve respeitar as características originais do teste (estrutura, material, formato, pontuação...), a menos que um acordo do detentor da propriedade intelectual permita estas características.

Sugestões para a prática. Desenvolvedores de teste devem respeitar a lei de direitos autorais e acordos que existam para o teste original. Eles devem ter um acordo assinado pelo dono da propriedade (por exemplo, o autor ou editora) antes de começar uma adaptação do teste. O acordo deve especificar as modificações no teste adaptado que serão aceitáveis a respeito das características do teste original e deixar claro quem será o dono dos direitos de propriedade intelectual da versão adaptada.

PC-2 (2) Avaliar se a quantidade de sobreposições na definição e conteúdo do construto medido pelo teste e o conteúdo do item nas populações alvo são suficientes para o uso (ou usos) pretendido dos resultados.

Explicação. Esta diretriz exige que o construto que vai ser avaliado seja entendido da mesma maneira em diferentes linguagens e grupos culturais, e esta é a base de comparações transculturais válidas. Neste estágio do processo, o teste ou instrumento ainda não foi adaptado, assim tendo a compilação de evidências empíricas anteriores com testes similares, julgamentos de correspondência de construção de itens e adequação para os grupos dos idiomas envolvidos no estudo, são desejáveis. Em última análise, esta diretriz importante deve ser avaliada com dados empíricos no decorrer das linhas de evidências exigidas no C-2 (10). O objetivo de qualquer análise não é estabelecer a estrutura de um teste, embora isto seja um subproduto de qualquer análise, mas sim confirmar a equivalência da estrutura em múltiplas versões idiomáticas.

Sugestões para a prática. Indivíduos que são especialistas a respeito do construto mensurado, e que estão familiarizados com os grupos culturais que estão sendo testados, devem ser recrutados para avaliar a legitimidade do construto mensurado em cada grupo cultural/linguístico. Eles podem tentar responder à seguinte questão: Este construto faz sentido nas culturas de ambos os grupos? Nós vimos com frequência em testes educacionais, por exemplo, que um comitê julgou que um construto mensurado por um teste não tinha sentido ou tinha seu significado reduzido em uma segunda cultura (por exemplo, qualidade de vida, depressão ou inteligência). Métodos como grupos focais, entrevistas e questionários podem ser utilizados para obter informações estruturadas sobre o grau do construto sobreposto.

PC-3 (3) Minimizar a influência de quaisquer diferenças culturais e linguísticas que são irrelevantes ao uso pretendido do teste nas populações de interesse.

Explicação. As características culturais e linguísticas irrelevantes para as variáveis que o teste pretende medir devem ser identificadas no estágio inicial do projeto. Elas podem estar relacionadas ao formato do item, material (por exemplo, uso de computador, figuras ou ideogramas...), tempo de aplicação, etc.

Uma forma de abordar o problema tem sido avaliar a "distância linguística e cultural" entre o idioma original e a linguagem de destino e nos grupos culturais. Avaliação da distância linguística e cultural pode incluir

considerações de diferenças do idioma, estrutura da família, religião, estilo de vida e valores (van de Vijver & Leung, 1997).

Esta diretriz baseia-se principalmente em métodos qualitativos e em especialistas familiarizados com a pesquisa sobre diferenças linguísticas e culturais específicas. Isso coloca uma pressão especial na seleção dos tradutores dos testes e exige que estes sejam nativos no idioma e cultura pretendido, pois conhecer o idioma de destino não é suficiente para identificar possíveis fontes de viés do método. Por exemplo, em um estudo comparativo de rendimento em matemática, na oitava série, conduzido por Hambleton, Yu e Slater (1999), problemas em relação ao formato e tamanho do teste foram identificados, juntamente com uma série de características culturais associadas ao teste de matemática da oitava série.

Sugestões para a prática. Esta é uma diretriz que será sempre difícil de descrever com dados empíricos. E será especialmente difícil nos estágios iniciais da adaptação de teste. Ao mesmo tempo, evidências qualitativas podem ser coletadas:

- Seja pela observação, entrevista, grupos focais ou questionários, devem ser determinados o nível motivacional dos participantes, sua compreensão das instruções, suas experiências com testes psicológicos, a rapidez associada à administração do teste, familiaridade com as escalas de avaliação e diferenças culturais (mesmo quando estas comparações podem ser problemáticas por causa de diferenças culturais na compreensão das próprias variáveis). Quando é problemático coletar esses dados de pesquisa dos participantes, obtenha o máximo possível de informações dos tradutores. Este trabalho pode ser iniciado antes de qualquer progresso com a adaptação do teste.

- Pode ser possível controlar essas "variáveis de ruído" em qualquer análise empírica subsequente, uma vez que o teste tenha sido adaptado e esteja pronto para estudos de validação por meio do uso de análise de covariância ou outras análises que combinem participantes em grupos linguísticos/culturais em variáveis como nível de motivação ou familiaridade com uma escala de classificação específica (por exemplo, Johnson, 2003; Javaras & Ripley, 2007).

Diretrizes para Desenvolvimento de Testes

TD-1 (4) Garanta que os processos de tradução e adaptação considerem as diferenças linguísticas, psicológicas e culturais nas populações alvo, por meio da escolha de especialistas com experiência relevante.

Explicação. Esta tem sido, ao longo dos anos, uma das diretrizes de maior impacto, porque há evidências consideráveis sugerindo que ela tem influenciado as editoras de testes para procurar tradutores com qualificação, além do conhecimento dos dois idiomas envolvidos na adaptação do teste (veja, por exemplo, Grisay, 2003). O conhecimento das culturas e, pelo menos, o conhecimento geral do tópico e de construção de testes, tem se tornado parte do critério de seleção dos tradutores. Além disso, esta diretriz parece ter influenciado as editoras na tradução e adaptação de testes para usar pelo menos dois tradutores em vários projetos (por exemplo, em projetos de tradução e tradução reversa). A velha prática de depender de apenas um

único tradutor para todas as decisões, por mais bem qualificada que a pessoa seja, foi eliminada da lista de práticas aceitáveis hoje em dia.

O conhecimento (*expertise*) na cultura de destino resulta no uso de tradutores que são nativos no idioma e que estão vivendo naquela cultura, sendo o primeiro essencial e o último altamente desejável. O nativo da língua alvo não irá apenas produzir uma tradução acurada, mas também uma leitura fluente e aparentemente oriunda do país. Viver no local de destino garantirá o conhecimento atualizado do uso do idioma atual.

Em nossa definição, um "especialista" é uma pessoa ou um grupo com conhecimento suficiente: (1) dos idiomas envolvidos, (2) das culturas, (3) do conteúdo do teste, e (4) dos princípios gerais da testagem, que possa produzir uma tradução ou adaptação de um teste com qualidade profissional. Na prática, pode ser eficaz utilizar equipes de pessoas com diferentes qualificações (por exemplo, tradutores com ou sem especialidade em um assunto específico, um especialista em testes, etc.) a fim de identificar áreas que os outros possam ignorar. Em todos os casos, o conhecimento de princípios gerais de testagem, além do conhecimento do conteúdo do teste, deve fazer parte do treinamento recebido pelos tradutores.

Sugestões para a prática. Nós sugerimos o seguinte:

- Escolha tradutores que são falantes nativos do idioma alvo e que tenham um conhecimento profundo da cultura na qual um teste será adaptado, preferivelmente, vivendo no local de destino. Um erro comum é identificar pessoas como tradutores que sabem o idioma, mas não sabem muito bem da cultura, porque um conhecimento profundo da cultura é essencial para manter a equivalência cultural. Ter o conhecimento cultural possibilitará identificar referências culturais (por exemplo, jogo de críquete, Torre Eiffel, Presidente Lincoln, canguru, etc.), com os quais os participantes locais podem não estar familiarizados.

- Selecione tradutores, se possível, com experiência no conteúdo do teste, e com conhecimento de princípios de avaliação (por exemplo, com itens de múltipla escolha, a resposta correta deve ser mais ou menos longa, e não mais longa ou mais curta que as demais escolhas; sinais gramaticais não devem ser úteis na localização da resposta correta; e, em itens de verdadeiro ou falso, declarações verdadeiras não devem ser notavelmente mais longas do que as declarações falsas).

- Tradutores com conhecimento em princípios de desenvolvimento de testes podem ser impossíveis de encontrar na prática e, portanto, seria essencial fornecer treinamento para tradutores para lhes proporcionar os princípios da redação de itens no formato que estarão trabalhando. Sem o treinamento, às vezes, tradutores excessivamente conscientes, introduzirão fontes de erro, o que pode diminuir a validade de um teste traduzido. Por exemplo, as vezes o tradutor pode adicionar uma observação esclarecedora para garantir que uma resposta pretendida seja de fato a resposta correta. Ao fazê-lo, o tradutor pode tornar o item mais fácil do que o pretendido, ou a resposta mais longa pode fornecer uma dica da resposta correta aos candidatos já experientes em testes.

TD-2 (5) Use procedimentos de tradução apropriados para maximizar a adequação da adaptação do teste para as populações alvo.

Explicação: Esta diretriz exige que as decisões tomadas por tradutores, ou grupos de tradutores, maximizem a adequação da versão adaptada à população pretendida. Isso significa que a linguagem deve ser natural e aceitável; focando no que é funcional e não na equivalência literal. Delineamentos comuns de tradução para atingir esses objetivos são as traduções e as traduções reversas. Brislin (1986) e Hambleton e Patsula (1999) discutem exaustivamente sobre estes delineamentos, incluindo suas definições, pontos fortes e fracos. Porém, deve-se notar que ambos os métodos têm falhas, e, portanto, raramente fornecem evidência suficiente para validar um teste traduzido e adaptado. A principal desvantagem do delineamento de tradução reversa é que, se esse método for implementado em sua forma mais restrita, não se conseguirá nenhuma revisão da versão de idioma de destino. Este delineamento resulta em uma versão de idioma de destino do teste que maximiza a facilidade da tradução reversa, mas às vezes produz uma versão estranha do idioma de destino do teste.
Um procedimento de dupla tradução e reconciliação visa abordar as deficiências e os riscos de confiar em idiossincrasias de traduções simplificadas. Nessa abordagem, um terceiro tradutor independente, ou um grupo de especialistas, identifica e resolve qualquer discrepância entre as traduções alternativas e as reconcilia em uma única versão. Em programas de avaliação transcultural em grande escala, como o PISA (*Programme for International Student Assessment* – Programa Internacional de Avaliação de Estudantes), duas versões de idiomas diferentes (por exemplo, inglês e francês) podem ser usadas como fontes separadas de tradução, que são então reconciliadas em uma única versão de idioma alvo (Grisay, 2003). Essa abordagem oferece vantagens importantes, como possíveis discrepâncias que são identificadas e revisadas diretamente no idioma de destino. Além disso, o uso de mais de um idioma de origem ajuda a minimizar o impacto das características culturais da fonte.

Diferenças na estrutura da linguagem podem causar problemas na tradução do teste. Por exemplo, em uma escala bem conhecida desenvolvida por Rotter e Rafferty (1950) em inglês, os examinandos são solicitados a preencher os espaços em branco no formato de itens incompletos, como: *"Eu gosto….."; "Eu me arrependo….."; "Eu não posso ....."*.

No entanto, o mesmo formato é inadequado na língua turca, onde o objeto de uma sentença deve vir antes do verbo e do sujeito. O uso de sentenças incompletas como na versão em inglês, portanto, mudaria completamente a resposta, já que os estudantes turcos deveriam primeiro olhar para o final da declaração antes de preencher o início.
Em quaisquer soluções alternativas para este problema, a versão traduzida (ou seja, a linguagem de destino) será de alguma forma diferente da versão do idioma de origem em termos de especificações de formato.

Sugestões para prática. A compilação de dados de julgamento por revisores parece especialmente valiosa para verificar se esta diretriz é atendida:

- Utilize as escalas avançadas de avaliação desenvolvidas por Brislin (1986), Jeanrie e Bertrand (1999), ou Hambleton e Zenisky (2010). Hambleton e Zenisky fornecem uma lista empiricamente validada de 25 características diferentes de um teste traduzido que devem ser verificadas durante o processo de adaptação. Exemplos de perguntas de Hambleton e Zenisky (2010) incluem "A linguagem do item traduzido é de dificuldade comparável e as palavras similares ao item na versão do idioma original?" e "A tradução introduz alterações no texto (omissões, substituições ou acréscimos) que possam influenciar a dificuldade do item do teste nas duas versões linguísticas?"

- Use delineamentos múltiplos de tradução, se for viável. Por exemplo, um delineamento de tradução

reversa pode ser usado para verificar a versão de destino criada por meio de tradução dupla e reconciliação por uma gama de especialistas.

- Se versões de um teste têm por objetivo ser usadas de forma transcultural, considere o desenvolvimento simultâneo ou concorrente de vários idiomas do teste desde o início, a fim de evitar problemas futuros com a tradução / adaptação da versão de origem. Mais informações sobre o desenvolvimento de testes simultâneos podem ser encontradas, por exemplo, em Solano-Flores, Trumbull e Nelson-Barber (2002). No mínimo, delineie a versão de origem que permita futuras traduções e evite possíveis problemas o máximo possível; especificamente, evitando referências culturais, itens idiossincráticos e formatos de resposta, etc.

- Considerando as diferenças sintáxicas entre os idiomas, o uso de formatos que dependam da estrutura rígida das sentenças deve ser evitado em avaliações internacionais de grande escala e, provavelmente, também com testes psicológicos, devido aos problemas de tradução que possam surgir.

TD-3 (6) Fornecer evidência de que as instruções do teste e o conteúdo do item têm um significado semelhante para todas as populações alvo.

Explicação. A evidência exigida pela diretriz pode ser obtida por meio de várias estratégias (ver, por exemplo, van de Vijver e Tanzer, 1997). Essas estratégias incluem (1) o uso de revisores nativos da cultura e da língua; (2) uso de amostras de respondentes bilíngues; (3) uso de pesquisas locais para avaliar o teste; e (4) uso de administrações de teste não padronizadas para aumentar a aceitabilidade e validade.
Realizar uma versão piloto para adaptação do teste é uma boa ideia. Esta amostra piloto pode ser utilizada não somente para a administração de teste, como também para a análise de dados. Também são importantes as entrevistas com os administradores dos testes e com os examinandos para obter suas críticas ao teste. Outros delineamentos que usam especialistas em conteúdo de diferentes contextos culturais, ou especialistas em conteúdo bilíngue são possíveis. Por exemplo, especialistas em conteúdo bilíngue poderiam ser solicitados para comparar a semelhança da dificuldade dos formatos de itens e do conteúdo dos dois testes. A entrevista cognitiva é outro método que está se mostrando promissor (Levin, et al., 2009).

Sugestões para a prática. Diversas sugestões foram oferecidas acima para abordar esta diretriz. Por exemplo,

- Use revisores nativos da cultura e idioma locais para avaliar a tradução / adaptação do teste.

- Use amostras de respondentes bilíngues para fornecer algumas sugestões sobre a equivalência das duas versões do teste, tanto nas instruções do teste, quanto em seus itens.

- Use pesquisas locais para avaliar o teste. Esses testes piloto podem ser muito valiosos. Certifique-se de entrevistar o administrador do teste e os entrevistados após a administração do teste, pois, frequentemente, os comentários do administrador e do respondente são mais valiosos do que as respostas reais dos respondentes do teste.

- Use administrações de teste adaptados para aumentar a aceitação e validade. Seguir instruções de testes semelhantes não faz sentido se forem mal interpretadas pelos entrevistados da segunda língua do

teste/grupo cultural.

TD-4 (7) Fornecer evidências de que o formato dos itens, escalas de classificação, categorias de pontuação, convenções de teste, modos de administração e outros procedimentos são adequados para todas as populações alvo.

Explicação. Formatos de itens como as escalas de avaliação de cinco pontos ou novos formatos de itens como "arrastar e soltar" (testes computadorizados) ou "responder a todos que estão corretos" ou até mesmo "responder uma, e, apenas uma opção de resposta" podem ser confusos para os participantes que nunca viram o formato desses itens anteriormente. Mesmo a aparência de itens, o uso de gráficos ou formatos de itens computadorizados podem ser confusos para os candidatos. Há muitos exemplos desses tipos de erros encontrados nos Estados Unidos com a iniciativa de transferir grande parte de teste padronizado para crianças para o computador. Através de exercícios práticos, estes problemas podem ser superados para a maioria das crianças. Esses novos formatos de item devem ser familiares aos entrevistados ou uma fonte de viés de teste é introduzida, o que pode distorcer qualquer resultado do teste individual e em grupo.

Um problema emergente pode estar associado às versões de um teste administrado por computador. Se os entrevistados não estiverem familiarizados com a plataforma do teste, será necessário um tutorial para garantir que esses respondentes obtenham a familiaridade necessária para que um teste administrado pelo computador forneça pontuações significativas.

Sugestões para a prática. Ambas as evidências qualitativas e quantitativas desempenham um papel na avaliação desta diretriz. Existem várias maneiras para verificar a adaptação de um teste:

- Avalie se os exercícios práticos são suficientes para equiparar os respondentes ao nível necessário para que eles forneçam respostas honestas e / ou que reflitam seu nível de domínio do material.

- Assegure-se de que os respondentes estejam familiarizados com quaisquer novos formatos de itens ou administrações de teste (como administração por computadores) que tenham sido incorporados ao processo do teste.

- Verifique se as orientações do teste (por exemplo, a colocação de quaisquer figuras ou a marcação de respostas em uma folha de respostas) serão claras para os entrevistados.

- Novamente, os formulários de avaliação fornecidos por Jeanrie e Bertrand (1999) e Hambleton e Zenisky (2010) são úteis. Por exemplo, Hambleton e Zenisky incluíram perguntas como "O formato do item, incluindo a aparência física, é o mesmo nas duas versões da linguagem?", E "Se uma forma de ênfase de palavra ou frase (negrito, itálico, sublinhado, etc.) foi usado no item de idioma de origem, essa ênfase foi usada no item traduzido? "

TD-5 (8) Coletar dados piloto sobre o teste adaptado para permitir a análise de itens, avaliação de confiabilidade e estudos de validade de pequena escala, para que quaisquer revisões necessárias ao teste adaptado possam ser feitas.

Explicação. Antes de iniciar quaisquer estudos de confiabilidade e validade de escores em larga escala e/ou estudos normativos que possam ser demorados e caros é importante ter evidências confirmadas sobre a qualidade psicométrica do teste adaptado. Existem muitas análises psicométricas que podem ser realizadas para fornecer evidências iniciais de confiabilidade e validade da pontuação. Por exemplo, no estágio de desenvolvimento de teste, uma análise de item usando pelo menos um tamanho de amostra modesto (por exemplo, 100) pode fornecer dados necessários sobre o funcionamento de itens de testes específicos. Itens que são muito fáceis ou difíceis em comparação a outros itens, ou que apresentam poder de discriminação baixa ou negativa, podem ser revistos para possíveis falhas de itens. Com itens de múltipla escolha, seria apropriado investigar a eficácia dos distratores de itens. Problemas podem ser encontrados e serem feitas revisões. Além disso, com os mesmos dados compilados para análise de itens, o coeficiente alfa ou coeficiente ômega (McDonald, 1999) fornece ao pesquisador informações valiosas que podem ser usadas para apoiar decisões sobre o tamanho adequado das versões de idioma de origem e destino do teste.

Em alguns casos, ainda podem existir dúvidas sobre certos aspectos da adaptação: As instruções do teste podem ser totalmente compreendidas? As instruções devem ser diferentes para orientar os participantes do teste no novo idioma e cultura? O teste administrado por computador causará problemas para os entrevistados selecionados (por exemplo, participantes de baixo nível socioeconômico) na população de interesse para o teste adaptado? Existem muitas perguntas sendo apresentadas no tempo disponível?
Todas essas perguntas e muitas outras poderiam ser respondidas com estudos de validade de tamanho modesto. O objetivo seria compilar dados suficientes para que se possa decidir se deve ou não avançar com o teste adaptado. Se a decisão é avançar, então uma série de estudos substancialmente mais ambiciosos pode ser planejada e executada (por exemplo, estudos de nível do item, Funcionamento Diferencial do Item (DIF, *Differential item functioning*), e estudos para investigar a estrutura fatorial do teste.

Sugestões para prática. Existem várias análises básicas que podem ser realizadas:

- Realizar um estudo clássico de análise de itens para obter informações sobre os índices de itens e os índices de discriminação de itens. Com itens de múltipla escolha ou itens de seleção semelhantes, realizar também uma análise dos distratores.

- Realizar uma análise de confiabilidade (por exemplo, KR-20 com itens dicotômicos ou coeficiente alfa ou coeficiente ômega com itens marcados em mais de duas opções).

- Conforme necessário, pode-se realizar um ou dois estudos para obter informações sobre a validade do teste adaptado. Por exemplo, suponha que o teste adaptado seja administrado por meio de um computador. Pode ser desejável realizar um estudo para avaliar o modo de administração do teste (ou seja, papel e lápis *versus* computador). Suponha que as instruções exijam que os participantes respondam a todas as perguntas. Pode ser necessário fazer alguma pesquisa para determinar as melhores instruções para atingir esse objetivo. Pesquisadores descobriram que é difícil conseguir que alguns participantes respondam a todas as perguntas, a não ser que a opção de marcar qualquer escolha for encorajada quando os entrevistados estiverem em dúvida.

Diretrizes para Confirmação

As Diretrizes para Confirmação são baseadas em análises empíricas de estudos de validade em larga escala.

C-1 (9) Selecionar uma amostra com características que sejam pertinentes para o uso pretendido do teste, o tamanho e relevância suficientes para as análises empíricas.

Explicação. O delineamento da coleta de dados refere-se à maneira como estes são coletados para estabelecer normas (se necessário) e equivalência entre as versões linguísticas de um teste, para conduzir estudos de validade e confiabilidade e estudos de DIF. Um primeiro requisito em relação à coleta de dados é que as amostras devem ser grandes o suficiente para permitir a disponibilidade de informação estatística. Embora esse requisito seja válido para qualquer tipo de pesquisa, é particularmente relevante no contexto de um estudo de validação de adaptação de teste. Tal fato ocorre porque as técnicas estatísticas necessárias para estabelecer equivalência de teste e item (por exemplo, análise fatorial confirmatória, abordagens de TRI para a identificação de teste/itens potencialmente enviesados) podem ser aplicados de forma mais significativa com amostras suficientemente grandes para estimar os parâmetros do modelo. O tamanho de amostra recomendado dependerá da complexidade do modelo e da qualidade dos dados.

Além disso, a amostra para o estudo de validade de uma escala deve ser representativa da população pretendida para o teste. Chamamos a atenção para o importante artigo de Van de Vijver e Tanzer (1997), e as contribuições metodológicas encontradas em Van de Vijver e Leung (1997), Hambleton, Merenda e Spielberger (2005), Byrne (2008), e Byrne e Van de Vijver (2014), para orientar a seleção de delineamentos e análises estatísticas apropriadas. Sireci (1997) discutiu os problemas e questões na vinculação de testes em vários idiomas a uma escala comum.

Às vezes, na prática, a população pretendida para a versão do idioma alvo de um teste pode pontuar muito abaixo ou acima e/ou ser mais ou menos homogênea do que o grupo do idioma de origem. Isso resulta grandes problemas para certos métodos de análise, como estudos de confiabilidade e validade. Uma solução é escolher uma subamostra do grupo de idiomas de origem para corresponder à amostra do grupo de idiomas de destino. Com amostras correspondentes, quaisquer diferenças nos resultados das amostras correspondentes que possam ser devidas a diferenças nas formas das distribuições nos dois grupos podem ser eliminadas (ver Sireci & Wells, 2010). Por exemplo, as comparações da estrutura de teste geralmente envolvem covariâncias, e elas variarão em função das distribuições de pontuação. Usando amostras correspondentes, qualquer finalidade que a distribuição das pontuações possa ter nos resultados é comparável nas duas amostras, e o impacto das distribuições de pontuação nos resultados pode ser descartado como uma explicação para quaisquer diferenças nos resultados.

Talvez mais um exemplo possa ajudar a explicar o problema de diferentes distribuições de pontuação nos grupos de idiomas de origem e de destino. Supondo que a confiabilidade do escore do teste seja 0,80 no grupo de idiomas de origem, mas apenas 0,60 no grupo de idiomas de destino. A diferença pode parecer preocupante e levantar questões sobre a adequação da versão em idioma de destino do teste. No entanto, muitas vezes é esquecido que a confiabilidade é uma característica conjunta do teste e da população (McDonald, 1999) porque

depende da variância do escore real (característica da população) e da variância do erro (característica do teste). Portanto, a mesma variação de erro pode levar a uma maior confiabilidade simplesmente devido à maior variação de pontuação verdadeira no grupo de idiomas de origem. McDonald (1999) mostra que o erro padrão de medida (que é a raiz quadrada da variação do erro) é, na verdade, um indicador mais apropriado para comparar as amostras e não a confiabilidade. Outra alternativa usando os coeficientes de confiabilidade seria delinear uma amostra comparativa de participantes do grupo de idiomas de origem e recalcular a confiabilidade da pontuação do teste.

Abordagens modernas para testar invariância de medida usando Análise Fatorial Confirmatória (CFA) com grupos múltiplos permite que amostras com diferentes distribuições das características latentes sejam avaliadas. Em tais modelos, enquanto parâmetros de medida, como cargas fatoriais de item e interceptações são consideradas iguais entre grupos, as médias, variâncias e covariâncias das características latentes podem variar entre os grupos. Isso permite o uso de amostras completas e apresenta um cenário mais realista de diferentes distribuições dos traços medidos entre diferentes populações.

Sugestões para prática. Em quase todas as pesquisas, há duas sugestões que são feitas ao descrever a(s) amostra(s):

- Coletar uma amostra tão grande quanto razoável, dado que estudos para identificar itens de teste potencialmente tendenciosos requerem um mínimo de 200 pessoas por versão do teste (Mazor, Clauser & Hambleton, 1992; Subok, 2017). Para realizar análises de teoria de resposta de itens e investigações de ajuste de modelos é necessária uma amostra de pelo menos 500 respondentes (Hulin, Lissak & Drasgow, 1982; Hambleton, Swaminathan & Rogers, 1991), enquanto estudos para investigar a estrutura fatorial de um teste requerem grandes amostras, 300 ou mais respondentes (Wolf, Harrington, Clark & Miller, 2013). Claramente, análises com amostras menores também são possíveis - mas a primeira regra é gerar grandes amostras sempre que possível.

- Escolher amostras representativas dos entrevistados sempre que possível. Generalizações de resultados com amostras não representativas dos entrevistados são limitadas. Para eliminar diferenças nos resultados devido aos fatores metodológicos, tais como variações nas distribuições de pontuação, coletar uma amostra do grupo de idiomas de origem para corresponder ao grupo de idiomas de destino, é geralmente uma boa ideia. Comparações dos erros padrão de medição podem ser mais apropriadas.

C-2. (10) Fornecer evidência estatística relevante sobre a equivalência do construto, equivalência do método e equivalência de item para todas as populações pretendidas.

Explicação. Estabelecer a equivalência de construto das versões de idioma de origem e de destino de um teste é importante, mas não é a única análise empírica importante a ser realizada. Além disso, abordagens para a equivalência de construto (PC-2) e equivalência de método (PC-3) foram descritas
brevemente nas diretrizes anteriores.

Os pesquisadores também precisam se referir à equivalência do nível do item. A equivalência de itens foi estudada sob o título “análise de funcionamento diferencial do item (DIF)”. Em geral, o DIF existe se dois indivíduos testados, de duas populações diferentes (cultural - linguística), tiverem o mesmo nível de traço medido, mas apresentaram uma probabilidade de resposta diferente a um item do teste. Diferenças gerais no

desempenho de um teste entre grupos podem ocorrer, mas isto não é um problema por si só. Considerando que, quando os membros de uma população correspondem a um construto medido pelo teste (tipicamente, o escore total, ou o escore total menos o escore do item em estudo), e diferenças de desempenho existem entre os grupos, o DIF está presente no item. Este tipo de análise é realizada para cada item do teste. Posteriormente, uma tentativa é feita para entender a razão para o DIF nos itens, e, baseado nesta revisão de julgamento, alguns itens podem ser identificados como defeituosos sendo alterados ou removidos completamente do teste.

Duas importantes fontes potenciais de DIF a serem avaliadas são problemas de tradução e diferenças culturais. Mais especificamente, o DIF pode ocorrer devido a (1) não equivalência de tradução que ocorre de versões do teste de origem para o idioma alvo, tais como familiaridade com o vocabulário usado, alteração na dificuldade do item, alteração na equivalência do significado, etc. e (2) diferenças contextuais culturais (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).

Durante a tradução, existe a possibilidade de ser utilizado vocabulário pouco comum no idioma de destino. Os significados poderiam ser os mesmos nas versões traduzidas, porém, em uma cultura, uma palavra poderia ser mais utilizada comparada à outra. Também é possível alterar o nível de dificuldade do item como resultado da tradução devido ao tamanho e complexidade da frase e ao uso de vocabulário fácil ou difícil. O significado também pode mudar no idioma de destino com a exclusão de algumas partes das frases, traduções imprecisas, mais de um significado no vocabulário usado no idioma de destino, significados não equivalentes de algumas palavras nas culturas, etc. Além de tudo, as diferenças culturais podem fazer com que os itens funcionem de maneira diferente nos idiomas. Por exemplo, palavras como "hambúrguer" ou "caixa registradora" podem não ser entendidas ou ter um significado diferente em duas culturas.

Existem pelo menos quatro grupos de análises para verificar se os itens estão funcionando de maneira diferente no idioma e/ou grupos culturais. Estes procedimentos são: (a) baseados no TRI (ver, Ellis, 1989; Thissen, Steinberg, & Wainer, 1988; 1993; Ellis & Kimmel, 1992), (b) procedimento e extensões Mantel-Haenszel (MH) (ver, Dorans & Holland, 1993; Hambleton, Clauser, Mazor, & Jones, 1993; Holland & Wainer, 1993; Sireci & Allalouf, 2003), (c) regressão logística (LR) (Swaminathan & Rogers, 1990; Rogers & Swaminathan, 1993), e (d) Análise Fatorial Restrita (RFA – *Retricted Factor Analysis*) (Oort & Berberoğlu, 1992).

Nas propostas baseadas no TRI, as amostras dos participantes dos testes nos dois idiomas são combinadas com base nos escores do traço latente. Nas metodologias de MH e LR, o escore do teste observado ou estimado é usado como critério de comparação antes de comparar o desempenho no item pelos dois grupos. Embora o escore da soma seja o critério de correspondência mais popular nesses procedimentos, outros escores estimados, por exemplo, a partir da análise fatorial, também podem ser usados. Essas pontuações também são "purificadas" pela exclusão de itens questionáveis. O critério de correspondência deve ser válido e confiável o suficiente para avaliar o DIF corretamente. Na RFA, cada item passa por regressão na variável de agrupamento (violador em potencial), bem como no traço latente. A carga de cada item é liberada e o ajuste ao modelo é avaliado com referência ao modelo nulo, no qual nenhum item carrega na variável de agrupamento (nenhum modelo DIF). Se o modelo fornece um ajuste significativamente melhor, isso indica que existe DIF no item.

Quando um teste é complexo em suas dimensões, encontrar um critério de correspondência apropriado é um problema (Clauser, Nungester, Mazor & Ripkey, 1996). O uso de critérios de multivariado de correspondência

tais como, diferenças nos resultados dos fatores obtidos através da análise fatorial, podem alterar as interpretações do nível de DIF do item. Assim, esta diretriz sugere que se o teste for multidimensional, os pesquisadores devem usar vários critérios para encontrar os itens com DIF. Devem ser avaliados os itens que sempre aparecem com DIF em relação a vários critérios correspondentes. A comparação multivariada pode reduzir o número de itens que exibem DIF entre os idiomas e grupos culturais.

Essas metodologias podem exigir diferentes tamanhos de amostra. MH, LR e RFA são modelos que podem ser utilizados de forma confiável e válida com amostras relativamente pequenas em comparação com técnicas baseadas em TRI, que requerem amostras maiores para estimativas de parâmetros. Outra consideração é o tipo de dados de resposta do item. MH, LR e RFA podem ser aplicados a dados com pontuação binária. Outros métodos, tais como o MH generalizado, requerem dados de resposta politômica.

Essa diretriz exige que os pesquisadores localizem possíveis fontes de viés no método do teste adaptado. Fontes de viés de método incluem: (1) diferentes níveis de motivação dos participantes do teste; (2) experiência diferenciada por parte dos entrevistados com testes psicológicos; (3) maior rapidez do teste no idioma de um grupo do que o outro; (4) familiaridade com o formato de resposta entre os grupos de idiomas; (5) heterogeneidade do estilo de resposta, etc. Os vieses nas respostas têm sido, por exemplo, uma grande preocupação na interpretação dos resultados do PISA e tem recebido atenção nas pesquisas.

Finalmente, mas ainda importante, essa diretriz exigirá que os pesquisadores discutam a equivalência do construto. Existem pelo menos quatro abordagens estatísticas para avaliar a equivalência de construto nas versões de idioma de origem e de destino de um teste: Análise Fatorial Exploratória (EFA), Análise Fatorial Confirmatória (CFA), Escalonamento Multidimensional (MDS) e comparação de redes nomológicas (Sireci, Patsula, & Hambleton, 2005).

De acordo com van de Vijver and Poortinga (1991), a análise fatorial (ambas, EFA e CFA) são as técnicas estatísticas mais utilizadas para avaliar se um construto em uma cultura é encontrado na mesma forma e frequência em outra cultura. Essa declaração de 1991 permanece verdadeira até hoje, embora as abordagens de modelagem estatística tenham avançado consideravelmente (ver, por exemplo, Hambleton & Lee, 2013, Byrne, 2008). Entretanto, com a EFA, é difícil comparar estruturas fatoriais separadas, e não há regras estabelecidas para decidir quando as estruturas podem ser consideradas equivalentes. Análises estatísticas tais como a CFA (ver, por exemplo, Byrne, 2001, 2003, 2006, 2008) e Escalonamento Multidimensional Ponderado (WMDS) são mais desejáveis, pois podem acomodar simultaneamente vários grupos (Sireci, Harter, Yang, & Bhola, 2003).

Existem muitos estudos nos quais o CFA foi utilizado para avaliar se a estrutura fatorial de uma versão original de um teste era consistente com suas versões adaptadas (por exemplo, Byrne & van de Vijver, 2014). A CFA é recomendada para avaliar a equivalência estrutural em testes adaptados, pois ela pode lidar com vários grupos simultaneamente, e os testes estatísticos de ajuste de modelo estão disponíveis, assim como os índices descritivos de ajuste de modelo (Sireci, Patsula, & Hambleton, 2005). A capacidade de lidar com múltiplos grupos é especialmente importante, pois está se tornando comum adaptar testes em muitos idiomas (por exemplo, algumas medidas de inteligência estão sendo traduzidas / adaptadas para mais de cem idiomas e, os testes TIMSS e OECD / PISA, estão adaptados em mais de 30 idiomas). Como a única exigência é zero para as cargas cruzadas no CFA, muitas vezes isso não acontece para ajustar bem os dados nos instrumentos multidimensionais e complexos. O modelo Exploratório de Equações Estruturais (ESEM, *Exploratory*

Structural Equational Modelling) está se tornando cada vez mais popular, especialmente com dados de personalidade ou variáveis mais complexas e inter-relacionadas (Asparouhov & Muthén, 2009).

O WMDS (*Weighted Multidimensional Scaling/* Escalonamento Multidimensional Ponderado*)* é outra abordagem atraente para avaliar a equivalência de construto em diferentes versões linguísticas de uma avaliação. Assim como o EFA, a análise de WMDS não requer especificação de estrutura de teste a priori e, como o CFA, permite a análise de múltiplos grupos (e.g., Sireci, et al., 2003).

Van de Vijver e Tanzer (1997) sugeriram que pesquisadores transculturais deveriam examinar a confiabilidade de cada versão cultural do teste de interesse e buscar evidências de validade convergente e discriminante em cada grupo cultural. Esses estudos podem ser mais práticos do que estudos de estrutura de teste que exigem tamanhos de amostra muito substanciais.

Deve-se reconhecer, no entanto, que a comparação do desempenho do participante em duas versões de um teste nem sempre é o objetivo de traduzir/adaptar um teste. Talvez, por exemplo, o objetivo seja simplesmente avaliar os participantes em um grupo de idiomas diferente em um construto. Neste caso, um exame cuidadoso da validade do teste no segundo grupo linguístico é essencial, mas a pesquisa para encontrar evidências da equivalência das duas formas não é tão crítica. A importância desta diretriz dependerá da finalidade ou propósitos do teste no segundo idioma (ou seja, o grupo do idioma alvo). Testes como aqueles usados no PISA ou no TIMSS exigem altas evidências de sobreposição de conteúdo, pois os resultados são usados para comparar o desempenho dos alunos em muitos países. O uso de um inventário de depressão traduzido do inglês para o chinês, para pesquisadores estudarem depressão, ou para conselheiros avaliarem a depressão de seus clientes, não exigiria alta sobreposição de conteúdo. Em vez disso, a validade para apoiar o inventário de depressão na China seria necessária.

Esta diretriz também pode estar relacionada com métodos estatísticos após o teste ter sido adaptado. Por exemplo, se pensa que os grupos culturais diferem em variáveis importantes irrelevantes para o construto medido, delineamentos abrangentes e análises estatísticas podem ser usados para controlar essas variáveis 'incômodas'. Análises de covariância, desenhos de blocos randomizados e outras técnicas estatísticas (análise de regressão, correlação parcial, etc.) podem ser usadas para controlar os efeitos de fontes indesejadas de variação entre os grupos.

Sugestões para prática. Esta é uma diretriz muito importante e existem muitas análises que podem ser realizadas. Para análises de equivalência, oferecemos as seguintes sugestões para prática:

- Se os tamanhos de amostra forem suficientes, realize um estudo comparativo da equivalência de construto das versões de idioma de origem e de destino do teste. Existem muitos pacotes de software para facilitar essas análises (ver Byrne, 2006).

- Conduza a análise exploratória (preferencialmente rodando para uma estrutura alvo - chamada “rotação alvo”) ou análise fatorial confirmatória, e/ou análise de Escalonamento Multidimensional Ponderado para determinar o nível de concordância na estrutura do teste de interesse por meio da linguagem e/ou grupos culturais. A exigência de amostras grandes (10 pessoas por variável) dificulta a realização dessas análises em muitos estudos transculturais. Um excelente modelo para um estudo desse tipo é Byrne e van de Vijver (2014).

- Procure por evidências de validade convergente e discriminante (essencialmente, procure evidências correlacionais entre um conjunto de construtos e verifique a estabilidade dessas correlações entre grupos linguísticos e/ou culturais) (ver Van de Vijver & Tanzer, 1997).

Para análises de DIF, algumas sugestões são identificadas abaixo. Para abordagens mais sofisticadas, os pesquisadores são encorajados a ler a literatura profissional sobre DIF:

- Realize uma análise de DIF usando um dos procedimentos padrão (se os itens tiverem pontuação binária, o procedimento de Mantel-Haenszel pode ser o mais simples; se forem pontuados com mais de um item o procedimento generalizado de Mantel-Haenszel é uma opção). Outras soluções mais trabalhosas incluem abordagens baseadas em TRI. Se os tamanhos das amostras forem mais modestos, um *delta plot* pode revelar itens potencialmente falhos. Comparações condicionais são outra possibilidade (para uma comparação de resultados com métodos para pequena amostragem, ver, Muñiz, Hambleton, & Xing, 2001).

C-3. (11) Fornecer evidências que apoiem as normas, a confiabilidade e a validade da versão adaptada do teste nas populações pretendidas.

Explicação. As normas, evidências de validade e evidências de confiabilidade de um teste em sua versão original não se aplicam automaticamente as outras possíveis adaptações do teste em diferentes culturas e idiomas. Portanto, evidências empíricas de validade e confiabilidade de quaisquer novas versões desenvolvidas também devem ser apresentadas. Todos os tipos de evidências empíricas que apoiam as inferências feitas a partir do teste devem ser incluídas no manual do teste. Atenção especial deve ser dada às cinco fontes de evidências de validade baseadas em: conteúdo do teste, processos de resposta, estrutura interna, relações com outras variáveis e consequências do teste (AERA, APA, NCME, 2014). Análise fatorial exploratória e confirmatória, modelagem de equações estruturais e análises múltiplas de traços são algumas das técnicas estatísticas que podem ser usadas para obter e analisar evidências de validade de análises de dados baseadas em estrutura interna.

Sugestões para prática. As sugestões são as mesmas que seriam necessárias para qualquer teste que esteja sendo considerado para uso:

- Se as normas desenvolvidas para a versão original do teste forem sugeridas para serem usadas para a versão adaptada, deve-se fornecer evidências de que esse uso é estatisticamente apropriado e justo. Se nenhuma evidência puder ser fornecida para tal uso das normas originais, devem ser desenvolvidas normas específicas para a versão adaptada de acordo com os padrões para desenvolvimento de normas.

- Compile uma quantidade suficiente de evidências de confiabilidade para justificar o uso da versão do idioma adaptado. A evidência pode normalmente incluir uma estimativa de consistência interna (por exemplo, KR-20 ou coeficiente alfa ou ômega).

- Compile o máximo de evidências de validade necessárias para determinar se a versão do idioma de destino do teste deve ser usada. O tipo de evidência compilada dependerá do uso pretendido das pontuações (por exemplo, validade de conteúdo para testes de desempenho, validade preditiva para

testes de aptidão etc.).

C-4. (12) Use um delineamento de equacionamento apropriado e procedimentos de análise de dados ao vincular escalas de pontuação de diferentes versões de idioma de um teste.

Explicação. Ao vincular duas versões de idiomas de um teste a uma única escala, várias opções são possíveis. Se um conjunto comum de itens for usado, o funcionamento desses itens comuns nos dois grupos de idiomas deve ser avaliado e, se o funcionamento diferencial for observado, sua remoção dos dados usados no estabelecimento do vínculo deve ser considerada. Os gráficos delta (Angoff & Modu, 1973) servem bem a esse propósito, e Cook e Schmitt-Cascallar (2005) fornecem uma boa ilustração de como usar gráficos delta para identificar itens que têm um significado diferente para os dois grupos de examinandos. Nem todos os tipos de item têm o mesmo potencial de vincular as versões de idioma. As estimativas de parâmetro de dificuldade e discriminação de itens, derivadas da estrutura da teoria da resposta ao item para itens comuns, podem ser colocados em gráfico para ajudar a identificar itens com desempenho inadequado (ver Hambleton, Swaminathan, & Rogers, 1991).

Porém, vincular (ou seja, "equacionar") as pontuações em duas versões de idiomas de um teste sempre será problemático, pois é necessário fazer fortes suposições sobre os dados. Às vezes, uma suposição altamente problemática é feita de que as diferentes versões de idioma do teste são equivalentes e, em seguida, as pontuações das duas versões do teste são usadas de forma intercambiável. Tal suposição pode ter mérito em testes de matemática, porque a tradução/adaptação é tipicamente direta. Pode ter mérito, também, se as duas versões do teste foram cuidadosamente construídas. Sendo assim, pode supor que a versão do idioma de origem do teste funciona com a população original, da mesma forma que a versão do idioma de destino do teste funciona. Essa suposição pode ter mérito se todas as outras evidências disponíveis sugerirem que as duas versões de idioma do teste são equivalentes e não há desvios de método que influenciem as pontuações na versão de idioma de destino do teste.

Existem outras duas soluções, porém, nenhuma é perfeita. Primeiro, a vinculação poderia ser feita com uma subamostra dos itens considerados equivalentes nas duas versões linguísticas do teste. Por exemplo, os itens podem ser os que foram julgados muito fáceis de traduzir/adaptar. A princípio, a solução poderia funcionar, mas requer que os itens de ligação e o restante dos itens do teste estejam medindo o mesmo construto. Uma segunda solução envolve a vinculação por meio de uma amostra de candidatos que são bilíngues. Com a mesma amostra tomando as duas versões do teste, seria possível estabelecer uma tabela de conversão de pontuação. A amostra não poderia ser muito pequena e, no delineamento, a ordem de apresentação das formas do teste seria contrabalançada. O grande pressuposto dessa abordagem é que os candidatos são realmente bilíngues e, portanto, além das dificuldades relativas a forma do teste, os candidatos devem ter um bom desempenho em ambas. Qualquer diferença pode ser usada para ajustar as pontuações na conversão de pontuações de uma versão do teste para a outra.

Sugestões para prática. Vincular pontuações em versões adaptadas de um teste será problemático na melhor das hipóteses, porque todos os delineamentos de equacionamento têm pelo menos uma lacuna importante. Provavelmente, a melhor estratégia é considerar todas as etapas para estabelecer a equivalência de pontuação. Se a evidência relacionada as três perguntas abaixo for forte, até mesmo as pontuações das duas versões do teste podem ser tratadas de maneira intercambiável:

- Há evidência de que o mesmo construto está sendo medido nas versões de idioma de origem e de destino do teste? O construto tem a mesma relação com outras variáveis externas na nova cultura?

- Há fortes evidências de que as fontes de viés de método foram eliminadas (por exemplo, nenhum problema de tempo, os formatos usados no teste são igualmente familiares aos candidatos, não há confusão sobre as instruções, nenhuma falta de representatividade sistemática em um grupo ou outro, instruções padronizadas, ausência de estilos de resposta (pontuações extremas, motivação diferenciada, etc.)?

- O teste está livre de itens potencialmente tendenciosos? Aqui, um gráfico de valores *p* ou valores delta de itens nas duas versões do teste, pode ser muito útil. Pontos que não caiam ao longo da linha de equação linear devem ser estudados para determinar se os itens associados são igualmente adequados em ambos os idiomas. As análises DIF fornecem evidências ainda mais fortes sobre a equivalência de itens entre os grupos linguísticos e culturais.

- Se a vinculação de pontuações for realizada, então um delineamento de vinculação apropriado precisa ser escolhido e implementado. Evidência da validade do delineamento deve ser fornecida.

Administração das Diretrizes

A-1 (13) Prepare os materiais e instruções de administração a fim de minimizar quaisquer problemas relacionados a cultura e idioma (que podem ser causados por procedimentos de administração) e no modo de resposta (que possam afetar a validade das inferências retiradas das pontuações).

Explicação. A implementação das diretrizes administrativas deve começar com uma análise de todos os fatores que podem ameaçar a validade dos resultados dos testes em um contexto cultural e linguístico específico. A experiência com a administração de um instrumento em um contexto com um único idioma ou cultura pode ser útil na antecipação de problemas que podem ser esperados em um contexto multilíngue ou multicultural. Por exemplo, os administradores de testes mais experientes geralmente sabem quais aspectos da instrução podem ser difíceis para os participantes. Esses aspectos podem permanecer difíceis após a tradução ou adaptação. Aplicações de instrumentos em um novo contexto linguístico ou cultural também pode apresentar problemas não encontrados anteriormente em aplicações em uma única cultura.

Sugestões para a prática. É importante, com essa diretriz, antecipar os fatores potenciais que podem criar problemas na administração do teste. Os seguintes fatores precisam ser estudados para garantir a postura ética na administração do teste:

- A clareza das instruções de teste (incluindo a tradução dessas instruções), o mecanismo de resposta (por exemplo, a folha de respostas), o tempo de resposta (uma fonte comum de erro é a falta de tempo suficiente para que os participantes concluam); motivação para os participantes concluírem o teste, conhecimento sobre o objetivo do teste e como ele será pontuado.

A-2 (14) Especifique as condições de teste que devem ser seguidas rigorosamente em todas as populações de interesse.

Explicação. O objetivo desta diretriz é incentivar os autores de teste a estabelecer instruções e procedimentos relacionados (por exemplo, condições de teste, limites de tempo, etc.) que possam ser seguidos rigorosamente em todas as populações de interesse. Esta diretriz é principalmente destinada a incentivar os aplicadores de testes a seguirem as instruções padronizadas. Ao mesmo tempo, adequações podem ser especificadas para lidar com subgrupos especiais de indivíduos dentro de cada população, que podem precisar de adaptações no teste, tais como tempo adicional, impressão com letra maior, condições de administração mais silenciosas e assim por diante. No campo atual de testes, essa é conhecida como "acomodações para testagem". O objetivo dessas adequações não é aumentar as pontuações dos participantes, mas sim criar um ambiente de teste para que eles possam mostrar o que sentem, e o que sabem fazer.

Variações das condições de teste padronizadas devem ser anotadas, de modo que essas variações e seu impacto nas generalizações e interpretações possam ser consideradas.

Sugestões para prática. Essa diretriz pode em parte se sobrepor com A-1 (13), mas é reafirmada aqui para destacar a importância dos participantes fazerem o teste sob condições tão similares quanto possível. Isso é essencial se as pontuações das duas versões de idioma forem usadas de forma intercambiável. Aqui estão algumas sugestões:

- Instruções e procedimentos do teste devem ser adaptados e reescritos de forma padronizada, adequada à nova língua e cultura.

- Se as instruções e os procedimentos do teste forem alterados para as novas culturas, os administradores devem ser treinados de acordo com os novos procedimentos; eles devem ser informados a respeito desses procedimentos e não sobre a versão original.

Escala de Pontuação e Diretrizes de Interpretação

SSI-1 (15) Interprete quaisquer diferenças de pontuação de grupo com referência a todas as informações relevantes disponíveis.

Explicação. Mesmo que um teste tenha sido adaptado através de procedimentos tecnicamente sólidos, e a validade dos resultados dos testes já tenha sido estabelecida até certo ponto, deve-se ter em mente que as divergências entre os grupos podem ser interpretadas de muitas maneiras por causa de diferenças entre os países e/ou culturas participantes. Sireci (2005) revisou o procedimento utilizado para avaliar a equivalência de duas versões de idiomas diferentes de um teste. Foram administradas as versões de idioma separadas do teste a um grupo de pessoas que

são proficientes em ambos os idiomas (bilíngue) e que vêm do mesmo grupo cultural ou de idioma. Ele concebeu algumas opções de delineamento de pesquisa para estudos de equivalência usando participantes bilíngues, listou as possíveis variáveis de confusão que precisam ser controladas e ofereceu algumas sugestões valiosas para interpretar os resultados.

Sugestões para prática. Segue uma sugestão para melhorar a prática:

- Dependendo da questão de pesquisa (ou contexto para o qual as comparações de grupo são feitas), um número de possíveis interpretações pode ser considerado antes de definir apenas uma. Por exemplo, é importante descartar a motivação diferencial para um bom desempenho no teste antes de inferir que um grupo teve melhor desempenho no teste do que outro. Também pode haver variáveis de contexto que afetaram significativamente o desempenho do teste. Por exemplo, um grupo de pessoas pode simplesmente fazer parte de um sistema educacional menos eficaz, e isso teria um impacto significativo no desempenho do teste.

SSI-2 (16) Apenas compare as pontuações entre as populações quando o nível de invariância foi estabelecido na escala em que as pontuações são relatadas.

Explicação. Quando estudos comparativos entre grupos linguísticos e culturais são o foco central da iniciativa de tradução e adaptação, as versões multilíngues de um teste precisam ser colocadas em uma escala de relatório comum, e isso é realizado através de um processo chamado "*linking*" (ligação) ou "*equating*" (equacionamento). Isso requer tamanhos amostrais substanciais e evidências de que o viés de construção, método e item não estão presentes na versão adaptada do teste.

Van de Vijver e Poortinga (2005) delinearam vários níveis de equivalência de teste entre grupos linguísticos e culturais. O trabalho dos autores é especialmente útil para entender esse conceito, visto que o conceito foi apresentado pelos mesmos. Por exemplo, eles apontaram que a equivalência da unidade de medida requer que as escalas de relatório em cada grupo tenham a mesma métrica, garantindo assim, que as diferenças entre as pessoas dentro dos grupos tenham o mesmo significado. (Por exemplo, diferenças entre homens e mulheres em uma amostra chinesa podem ser comparadas a uma amostra francesa). No entanto, comparações válidas de escore direto só podem ser feitas quando as pontuações mostram o mais alto nível de equivalência, chamado equivalência escalar ou equivalência de escore total, o que requer que as escalas de cada grupo tenham a mesma unidade de medida e a mesma origem entre os grupos.

Numerosos métodos (tanto no âmbito da Teoria Clássica dos Testes como na Teoria de Resposta ao Item) foram apresentados para vincular ou equacionar pontuações de dois grupos (ou versões linguísticas de um teste). Os leitores interessados podem consultar Angoff (1984) e Kolen e Brennan (2004) para obter uma compreensão mais profunda deste tópico. Cook e Schmitt-Cascallar (2005) sugerem uma base para a compreensão de métodos estatísticos atualmente disponíveis para equacionar e dimensionar testes educacionais e psicológicos. Os autores descrevem e criticam os procedimentos de vinculação de escala específicos usados em estudos de adaptação de testes e ilustram procedimentos e questões de vinculação selecionados, descrevendo e criticando três estudos

que foram realizados nos últimos vinte anos para vincular os escores do *Scholastic Assessment Test* ao *Prueba te Aptitude Academica*.

Sugestões para Prática. O ponto chave aqui é que os resultados dos testes não devem ser super interpretados:

- Interprete os resultados com base no nível de evidência de validade disponível. Por exemplo, não faça declarações comparativas sobre os níveis de desempenho do participante nos dois grupos de idiomas, a menos que tenha sido estabelecida uma invariância de medida para as pontuações dos testes que estão sendo comparadas.

Diretrizes de Documentação

Doc-1 (17) Forneça documentação técnica de quaisquer alterações, incluindo uma descrição das evidências obtidas para apoiar a equivalência quando um teste é adaptado para uso em outra população.

Explicação. A importância dessa diretriz foi percebida e enfatizada por muitos pesquisadores (ver, por exemplo, Grisay, 2003). TIMSS e PISA foram muito bem-sucedidos em observar essa diretriz, documentando cuidadosamente as mudanças ao longo do trabalho de adaptação. Com essa informação, é possível focar na adequação das mudanças que foram feitas.

A documentação técnica também deve conter detalhes suficientes da metodologia para futuros pesquisadores replicarem os procedimentos usados na mesma ou em outras populações. Ela deve conter informações suficientes da evidência de equivalência de construção e equivalência de escala (se realizada) para apoiar o uso do instrumento na nova população. A documentação deve relatar a evidência usada para determinar o equacionamento de pontuações entre populações onde as comparações interpopulacionais forem feitas.

Às vezes, surge a pergunta sobre o público de destino da documentação técnica. A documentação deve ser escrita para o especialista técnico e para pessoas que serão obrigadas a avaliar a utilidade do teste para uso na nova ou em outras populações. (Um breve documento suplementar pode ser adicionado para o benefício de uma pessoa que não seja especialista).

Sugestões para a prática. Os testes adaptados devem apresentar um manual técnico que documente todas as evidências qualitativas e quantitativas associadas ao processo de adaptação. É especialmente útil documentar quaisquer alterações feitas para adequar o teste em um segundo idioma e cultura. Basicamente, técnicos e editores de periódicos vão querer uma documentação sobre o processo realizado para produzir e validar a versão do idioma de destino do teste. Eles também vão querer ver os resultados de todas as análises. Aqui estão os tipos de perguntas que precisam ser respondidas:

- Quais evidências estão disponíveis para apoiar a utilidade do construto e do teste adaptado na nova população?

- Quais dados foram coletados sobre os itens e de quais amostras?

- Quais outros dados foram obtidos para avaliar o conteúdo, a validade de critério e de construto?

- Como os vários conjuntos de dados foram analisados?

- Quais foram os resultados?

Doc-2 (18) Forneça aos usuários uma documentação que apoiará as boas práticas no uso de um teste adaptado com pessoas no contexto da nova população.

Explicação. A documentação deve ser escrita para as pessoas que usarão o teste em situações práticas de avaliação. Ela deve ser coerente com as boas práticas definidas pelas Diretrizes para Uso de Teste do *International Test Commission* (veja www.InTestCom.org).

Sugestões para a prática. O autor do teste deve fornecer informações específicas sobre as maneiras pelas quais os contextos sócio-culturais e ecológicos das populações podem afetar o desempenho no teste. O manual do usuário deve:

- Descrever o(s) construto(s) medido pelo teste e resumir a informação; descrever o processo de adaptação do teste.

- Resumir as evidências que apoiam a adaptação, incluindo evidências de adequação cultural do conteúdo do item, instruções de teste, formato de resposta, etc.

- Definir a adequação do uso do teste com vários subgrupos dentro da população e quaisquer outras restrições de uso.

- Explicar quaisquer problemas que precisem ser considerados em relação à boa prática na administração de testes.

- Explicar se, e como, as comparações entre populações podem ser feitas.

- Fornecer as informações necessárias para pontuar e corrigir tabelas relevante às normas de pesquisa ou descrever como os usuários podem acessar os procedimentos de pontuação (por exemplo, onde eles são computadorizados).

- Fornecer diretrizes para a interpretação dos resultados, incluindo informações sobre as implicações dos dados de validade e confiabilidade sobre as inferências que podem ser feitas nos resultados dos testes.

PALAVRAS FINAIS

Fizemos o nosso melhor para fornecer um conjunto de diretrizes para ajudar desenvolvedores e usuários de testes em seu trabalho. No entanto, para que as diretrizes e outros esforços para mudar as práticas deficitárias tenham efeito, deve haver bons mecanismos de disseminação em vigor. Uma recente revisão sistemática de Rios e Sireci (2014) demonstrou que a maioria dos projetos de adaptação de testes na literatura publicada não seguiu, de fato, as Diretrizes do ITC que estão disponíveis há cerca de 20 anos. Por isso, encorajamos os leitores a fazer todos os esforços para aumentar a conscientização entre os colegas desta Segunda Edição como fonte primária de melhores práticas, às quais tantos profissionais do mundo contribuíram.

Ao mesmo tempo, sabemos que, assim como a primeira edição dessas diretrizes foram substituídas, essas diretrizes da segunda edição também serão substituídas. Os conhecidos padrões de teste como AERA, APA e NCME estão agora em sua sexta edição (AERA, APA, & NCME, 2014). Esperamos que as Diretrizes do ITC para adaptação de testes passem por outra revisão também nos próximos anos. Se você souber de novos estudos que devam ser citados, ou que possam influenciar a terceira edição, ou ainda, se você quiser oferecer novas diretrizes ou revisões às 18 diretrizes apresentadas aqui, informe o ITC. Você pode entrar em contato com o atual presidente do comitê de Pesquisa e Diretrizes que produziu a segunda edição e / ou o secretário do ITC no endereço de e-mail encontrado em: www.InTestCom.org.

REFERÊNCIAS

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.

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.

APÊNDICE A. *CHECKLIST* DAS DIRETRIZES DO ITC PARA A TRADUÇÃO E ADAPTAÇÃO DE TESTES

Aqui está um *checklist* para lembrá-lo das 18 Diretrizes do ITC. Nós recomendamos que você marque aqueles que você acha que conseguiu atingir satisfatoriamente em seu projeto de tradução ou adaptação de testes e, em seguida, anote aqueles que ainda não foram abordados.

Diretrizes de Pré-Condição

[ ] **PC-1 (1) Obtenha as permissões necessárias do titular dos direitos de propriedade intelectual relativos ao teste antes de realizar qualquer adaptação.**

[ ] **PC-2 (2) Avalie que a quantidade de sobreposição na definição e conteúdo do construto medido pelo teste nas populações de interesse é suficiente para o uso pretendido (ou usos) dos resultados.**

[ ] **PC-3 (3) Minimize a influência de quaisquer diferenças culturais e linguísticas que são irrelevantes para os usos pretendidos do teste nas populações de interesse.**

Diretrizes para o Desenvolvimento do Teste

[ ] **TD-1 (4) Garanta que o processo de adaptação considere as diferenças linguísticas, psicológicas e culturais nas populações pretendidas, através da escolha de especialistas com experiência.**

[ ] **TD-2 (5) Utilize delineamentos e procedimentos de tradução apropriados para maximizar a adequação da adaptação de teste nas populações pretendidas.**

[ ] **TD-3 (6) Forneça evidência de que as instruções do teste e o conteúdo do item têm um significado semelhante para todas as populações pretendidas.**

[ ] **TD-4 (7) Forneça evidências de que os formatos de itens, escalas de classificação, categorias de pontuação, modos de administração e outros procedimentos são adequados para todas as populações pretendidas.**

[ ] **TD-5 (8) Colete dados de uma aplicação piloto do teste adaptado para permitir a análise de itens, a avaliação de confiabilidade e outros estudos de validade de pequena escala, para que quaisquer revisões necessárias ao teste adaptado possam ser feitas.**

Diretrizes de Confirmação

[ ] **C-1 (9) Selecione uma amostra com características que sejam relevantes para o uso pretendido do teste com tamanho e relevância suficientes para as análises empíricas**

[ ] **C-2 (10) Forneça evidência estatística relevante sobre a equivalência do construto, equivalência de método e equivalência de item para todas as populações pretendidas.**

[ ] **C-3 (11) Forneça evidências que apoiem as normas, a confiabilidade e a validade da versão adaptada do teste nas populações pretendidas.**

[ ] **C-4 (12) Use um delineamento de equacionamento apropriado e procedimentos de análise de dados ao vincular escalas de pontuação de diferentes versões de idioma de um teste.**

Diretrizes de Administração

[ ] **A-1 (13) Prepare os materiais e instruções de administração a fim de minimizar quaisquer problemas relacionados à cultura e à linguagem causados por procedimentos de administração e modos de resposta que possam afetar a validade das inferências extraídas das pontuações.**

[ ] **A-2 (14) Especifique as condições de teste que devem ser seguidas com rigor em todas as populações de interesse.**

Diretrizes de Interpretação e Escalas de Pontuação

[ ] **SSI-1 (15) Interprete quaisquer diferenças de pontuação de grupo tendo como referência todas as informações relevantes disponíveis.**

[ ] **SSI-2 (16) Compare pontuações entre populações apenas quando o nível de invariância for estabelecido na escala em que as pontuações são relatadas.**

Diretrizes de Documentação

[ ] **Doc-1 (17) Forneça a documentação técnica de quaisquer alterações realizadas, incluindo um relato das evidências obtidas para apoiar a equivalência, quando um teste é adaptado para uso em outra população.**

[ ] **Doc-2 (18) Forneça documentação para usuários de teste que visem dar suporte às boas prática no uso de um teste adaptado com pessoas no contexto da nova população.**

APÊNDICE B. GLOSSÁRIO DE TERMOS

Alpha (também chamado de "Coeficiente Alpha" ou "Alpha de Cronbach"). O coeficiente de confiabilidade de um teste cujos itens são adotados para medir um atributo em comum e têm discriminações iguais (portanto, é um caso especial do Ômega - veja abaixo). Em condições mais gerais, é um limite inferior para a confiabilidade.

Análise Fatorial Confirmatória. Uma hipótese sobre a estrutura de um teste é feita e, em seguida, análises são realizadas para obter a estrutura de teste da matriz a partir da correlação de itens no teste. Um teste estatístico é realizado para verificar se a estrutura hipotética e estimada do teste está próxima o suficiente para que a hipótese nula de que as duas estruturas sejam iguais não possa ser rejeitada.

Análise Fatorial Exploratória. A análise fatorial é um procedimento estatístico que é aplicado, por exemplo, com a matriz de correlação produzida pelas inter-correlações entre um conjunto de itens em um teste (ou um conjunto de testes). O objetivo é tentar explicar as inter-correlações entre os itens de teste (ou testes) em termos de um pequeno número de fatores que se acredita serem medidos pelo teste (ou testes). Por exemplo, com um teste de matemática, uma análise fatorial pode identificar o fato de que os itens se enquadram em três grupos - itens para cálculo, conceitos e solução de problemas. Pode-se dizer, então, que o teste de matemática está medindo três fatores - cálculos, conceitos matemáticos e resolução de problemas matemáticos.

Delineamento de Tradução Avançada. Com esse delineamento, um teste é adaptado para o idioma de interesse desejado por um tradutor ou, mais frequentemente, por um grupo de tradutores e, em seguida, um tradutor ou grupo de tradutores diferente avalia a equivalência das versões original e de destino do teste.

Delineamento de Tradução Reversa. Com esse delineamento, um teste é traduzido da versão do idioma de origem para a versão do idioma de destino por um grupo de tradutores, e a versão do idioma de destino é traduzida novamente para o idioma de origem, por um segundo tradutor ou grupo de tradutores. O modelo original e as versões de tradução reversa são comparadas, e um julgamento é feito sobre a adequação da versão do idioma original do teste. Se as duas versões do idioma original estiverem muito próximas, presume-se que a versão do idioma de destino do teste é aceitável.

Desenvolvimento Simultâneo de Teste. Desenvolvimento simultâneo de questionários em idioma de origem e de destino, utilizando procedimentos padronizados de controle de qualidade de tradução. Projetos internacionais de larga escala usam cada vez mais o desenvolvimento simultâneo para evitar o problema de que a versão desenvolvida em uma língua não possa ser traduzida / adaptada para

todas as línguas do estudo

Examinados. Usado de maneira intercambiável no campo da testagem como sinônimo de “participantes de um teste”, “candidatos” e “estudantes” (se estiver envolvido na realização de um teste).

Fórmula Kuder-Richardson 20. (Também chamada de "KR-20."). O coeficiente de confiabilidade de um teste formado a partir de itens binários, que são assumidos para medir um atributo em comum e têm discriminações iguais.

Funcionamento Diferencial do Item (DIF). Existe uma classe de procedimentos estatísticos que podem determinar se um item está funcionando mais ou menos da mesma forma em dois grupos diferentes. As comparações de desempenho são feitas, inicialmente, pela análise de correspondência dos examinandos sobre o traço medido pelo teste. Quando diferenças são observadas, diz-se que o item é potencialmente tendencioso. Um esforço é feito para explicar as diferenças condicionais no desempenho para os participantes nos dois grupos correspondentes na característica medida pelo item.

Modelagem de Equações Estruturais. Um conjunto de modelos estatísticos complexos que são usados para identificar a estrutura subjacente de um teste ou um conjunto de testes. Frequentemente, esses modelos são usados para investigar inferências causais sobre as relações entre um conjunto de variáveis.

PISA. Significa "Programa Internacional de Avaliação de Estudantes." Esta é a avaliação internacional de competência acadêmica patrocinada pela Organização para Cooperação e Desenvolvimento Econômico (OCDE) com mais de 40 países participantes.

Procedimento de Regressão Linguística para Identificação do Funcionamento Diferencial do Item. Este procedimento estatístico é mais uma maneira de realizar análises de DIF. Uma curva logística é ajustada aos dados de desempenho de cada grupo e, em seguida, as duas curvas logísticas, uma para cada grupo de idiomas, são comparadas estatisticamente.

Procedimento Mantel-Haenszel para Identificação do Funcionamento Diferencial do Item. Um procedimento estatístico para comparar o desempenho de dois grupos de examinandos em um item de teste. As comparações são feitas para os examinandos em cada grupo que são comparados na mesma característica ou construto medida pelo teste.

Ômega (também chamado de "Coeficiente de Ômega" ou "Ômega de McDonald"). O coeficiente de confiabilidade de um teste cujos itens são assumidos para medir um atributo em comum (ajuste ao modelo de fator geral). Geralmente mais aplicável do que o coeficiente Alpha.

Teoria de Resposta ao Item. É uma classe de modelos estatísticos para vincular

respostas de itens a um traço ou conjunto de traços que estão sendo medidos pelos itens no teste. Modelos específicos de TRI podem manipular dados de resposta binários e com mais de duas opções. Os dados binários podem vir da classificação de itens de múltipla escolha ou itens verdadeiro-falso em uma escala de personalidade. Os dados de resposta com mais de duas opções podem vir da pontuação de tarefas de desempenho ou ensaios em um teste de desempenho ou de escalas de classificação como "Likert".

Teste de Dimensionalidade. Isso se refere ao número de dimensões ou fatores que um teste está medindo. Muitas vezes, essa análise é realizada estatisticamente usando um dos muitos procedimentos, incluindo gráficos *eingvalue* ou modelagem de equações estruturais.

Teste de Pontuação Equivalente. Um procedimento estatístico para vincular pontuações em dois testes que medem o mesmo construto, mas no qual os testes não são estritamente paralelos.

TIMSS. Significa "*Trends in International Mathematics and Science Studies*" e é uma avaliação internacional de alunos de 4, 8 e 12 anos em países nas áreas de matemática e ciências e patrocinada pela AIE.

Tradução Dupla e Reconciliação. Neste delineamento de tradução, um tradutor independente ou um grupo de especialistas identifica e resolve qualquer discrepância entre as traduções alternativas e reconcilia-as em uma única versão.

Valores Delta. Os valores delta são simplesmente valores não lineares de *p* transformados e aplicados a itens com pontuação binária. Um valor delta de item é o desvio normal correspondente à área sob uma distribuição normal (média = 0,0, DP = 1,0), em que a área sob a distribuição normal é igual à proporção de candidatos que respondem corretamente ao item. Portanto, se $p = 0,84$, o valor delta para o item seria -1,0. Essa transformação é feita com a crença de que os valores delta têm maior probabilidade de estar em uma escala de intervalo igual a valores *p*.

Versão do Idioma de Destino. O idioma para a qual um teste é traduzido / adaptado. Por exemplo, se um teste for traduzido do inglês para o espanhol, a versão em inglês é geralmente chamada de "versão do idioma de origem" e a versão em espanhol é chamada de "versão do idioma de destino".

Versão do Idioma de Origem. O idioma no qual o teste original foi escrito.

WDMS. Significa "Escalonamento Multidimensional Ponderado" É outro procedimento estatístico para estudar a dimensionalidade do teste.