Recommandations de l'ITC pour la traduction et l'adaptation de tests (seconde édition)

K. Gana, Guillaume Broc, N.E. Boudouda, N. Calcagni, S. Ben Youssef

Recommandations de l'ITC pour la traduction et l'adaptation de tests (Seconde édition)

K.Gana

G.Broc

N.E.Boudouda

N.Calcagni

S.Ben Youssef

The ITC Guidelines for Translating and Adapting Tests (Second edition)

International Test Commission. (2017). *The ITC Guidelines for Translating and Adapting Tests* (Second edition). [www.InTestCom.org]

TÉMOIGNAGE DE RECONNAISSANCE

Le Conseil de la Commission Internationale des Tests (ITC) souhaite remercier le comité de six personnes qui a travaillé pendant plusieurs années à la production de la deuxième édition des *Recommandations pour la traduction et l'adaptation de tests* : David Bartram, SHL, Royaume-Uni; Giray Berberoglu,

Université Technique du Moyen-Orient, Turquie; Jacques Grégoire, Université Catholique de Louvain, Belgique; Ronald Hambleton, Président du comité, Université de Massachusetts Amherst, États-Unis; Jose Muniz, Université d'Oviedo, Espagne; et Fons van de Vijver, Université de Tilburg, Pays-Bas.

La Commission Internationale des Tests souhaite également remercier Chad Buckendahl (États-Unis), Anne Herrmann et ses collègues de l'OPP Ltd. (Royaume-Uni) et April Zenisky de l'Université du Massachusetts (États-Unis) pour leur lecture attentive d'une version antérieure du document. L'ITC remercie également tous les autres experts du monde entier qui ont contribué, directement ou indirectement, à la deuxième édition des *Recommandations* de l'ITC pour la traduction et l'adaptation de tests.

Résumé

La seconde édition des recommandations de l'ITC pour la traduction et l'adaptation de tests a été élaborée entre 2005 et 2015 afin de mettre à jour la première édition, et de répondre aux progrès des techniques et pratiques du testing. Les 18 recommandations sont organisées en six sections afin de faciliter leur utilisation : Conditions préalables (3), élaboration de test (5), Confirmation/validation (4),

administration de test (2), cotation et interprétation (2) et documentation (2). Pour chaque recommandation, une explication est fournie ainsi que des suggestions pour la mise en pratique. Une liste récapitulative (checklist) est également fournie pour améliorer la mise en œuvre des recommandations.

CONTEXTE GENERAL

Le domaine de la traduction et de la méthodologie d'adaptation de tests a connu un progrès fulgurant au cours des 25 dernières années, comme en témoigne la publication de plusieurs ouvrages et de nombreuses études et exemples de travaux exceptionnels d'adaptation de tests (voir, par exemple, van de Vijver & Leung, 1997, 2000; Hambleton, Merenda, & Spielberger, 2005 ; Grégoire & Hambleton, 2009 ; Rios & Sireci, 2014). Ces progrès ont été nécessaires en raison de l'intérêt croissant pour (1) la psychologie interculturelle, (2) les études comparatives internationales menées à grande échelle sur la réussite scolaire (par exemple, TIMSS et OCDE/PISA), (3) les tests d'accréditation utilisés dans le monde entier (par exemple, dans le domaine de l'informatique par des entreprises comme Microsoft et Cisco), et (4) les considérations relatives à l'équité dans le testing qui permet aux candidats de faire le choix de la langue dans laquelle ils seront évalués (par exemple, pour les inscriptions universitaires en Israël, les candidats peuvent passer plusieurs tests dans une des six langues proposées).

Des avancées techniques ont été réalisées dans les approches qualitatives et quantitatives concernant la mesure des construits psychologiques, les méthodes et les divers biais dans les tests et questionnaires adaptés, y compris l'utilisation de procédures statistiques complexes comme la théorie de réponse à l'item, la modélisation par équations structurales et la théorie de la généralisabilité (voir Hambleton et al., 2005; Byrne, 2008). L'OCDE/PISA a proposé de nouveaux modèles de traduction (voir, Grisay, 2003); un ensemble de démarches a été proposé pour mener à bien un projet d'adaptation de tests (voir, par exemple, Hambleton et Patsula, 1999; des projets exemplaires sont disponibles pour guider les pratiques d'adaptation de tests - par exemple, les projets OCDE/PISA et TIMSS), et bien d'autres progrès ont été réalisés.

La première édition de ces recommandations (voir van de Vijver & Hambleton, 1996 ; Hambleton, 2005) est née sous l'angle d'une perspective comparative, reflet du but de l'adaptation de tests, à savoir permettre ou faciliter les comparaisons entre groupes de répondants. Le modèle sous-jacent auquel ces recommandations étaient destinées se basait sur le développement successif d'instruments dans un

contexte comparatif (l'instrument existant doit être adapté pour son usage dans un nouveau contexte culturel). Cependant, il est de plus en plus évident que l'adaptation de tests couvre un champ d'application plus vaste. L'exemple le plus frappant est l'usage d'outils, nouveaux ou préexistants, au sein de groupes multiculturels, tels que les consultations de clients issus de différents groupes ethniques, les examens académiques auprès de divers groupes ethniques maîtrisant différemment la langue d'évaluation, ou encore les recrutements internationaux pour des postes de management dans des entreprises multinationales. Cette évolution dans les champs d'applications n'est pas sans conséquences sur l'élaboration, l'administration, la validation et la documentation d'outils de mesure. Une des répercussions potentielles pourraient être la nécessité d'adapter les items d'un test existant afin d'accroître sa compréhension par les participants pour qui la langue du test n'est pas la langue maternelle (par exemple, en simplifiant la langue). Une autre extension importante des recommandations consisterait à permettre le développement simultané d'outils (c'est-à-dire le développement combiné de questionnaires dans les langues sources et cibles). Les recherches internationales à grande échelle ont de plus en plus recours à l'élaboration simultanée de tests et ce, afin d'éviter que l'outil développé dans une certaine langue ne puisse être traduit ou adapté à toutes les autres langues de l'étude.

La première édition des recommandations de l'ITC pour la traduction et l'adaptation de tests a été publiée par van de Vijver et Hambleton (1996), par Hambleton (2002), et par Hambleton, Merenda et Spielberger (2005). Cette version n'a connu que quelques modifications mineures au niveau rédactionnel entre 1996 et 2005. Entre-temps, de nombreux progrès ont été réalisés. Il y a eu, premièrement, un certain nombre de commentaires judicieux concernant les recommandations de l'ITC, comprenant les articles de Jeanrie et Bertrand (1999), Tanzer et Sim (1999), et Hambleton (2002), qui ont souligné la qualité des recommandations, tout en proposant aussi des pistes d'améliorations. Hambleton, Merenda et Spielberger (2005) ont publié les actes d'une conférence internationale de l'ITC tenue en 1999 à l'Université de Georgetown aux Etats-Unis. Nombreux sont les auteurs de ces actes, parmi lesquels Cook et Schmitt-Cascallar (2005) et Sireci (2005), qui ont proposé de nouveaux paradigmes et de nouvelles méthodologies pour l'adaptation de tests.

En 2006, l'ITC a organisé une conférence internationale à Bruxelles (Belgique), en vue de se pencher sur les recommandations de l'ITC pour la traduction et l'adaptation de tests. Plus de 400 participants de plus de 40 pays différents se sont penchés sur la question de l'adaptation de tests, de nombreuses nouvelles orientations méthodologiques ont été avancées, de nouvelles recommandations ont été également suggérées et des exemples de mise en œuvre réussie ont été partagés. Les communications présentées aux symposiums des rencontres internationales de 1996 à 2009 ont été nombreuses (voir, par exemple,

Grégoire & Hambleton, 2009; et nous invitons les lecteurs à lire le papier de Muniz, Elosua et Hambleton (2013) relatif à la version préliminaire en espagnol de la deuxième édition des recommandations de l'ITC).

En 2007, le Conseil de l'ITC (ITC Council) a mis en place un comité de six personnes lui confiant la tâche de de la mise à jour des recommandations de l'ITC afin de mettre à profit les nouvelles avancées techniques et les nombreuses expériences acquises par les chercheurs dans ce domaine. Ces avancées comprennent (1) l'essor de la modélisation par équations structurales permettant d'évaluer l'équivalence factorielle d'un test à travers de différents groupes linguistiques, (2) des approches nouvelles permettant de déterminer l'ampleur du fonctionnement différentiel des items polytomiques dans différents groupes linguistiques et (3) de nouveaux modèles d'adaptation de tests mis au point par des projets internationaux d'évaluation comme OECD/PISA et TIMSS. Le comité a également présenté et rédigé les versions préliminaires des recommandations lors des rencontres internationales des psychologues à Prague (en 2008) et Oslo (en 2009), suscitant de nombreux commentaires forts utiles.

La section des recommandations relative à l'Administration de test a été conservée dans la seconde édition. Toutefois, et afin d'éviter les redondances, le nombre total de recommandations dans cette section a été ramené de six à deux. La dernière section de la première édition était dédiée à la "Documentation/interprétations des scores". Dans la deuxième édition, nous l'avons divisé en deux sections distinctes - l'une ayant trait aux scores et leurs interprétations, et l'autre à la documentation sur le test. De plus, deux des quatre recommandations initiales de cette section ont fait l'objet de modifications substantielles.

Pareillement à la première édition, nous voulions que les lecteurs soient au clair quant à la distinction entre la traduction du test et son adaptation. La traduction désigne probablement le terme le plus couramment utilisé, mais l'adaptation de test est un terme plus large désignant la transposition d'un test d'une langue et d'une culture à une autre. L'adaptation de test fait référence à une procédure comprenant des activités telles que: déterminer si un test dans une seconde langue et culture peut ou non mesurer le même construit que la langue d'origine; choisir des traducteurs; choisir une méthode pour évaluer le travail des traducteurs du test (p. ex., traductions bidirectionnelles/à rebours/inversées; traduction unidirectionnelle); décider de toutes les adaptations nécessaires; modifier le format du test; superviser la traduction; vérifier l'équivalence linguistique du test et mener les études de validité nécessaires et supplémentaires. En revanche, la traduction d'un test a un sens plus restreint, limité au choix de faire

passer le test d'une langue et d'une culture à une autre en veillant à préserver le sens linguistique. La traduction d'un test n'est qu'une partie du processus d'adaptation, considérée, en tant que tel, comme une approche très simpliste de transposition d'un test d'une langue à une autre sans égard pour les équivalences académiques ou psychologiques.

LES RECOMMANDATIONS

Introduction

Les recommandations sont définies dans le présent document comme des lignes directrices guidant la pratique de réalisation et d'évaluation de l'adaptation de tests ou de développement simultané de tests psychologiques et éducatifs en vue de les utiliser auprès de différentes populations. Dans le texte qui suit, 18 recommandations sont organisées autour de six grandes sections: Conditions préalables (3), Développement/élaboration de test (5), Confirmation/Validation [analyses empiriques] (4), Administration de test (2), Cotation et Interprétation (2), et Documentation sur le test (2).
La première section intitulée "Conditions préalables" insiste sur le fait que des décisions importantes doivent être prises avant de commencer tout processus de traduction/adaptation de tests. La deuxième section "Recommandations pour l'élaboration de tests" est axée sur le processus d'adaptation d'un test. La troisième section "Confirmation/Validation" comprend les recommandations ayant trait à la compilation de preuves empiriques portant sur l'équivalence linguistique, la fiabilité et la validité d'un test dans plusieurs langues et cultures. Les trois dernières sections portent sur "L'administration du test", "La cotation et interprétation des scores" et "La documentation sur le test". La documentation a été un aspect particulièrement négligé dans les initiatives d'adaptation de tests en psychologie et en éducation. Aussi, nous souhaiterions que les rédacteurs de revues scientifiques et les organismes de financement exigent davantage de documentation sur les processus d'adaptation des tests.
Pour chaque recommandation nous avons fourni des explications et des suggestions pour sa mise en œuvre dans la pratique.

Recommandations concernant les Conditions Préalables (CP)

CP-1 (1) Obtenir l'autorisation nécessaire auprès du titulaire des droits de propriété intellectuelle du test avant d'entreprendre son adaptation.

Explication. Les droits de propriété intellectuelle renvoient à un ensemble de droits que les individus ont sur leurs propres créations, inventions ou productions. Il s'agit de protéger les intérêts des créateurs en leur accordant des droits moraux et économiques sur leurs propres créations. Selon l'Organisation mondiale de la propriété intellectuelle (www.wipo.int),
"la législation sur le droit d'auteur fait partie du secteur juridique et
plus largement de la propriété intellectuelle, qui vise d'une manière générale à protéger les œuvres de l'esprit. Les droits de propriété intellectuelle protègent les intérêts des innovateurs
et des créateurs en leur conférant des droits sur leurs œuvres".
Il existe deux branches de la propriété intellectuelle : Propriété industrielle et droits d'auteur (copyright). Le premier concerne les brevets protégeant les inventions, les dessins et modèles industriels, les marques de produits et les noms commerciaux. Le droit d'auteur se rapporte aux créations artistiques (y compris les œuvres fondées sur la technologie) et littéraires. Le créateur (l'auteur) a des droits spécifiques sur sa création (e.g., la prévention de certaines distorsions lorsqu'elle est copiée ou adaptée). D'autres droits (e.g., faire des copies) peuvent être exercés par d'autres personnes (e.g., un éditeur) qui ont obtenu une licence de l'auteur ou du détenteur du droit d'auteur. Pour de nombreux tests, comme pour d'autres œuvres écrites, le droit d'auteur peut être cédé par l'auteur à l'éditeur ou au distributeur.

Étant clairement considérés comme des créations de l'esprit humain, les tests éducatifs et psychologiques sont protégés par des droits de propriété intellectuelle. Souvent, le droit d'auteur ne concerne pas le contenu spécifique des items (e.g., personne n'a de droit sur des items tels que "1+1 = ..." ou "je me sens triste"), mais plutôt l'organisation originale du test (structure du test, système de cotation, organisation du matériel du test, etc.). Par conséquent, imiter un test existant, c'est-à-dire conserver la structure originale et son système de cotation tout en créant de nouveaux items, constitue une violation des droits de propriété intellectuelle du test original. Lorsqu'il est autorisé à procéder à une adaptation, l'auteur de l'adaptation du test doit respecter les caractéristiques originales du test (structure, matériel, format, cotation...), à moins qu'un accord du titulaire de la propriété intellectuelle ne permette d'en modifier les caractéristiques.

Suggestions pour la pratique. Les auteurs de l'adaptation d'un test devraient respecter toute les lois et tous les accords sur le droit d'auteur protégeant le test original. Ils doivent obtenir un accord signé par le détenteur de la propriété intellectuelle (c.-à-d. l'auteur ou l'éditeur) avant d'entreprendre l'adaptation d'un test. L'accord devrait préciser les modifications acceptables concernant les caractéristiques du test original

et devrait préciser le détenteur des droits de propriété intellectuelle de la version adaptée.

CP-2 (2) Évaluer auprès de la population cible que le degré d'adéquation/concordance entre la définition et le contenu du construit mesuré par le test original et le contenu de chaque item soit suffisant en vue de l'utilisation prévue (ou les utilisations prévues) des scores au test.

Explication. Cette recommandation exige que l'objet évalué soit compris de la même manière par tous les groupes linguistiques et culturels en lice/en présence, ce qui constitue le fondement de comparaisons interculturelles valides. A ce stade de la procédure, le test ou l'instrument de mesure n'est pas encore adapté. Aussi, il serait souhaitable de compiler des preuves empiriques en se documentant sur les tests similaires, et d'évaluer l'adéquation/concordance construit-item pour les groupes linguistiques cibles de l'étude. Cependant, cette recommandation doit être évaluée à l'aide de données empiriques, conformément aux preuves exigées dans C-2 (10). Le but de toute analyse n'étant pas de déterminer la structure d'un test, bien qu'il s'agisse d'un élément important de toute analyse, mais de confirmer l'invariance de la structure à travers les différentes versions linguistiques du test.

Suggestions pour la pratique. Il convient de recruter des experts connaissant aussi bien le construit mesuré que les groupes culturels cibles afin d'évaluer la pertinence/bien-fondé du construit auprès de chacun de ces groupes. Ces experts tenterons de répondre à la question suivante: le construit a-t-il une signification dans chaque culture? Il arrive, nous l'avons vu à maintes reprises dans les tests éducatifs, par exemple, qu'un comité juge que le construit mesuré par un test n'ait pas de sens ou qu'il ait perdu du sens dans une culture différente (par exemple, la qualité de vie, la dépression ou l'intelligence). Des méthodes telles que les groupes de discussion (focus groups), les entretiens et les enquêtes peuvent être utilisées pour obtenir un corpus d'information sur le degré de concordance interculturelle du construit.

CP-3 (3) Réduire drastiquement l'influence des différences culturelles et linguistiques préjudiciables/non désirables/inutiles à l'utilisation prévue du test dans les populations cibles.

Explication. Les caractéristiques culturelles et linguistiques non pertinentes aux variables que le test est censé mesurer doivent être identifiées dès le début du projet. Elles peuvent avoir trait au format des items, au matériel (par exemple l'utilisation d'un ordinateur, d'images ou d'idéogrammes...), à la durée de passation, etc.

L'approche pour y parvenir consiste à évaluer la "distance linguistique et culturelle" entre la langue source et la langue cible du test. L'évaluation de la distance linguistique et culturelle peut inclure des considérations relatives aux différences de langues (linguistiques), de structure familiale, de religion, de style de vie et de valeurs (van de Vijver & Leung, 1997).
Cette recommandation repose principalement sur des méthodes qualitatives et fait appel à des spécialistes familiers avec une recherche sur des différences culturelles et linguistiques spécifiques. Elle met particulièrement l'accent sur le choix des traducteurs de test et exige que ces derniers soient natifs de la langue et de la culture cible, car la simple connaissance de la langue cible ne suffit pas pour parvenir à ~~pouvoir~~ identifier les sources possibles de biais de méthode. Par exemple, dans l'étude comparative sino-américaine de Hambleton, Yu et Slater (1999) sur les connaissances en mathématiques des élèves de quatrième, on a relevé des problèmes de format et de longueur du test, ainsi qu'une foule de caractéristiques culturelles associées au test de mathématiques conçu pour des élèves de quatrième.

Suggestions pour la pratique. Il s'agit d'une recommandation difficile à satisfaire à tout moment à l'aide de données empiriques. Ceci est particulièrement vrai aux premières étapes de l'adaptation de tests. Toutefois, il est souvent possible de recueillir des preuves qualitatives :

- Que ce soit par l'entremise de l'observation, l'entretien, le focus group ou encore l'enquête, il s'agit ici de déterminer le niveau de motivation des participants, leur compréhension des consignes, leur expérience des tests psychologiques, la rapidité de passation de tests, la familiarité avec les échelles de réponses et les différences culturelles (combien même ces comparaisons pourraient être problématiques en raison des différences culturelles dans la compréhension des variables elles-mêmes). Lorsque la collecte de telles données auprès des participants se révèle problématique, il convient d'obtenir autant d'informations que possible auprès des traducteurs. Une partie de ce travail pourrait être effectuée avant de progresser dans l'adaptation du test.

- Il demeure possible de tenir compte de ces "variables parasites" dans toute analyse empirique subséquente une fois que le test aura été adapté et qu'il est prêt pour à être soumis à des études de validation. cette démarche est réalisé au moyen d'analyse de covariance ou d'autres analyses qui permettent de contrôler, entre autres, le niveau de motivation ou la familiarité avec une échelle de réponse particulière auprès de participants issus de langues et cultures différentes (voir Johnson, 2003 ; Javaras & Ripley, 2007).

ET-1 (4) Veiller à ce que la procédure de traduction et d'adaptation tient compte des différences linguistiques, psychologiques et culturelles des populations visées en choisissant des experts possédant l'expertise nécessaire.

Explication. Il s'agit de l'une des recommandations qui au fil des années a eu le plus d'impact, car de nombreuses données probantes indiquent qu'elle a eu une influence considérable sur la recherche par les agences d'évaluation de traducteurs possédant des qualifications dépassant la simple connaissance des deux langues impliquées dans l'adaptation du test (voir, par exemple, Grisay, 2003). La connaissance des cultures, et au moins la connaissance générale du construit en question et de la construction de tests, font désormais partie des critères de sélection des traducteurs. De plus, cette recommandation semble avoir joué un rôle important en encourageant les organismes de traduction et d'adaptation de tests à faire appel à au moins deux traducteurs en fonction du modèle retenu (p. ex., modèle de la traduction bidirectionnelle/inversée). L'ancienne pratique consistant à s'en remettre pour toutes les décisions à un seul traducteur, aussi qualifié soit-il, a disparu de la liste des pratiques acceptables aujourd'hui.

La connaissance/expertise de la culture cible amène à faire appel à des traducteurs dont la langue cible est la langue maternelle, vivant de préférence dans la région cible. Un autochtone ne produira pas seulement une traduction précise, mais également une traduction facile à lire et en phase avec le contexte local. Vivre dans la région cible, assurera la mise à jour en matière de l'usage courant d'une langue.

Notre définition d'un "expert" est donc une personne ou une équipe ayant des connaissances combinées suffisantes (1) des langues concernées, (2) des cultures concernées, (3) du contenu du test concerné, et (4) des principes généraux du testing, le but étant de produire une traduction/adaptation professionnelle de qualité. Dans la pratique, il serait judicieux de faire appel à des équipes ayant des qualifications différentes (par exemple, des traducteurs avec et d'autres sans expertise dans le domaine spécifique du test, un expert en tests, etc.) afin d'identifier les aspects susceptibles d'être négligés par les uns et les autres. Dans tous les cas, la connaissance des principes généraux du testing, en plus de la connaissance du contenu de tests, devrait faire partie intégrante de la formation des traducteurs.

Suggestions pour la pratique. Nous suggérons ce qui suit :

- Il convient de choisir des traducteurs locaux dont la langue maternelle est impérativement la langue cible et qui ont une connaissance approfondie de la culture à laquelle le test est adapté. Une erreur

courante consiste à qualifier comme traducteurs des personnes qui connaissent la langue, mais pas très bien la culture locale, car une connaissance approfondie de la culture est souvent essentielle pour garantir l'équivalence culturelle des versions du test. Posséder ces connaissances culturelles permettra d'identifier des références culturelles (par exemple, le cricket, la Tour Eiffel, le Président Lincoln, le kangourou, etc.) dont les participants auxquels s'adressent l'adaptation ne sont peut-être pas familiers.

- Choisir des traducteurs possédant, si possible, de l'expérience en lien avec le contenu du test et connaissant les principes du testing (p. ex. concernant les items à choix multiples, la bonne réponse devrait être ni plus longue ni plus courte que les autres, choix de distracteurs. Les indices grammaticaux ne devraient pas suggérer la bonne réponse; concernant les items vrai/faux, les énoncés « vrais » ne devraient pas être ostensiblement plus longs que les énoncés « faux »).
- Dans la pratique il est quasiment impossible de trouver des traducteurs ayant une connaissance des principes de construction/développement de tests. Il est donc, essentiel de former les traducteurs aux principes de base, en matière de rédaction d'items et de format d'item avec lequel ils vont travailler. En l'absence de formation, des traducteurs parfois trop consciencieux deviennent sources d'erreur de nature à compromettre la validité du test traduit. Par exemple, un traducteur peut parfois ajouter une remarque de clarification induisant la bonne réponse. Ce faisant, le traducteur peut rendre la question plus facile que prévu, ou une bonne réponse d'un QCM plus longue fournissant ainsi, un indice aux candidats qui passent le test.

ET-2 (5) Utiliser des designs et procédures de traductions appropriées, pour maximiser l'adéquation/convenance de l'adaptation du test aux populations cibles.

Explication. Cette recommandation exige que les décisions prises par les traducteurs ou les groupes de traducteurs maximisent la convenance de la version adaptée à la population visée. Cela signifie que les mots employés doivent être naturels et acceptables et ce, en privilégiant l'équivalence linguistique fonctionnelle plutôt que l'équivalence littérale. Les modèles de traduction les plus populaires pour atteindre ces objectifs sont les traductions unidirectionnelles et les traductions bidirectionnelles/ à rebours/inversées. Brislin (1986) et Hambleton et Patsula (1999) présentent une analyse complète des deux modèles, comprenant leurs définitions, leurs forces et leurs faiblesses. Toutefois, il convient de noter que les deux modèles ont des limites, fournissant rarement de preuves pour valider un test traduit et adapté. Le principal inconvénient de la traduction bidirectionnelle, si elle est réalisée sous sa forme la plus stricte, est qu'elle exclut de facto tout examen/vérification de la version traduite. Cette

procédure génère trop souvent une version traduite du test, maximisant la facilité de la traduction inversée/à rebours, en aboutissant parfois à des traductions maladroites.

Une procédure de double traduction et de recoupement vise à remédier aux insuffisances et aux risques inhérents aux traductions uniques. Dans cette approche, un troisième traducteur indépendant ou un panel d'experts identifie et résout toutes les divergences entre les traductions alternatives et les unifie en une seule version. Dans les programmes d'évaluation transculturelle à grande échelle tel que PISA, deux versions de langues différentes (par exemple, l'anglais et le français) peuvent être utilisées comme sources distinctes, générant deux traductions, qui seront ensuite rassemblées en une seule version linguistique cible (Grisay, 2003). Cette approche offre d'importants avantages tels que l'identification et l'examen direct des divergences éventuelles dans la langue cible. De plus, utiliser plus qu'une langue source aide à minimiser l'impact des caractéristiques culturelles.

Les différences de structure linguistique peuvent induire des problèmes dans la traduction de tests. Par exemple, dans une échelle bien connue développée en anglais par Rotter et Rafferty (1950), les items sont des phrases à compléter : "J'aime..." ; "Je regrette..." ; "Je ne peux...".
Cependant, le même format s'avère inapproprié en langue turque, où l'objet d'une phrase doit précéder le verbe et le sujet. L'utilisation de phrases à compléter, comme dans la version anglaise, modifierait donc complètement le comportement de réponse, puisque les élèves turcs devraient d'abord regarder la fin de l'énoncé avant d'en remplir le début.
Quel que soit la solution retenue à ce problème, la version traduite (c.-à-d. la langue cible) sera quelque peu différente de la version issue de la langue source en termes de spécificité du format.

Suggestions pour la pratique. La compilation des jugements des experts semble particulièrement utile pour vérifier/s'assurer que cette recommandation est respectée :
- Utiliser les échelles d'évaluation proposées par Brislin (1986), Jeanrie et Bertrand (1999) ou Hambleton et Zenisky (2010). Hambleton et Zenisky mettent à disposition une liste validée empiriquement de 25 caractéristiques d'un test traduit qui doivent être vérifiées pendant le processus d'adaptation. En voici quelques exemples de questions de cette liste : «les mots de l'item traduit présentent-ils les mêmes difficultés et des similitudes en comparaison avec les mots de l'item dans sa version originale?» et «la traduction introduit-elle des changements dans le texte (omissions, substitutions ou additions) qui pourraient impacter la difficulté de l'item du test dans les deux versions linguistiques? »
- Utiliser, si possible, plusieurs modèles de traduction. Par exemple, un plan de traduction inversée

peut être utilisé pour revérifier la version créée par un panel d'experts à l'issue d'une double traduction et une synthèse des deux.

- Si un test est destiné à un usage transculturel, il convient d'envisager dès le début le développement simultané de versions multilingues, afin d'éviter d'éventuels problèmes de traduction/adaptation de la version source. On trouvera, par exemple, dans Solano-Flores, Trumbull et Nelson-Barber (2002) plus de détails sur l'élaboration simultanée de tests. Il convient tout au moins, d'élaborer une version source, pour faciliter les traductions éventuelles, permettant d'éviter, autant que possible, les problèmes potentiels. En particulier, il convient d'éviter les références culturelles, les items idiosyncrasiques/dialogiques/singuliers et les formats de réponse inappropriés, etc.
- Compte tenu des différences de syntaxe entre les langues, l'utilisation de formats qui reposent sur une structure rigide des phrases ; devrait être évitée dans les évaluations internationales à grande échelle et probablement aussi dans les tests psychologiques et ce, en raison des problèmes de traduction que ces formats rencontrent.

ET-3 (6) Fournir les preuves que la consigne du test et le contenu des items ont la même signification pour toutes les populations visées.

Explication. Les preuves exigées dans le cadre de cette recommandation peuvent être recueillies au moyen de diverses stratégies (voir, par exemple, van de Vijver et Tanzer, 1997) qui comprennent (1) le recours à des experts autochtones ; (2) l'utilisation d'échantillons de répondants bilingues ; (3) l'utilisation d'enquêtes locales pour évaluer le test ; et (4) l'utilisation d'administrations non-standardisées du test afin d'en accroître l'acceptabilité et la validité.
Réaliser une étude préliminaire/pilote de la version adaptée du test est une bonne idée. Cette étude consiste à effectuer non seulement l'administration du test et l'analyse des données, mais aussi et surtout, des entretiens avec aussi bien ceux qui administrent le test que les participants au test afin recueillir leurs jugements concernant le test. D'autres approches sont envisageables telle que faire appel à des spécialistes issus de différentes langues ou à des spécialistes bilingues pour expertiser le contenu. Par exemple, on pourrait demander à des spécialistes bilingues d'évaluer la similitude de la difficulté due au format d'items et au contenu de chacune des versions. L'entretien cognitif est une autre méthode prometteuse (Levin et coll., 2009).

Suggestions pour la pratique. Plusieurs suggestions ont été présentées ci-dessus pour satisfaire cette

recommandation. En voici quelques exemples:

- Recourir à des experts autochtones issus de la culture et de la langue locale pour évaluer la traduction/adaptation du test.
- Utiliser des échantillons de répondants bilingues afin de recueillir des suggestions sur l'équivalence des deux versions du test, tant sur les consignes que sur les items du test.
- Réaliser des enquêtes locales pour évaluer le test. Ces études préliminaires peuvent être fort utiles. Assurez-vous d'interroger après la passation du test aussi bien ceux qui l'administrent que ceux qui y répondent, car leurs commentaires sont souvent plus précieux que les simples réponses des participants aux items du test.
- Adapter l'administration du test pour accroître son acceptabilité et sa validité.
- Se conformer à des consignes identiques n'a aucun sens, si elles sont mal comprises par les répondants de la deuxième langue/groupe culturel.

ET-4 (7) Démontrer que les formats d'items, les échelles de réponse/cotation, les catégories de cotation, les conventions relatives aux tests, les modes d'administration et toute autre procédure conviennent à toutes les populations visées.

Explication. Les formats d'items tels que les échelles de réponse à cinq points ou les nouveaux formats d'item tel que "glisser-déposer" ou "répondre à tout ce qui est correct" ou même "choisir une seule réponse" peuvent être source de confusion pour les répondants qui n'ont jamais rencontré ces formats d'items. Même la mise en page des items, l'utilisation de graphiques ou l'émergence rapide de formats d'items informatisés peuvent être source de confusion pour les candidats. Il existe de nombreux exemples de ce type d'erreurs aux États-Unis qui ont pris l'initiative d'informatiser une grande partie des tests standardisés pour enfants. Grâce aux items d'entraînement, les problèmes peuvent être surmontés pour la plupart des enfants. Ces nouveaux formats d'items doivent être familiers aux répondants sous peine d'introduire une source de biais de nature à fausser les résultats aux tests.

Suggestions pour la pratique. Les preuves fondées sur des données qualitatives et quantitatives ont toutes deux un rôle à jouer dans la prise en compte de cette recommandation. Il y a plusieurs caractéristiques d'un test adapté qui appellent à une vérification:

- Vérifiez que les items d'entraînement sont suffisants pour amener les répondants au niveau requis, leur permettant de fournir des réponses honnêtes et/ou des réponses qui reflètent leur niveau de maîtrise du matériel du test.

- S'assurer que les répondants connaissent les nouveaux formats d'item ou les nouveaux modes d'administration du test (tels que les tests informatisés) qui font désormais partie des procédures de testing.
- Vérifier que toutes les conventions de tests (p. ex. le placement des illustrations ou le marquage des réponses sur une feuille de réponses) soient claires pour les répondants.
- Ici aussi, la grille d'évaluation fournie par Jeanrie et Bertrand (1999) et Hambleton et Zenisky (2010) peut servir. Par exemple, Hambleton et Zenisky ont inclus des questions telles que "Est-ce que le format de l'item, y compris sa présentation matérielle, est le même dans les deux versions linguistiques ? et "Si une forme d'accentuation de mots ou de phrases (gras, italique, souligné, etc.) a été utilisée dans l'item source, cette accentuation a-t-elle été respectée dans l'item traduit ?

ET-5 (8) Collecter des données d'études pilotes sur le test adapté afin de réaliser l'analyse d'items, l'évaluation de la fiabilité et la validité permettant d'effectuer les révisions s'avérant nécessaires.

Explication. Avant d'entreprendre des études de grande envergure sur la fiabilité et la validité des scores aux tests et/ou des études d'étalonnage qui sont chronophages et couteuses, il est important d'avoir des preuves préalables confirmant la qualité psychométrique du test adapté. De nombreuses analyses psychométriques peuvent être effectuées pour produire des preuves préalables de fiabilité et de validité des scores. Par exemple, à l'étape de l'élaboration du test, une analyse d'items utilisant un échantillon de taille modeste (n = 100) peut fournir des données très utiles sur le fonctionnement de certains items du test. Les items qui sont très faciles ou très difficiles en comparaison aux autres, ou qui affichent un potentiel de discrimination faible voire négatif, peuvent être révisées. Dans le cas d'items à choix multiples, il serait judicieux d'examiner l'efficacité des distracteurs. Ainsi, les problèmes peuvent être repérés et des modifications apportées. De plus, avec les mêmes données recueillies pour l'analyse d'items, le calcul du coefficient alpha ou le coefficient oméga (McDonald, 1999) fournit des indications précieuses qui pourraient être utilisées pour étayer les décisions concernant la longueur appropriée des versions linguistiques source et cible du test.
Dans certains cas, des interrogations peuvent subsister sur certains aspects de l'adaptation : les consignes du test seront-elles bien comprises? Les consignes devraient-elles être différentes pour guider efficacement les candidats issus de la nouvelle langue et nouvelle culture auxquels est destiné le test adapté? Une administration sur ordinateur pénalisera-t-elle certains répondants (p. ex., les répondants de faible statut socioéconomique) dans la population cible pour le test adapté? Y a-t-il trop de questions pour la durée du test ? On pourrait répondre à toutes ces questions et à bien d'autres encore au moyen d'études

préliminaires de validité. L'objectif serait de compiler suffisamment de données pour qu'une décision puisse être prise quant à l'opportunité de poursuivre la procédure d'adaptation du test. Si l'on décide de la poursuite de la procédure, une série d'études d'envergure peuvent être planifiées et réalisées (p. ex. études pour examiner l'ampleur du fonctionnement différentiel des items, et la structure factorielle du test).

Suggestions pour la pratique. Un certain nombre d'analyses basiques peuvent être effectuées :

- Effectuer une étude classique d'analyse d'items pour obtenir des informations sur les moyennes et les indices de discrimination des items, et effectuer également une analyse des distracteurs pour les items à choix multiple.
- Effectuer une analyse de fiabilité (p. ex., KR-20 avec des items dichotomiques, ou coefficient alpha ou coefficient oméga avec des items polytomiques).
- Au besoin, effectuez une ou deux études préliminaires (pilotes) pour mieux cerner la validité du test adapté. Supposons, par exemple, que le test adapté soit administré sur ordinateur. Il peut être judicieux de réaliser une étude pour évaluer le mode d'administration du test (c.-à-d. format papier-crayon vs. test sur ordinateur). Supposons que les consignes du test invitent les répondants à répondre à tous les items. Il peut être nécessaire de faire des recherches pour déterminer les meilleures façons de formuler de telles consignes permettant d'atteindre cet objectif. Les chercheurs ont constaté qu'il est étonnamment difficile d'amener certains répondants à répondre à tous les items lorsqu'on les encourage à deviner les réponses.

Recommandations pour la Validation/confirmation

Les recommandations relatives à la confirmation sont celles qui ont trait aux analyses empiriques d'études de validité d'envergure.

C-1 (9) Choisir un échantillon dont les caractéristiques sont appropriées à l'utilisation prévue du test et dont la taille et la pertinence sont suffisantes pour les analyses empiriques.

Explication. La préparation de la collecte des données renvoie à la façon dont les données sont recueillies pour établir des normes (si nécessaire) et l'équivalence entre les versions linguistiques d'un test, et pour mener des études de validité et de fiabilité ainsi que des études sur fonctionnement différentiel des items. Une première exigence en ce qui concerne la collecte des données est que les échantillons doivent être suffisamment larges pour permettre l'obtention d'indices statistiques stables. Bien que cette exigence s'applique à tout type de recherche, elle est particulièrement en vigueur dans le contexte d'une étude de

validation d'adaptation d'un test parce que les techniques statistiques nécessaires pour établir l'équivalence du test et des items à travers les groupes/langues (p. ex. l'analyse factorielle confirmatoire, les modèles de réponse à l'item pour l'identification des items potentiellement biaisés) ne peuvent être appliquées de façon efficace qu'à des échantillons suffisamment larges permettant d'estimer avec fiabilité les paramètres du modèle (la taille recommandée pour un échantillon dépend de la complexité et de la nature des données).

De plus, l'échantillon d'une étude de validité d'envergure doit être représentatif de la population à laquelle est destiné le test. Nous attirons l'attention sur l'important document de van de Vijver et Tanzer (1997) et sur les contributions méthodologiques de van de Vijver et Leung (1997), Hambleton, Merenda et Spielberger (2005), Byrne (2008) et Byrne et van de Vijver (2014), pour guider le choix des plans et analyses statistiques appropriés. Sireci (1997) a discuté des problèmes et des enjeux liés à l'établissement d'un lien entre les tests multilingues et une mesure (test) commune.

Parfois, dans la pratique, la population cible à laquelle est destiné le test adapté peut obtenir des résultats beaucoup plus faibles ou plus élevés, et/ou être plus ou moins homogène que la population de la version source. Cela crée des problèmes majeurs pour certaines méthodes d'analyse, comme les études de fiabilité et de validité. Une solution consiste à choisir un sous-échantillon de la population source correspondant à la population cible. Avec les échantillons appariés, toute différence qui pourrait être due à des différences dans la forme des distributions dans les deux groupes peut être éliminée (voir Sireci et Wells, 2010). Par exemple, les comparaisons de structure de test impliquent généralement des covariances, qui varient en fonction de la distribution des scores. En utilisant des échantillons appariés, on peut exclure que le rôle de la distribution des scores sur les résultats puisse en expliquer toute différence.

Un autre exemple pourrait peut-être aider à expliquer le problème des différentes distributions de scores dans les groupes linguistiques source et cible. Supposons que la fiabilité des scores au test est de .80 dans le groupe de la langue source, mais seulement de .60 dans le groupe de la langue cible. La différence peut sembler inquiétante et soulever des questions quant à la pertinence de la version de la langue cible du test. Cependant, on oublie souvent que la fiabilité est une caractéristique commune du test et de la population (McDonald, 1999) parce qu'elle dépend à la fois de la variance du score vrai (caractéristique de la population) et de la variance d'erreur (caractéristique du test). Par conséquent, la même variance d'erreur peut mener à une plus grande fiabilité simplement en raison de la plus grande variance du score vrai dans le groupe de la langue source. McDonald (1999) montre que l'erreur-type de mesure (qui est la racine

carrée de la variance d'erreur) est en fait une quantité plus appropriée pour comparer les échantillons et non la fiabilité. Une autre solution en utilisant les coefficients de fiabilité consisterait à prélever un échantillon apparié de candidats du groupe linguistique source et à recalculer la fiabilité des scores au test.

Les techniques modernes destinées à éprouver l'invariance métrique et structurale grâce à l'analyse factorielle confirmatoire (AFC) multigroupes permettent d'évaluer des échantillons avec différentes distributions des traits latents. Dans de tels modèles, tout en contraignant à l'égalité entre les groupes des paramètres métriques tels que les saturations factorielles des items et leurs intercepts (moyennes estimées), les moyennes, les variances et les covariances des traits latents peuvent, elles, varier d'un groupe à l'autre. Cela permet l'utilisation d'échantillons complets et tient compte du scénario plus réaliste de différentes distributions des traits mesurés dans différentes populations.

Suggestions pour la pratique. Dans presque toutes les recherches, deux suggestions sont faites pour décrire le ou les échantillons :

- Constituer un échantillon aussi large que possible, étant donné que les études visant à identifier les items potentiellement biaisés nécessitent un minimum de 200 participants par version du test (Mazor, Clauser & Hambleton, 1992 ; Subok, 2017). Pour utiliser la théorie de réponse à l'item et les analyses confirmatoires un échantillon d'au moins 500 répondants est nécessaire (Hulin, Lissak et Drasgow, 1982 ; Hambleton, Swaminathan et Rogers, 1991), tandis que les études portant sur la structure factorielle d'un test exigent un échantillon un peu moins large, peut-être 300 participants ou plus (Wolf, Harrington, Clark et Miller, 2013). Il est clair que des analyses avec des échantillons plus petits sont également possibles, mais la première règle est de constituer de grands échantillons de participants chaque fois que cela est possible.

- Dans la mesure du possible, constituez des échantillons représentatifs. Les généralisations de résultats tirés d'échantillons non représentatifs sont limitées. Pour éliminer les différences dans les résultats dues à des facteurs méthodologiques telles que les variances dans la distribution des scores, il est souvent judicieux de prélever un échantillon du groupe linguistique source pour le faire correspondre au groupe linguistique cible. Des comparaisons des erreurs-types de mesure pourraient être plus appropriées.

C-2 (10) Fournir des preuves statistiques satisfaisantes/crédibles sur les équivalences relatives au

construit, aux méthodes et aux items à travers les populations visées.

Explication. Il est important d'établir l'équivalence du construit des versions linguistiques source et cible d'un test. Toutefois, ce n'est pas la seule analyse empirique importante à effectuer. De plus, les approches pour l'équivalence du construit (CP-2) et l'équivalence de méthode (CP-3) ont été abordées brièvement plus tôt dans les recommandations.
Les chercheurs doivent également se pencher sur l'équivalence des items à travers les différents groupes linguistiques. L'équivalence des items est étudiée sous le titre "analyse du fonctionnement différentiel des items (FDI)". En général, un item présente un fonctionnement différentiel (FDI) lorsque deux personnes qui passent le test, provenant de deux populations différentes (culturelles et linguistiques) et bien qu'ayant un niveau égal sur trait/compétence mesuré, ont une probabilité de réponse/réussite différente à cet item. Il est possible que des différences globales entre groupes dans des tests d'acquisitions puissent se produire, mais cela ne pose pas de problème en soi. Alors que, lorsque les membres des populations sont appariés par rapport au construit mesuré par le test (généralement un score total au test, ou score total au test moins le score à l'item en question), et que des différences de performance à l'item existent entre les groupes, cet item affiche donc un fonctionnement différentiel. Ce type d'analyse est effectué pour chaque item du test. Par la suite, on tente de comprendre les raisons d'un tel fonctionnement différentiel de certains items permettant ainsi d'en identifier ceux qui sont défectueux, ceux qui sont modifiables ou ceux qui méritent d'être retirés du test.

Les problèmes de traduction et les différences culturelles sont deux sources potentielles importantes du fonctionnement différentiel des items, et qui méritent d'être évaluées. Plus précisément, le fonctionnement différentiel des items peut être dû (1) à la non-équivalence de la traduction entre la langue source et la langue cible du test, touchant la familiarité avec le vocabulaire utilisé, la difficulté de l'item, l'équivalence sémantique, etc, et (2) les différences culturelles contextuelles (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).

Pendant la traduction, il paraît probable d'utiliser un vocabulaire moins courant dans la langue cible. Les significations pourraient être les mêmes dans les versions traduites, mais, dans une culture, un mot pourrait être plus courant que dans l'autre. Il est également possible que la traduction modifie le niveau de difficulté de l'item en raison de la longueur des phrases, de leur complexité et de l'utilisation d'un

vocabulaire facile ou difficile. Le sens peut également changer dans la langue cible avec la suppression de certaines parties de phrases, des traductions inexactes, la polysémie du vocabulaire utilisé dans la langue cible, des impressions du sens de certains mots différentes d'une culture à l'autre, etc. Les différences culturelles peuvent faire en sorte que les items fonctionnent différemment selon les langues. Par exemple, des mots comme "hamburger" ou "caisse enregistreuse" peuvent ne pas être compris ou bien avoir une signification différente dans deux cultures différentes.

Il existe au moins quatre groupes d'analyses permettant de vérifier si les items fonctionnent différemment selon la langue et/ou le groupe culturel: (a) les procédures basées sur la théorie de réponse à l'item (TRI) (Ellis, 1989 ; Thissen, Steinberg et Wainer, 1988 ; 1993 ; Ellis et Kimmel, 1992), (b) la procédure de Mantel-Haenszel (MH) et ses extensions (voir, p. ex, Dorans et Holland, 1993 ; Hambleton, Clauser, Mazor et Jones, 1993 ; Holland et Wainer, 1993 ; Holland et Wainer, 1993 ; Sireci et Allalouf, 2003), (c) les procédures de régression logistique (RL) (Swaminathan et Rogers, 1990 ; Rogers et Swaminathan, 1993) et (d) l'analyse factorielle restrictive (AFR) (Oort et Berberoğlu, 1992).

Dans les approches basées sur la théorie de réponse à l'item, les répondants des deux langues sont appariés en fonction des scores aux traits latents. Dans les méthodologies MH et RL, le score observé ou estimé au test est utilisé comme critère d'appariement avant de comparer la performance à l'item des répondants des deux groupes. Bien que le score total observé soit le critère d'appariement le plus populaire dans ces procédures, d'autres scores estimés, par exemple à partir de l'analyse factorielle, peuvent également être utilisés. Ces scores sont également "purifiés" de façon itérative en supprimant les items problématiques. Le critère d'appariement doit être suffisamment valide et fiable pour permettre d'évaluer correctement le fonctionnement différentiel des items. Dans l'analyse factorielle restrictive, chaque item dépend de la variable latente indiquant l'appartenance à un groupe (i.e., variable potentiellement perturbatrice de l'uniformité des items) ainsi que du trait latent. Chaque saturation factorielle constitue un paramètre libre à estimer dans le modèle. L'adéquation de celui-ci est évaluée en comparaison au modèle nul où la variable latente de l'appartenance à un groupe ne sature pas l'item en question. Si le modèle offre un ajustement nettement meilleur, cet item présente un FDI (c'est-à-dire il est biaisé).

Lorsque la dimensionnalité d'un test est complexe, il est difficile de trouver un critère d'appariement approprié (Clauser, Nungester, Mazor et Ripkey, 1996). L'utilisation de critères d'appariement multivariés, comme les différents scores factoriels obtenus à la suite de l'analyse factorielle, pourrait également modifier les interprétations du fonctionnement différentiel des items. Par conséquent, les

présentes recommandations suggèrent que, si le test est multidimensionnel, les chercheurs pourraient utiliser divers critères pour détecter et évaluer les items affichant un fonctionnement différentiel. L'appariement multivarié peut réduire le nombre d'items présentant des FDI à travers les groupes linguistiques et culturels

Cette recommandation invite les chercheurs à détecter les éventuelles sources de biais de méthode dans le test adapté. Les sources de biais de méthode comprennent (1) les différents degrés de motivation des participants au test, (2) la différence de familiarité des répondants avec les tests psychologiques, (3) Passation plus rapide du test dans un groupe linguistique que dans l'autre, (4) une différence de familiarité avec le format de réponse entre les groupes linguistiques, (5) l'hétérogénéité du style de réponse, etc. Les biais entachant les réponses ont été, par exemple, une préoccupation majeure dans l'interprétation des résultats de l'enquête PISA, et ont fait l'objet d'un certain nombre de recherches.

Enfin, et c'est un point important, cette recommandation exigera que les chercheurs se penchent sur la question de l'équivalence de construit. Il existe au moins quatre approches statistiques pour évaluer l'équivalence du construit à travers les versions linguistiques source et cible d'un test : l'analyse factorielle exploratoire (AFE), L'analyse factorielle confirmatoire (AFC), l'échelonnement multidimensionnel (EMD), et les comparaisons des réseaux nomologiques (Sireci, Patsula, & Hambleton, 2005).

Selon van de Vijver et Poortinga (1991), l'analyse factorielle (AFE et AFC) est la technique statistique la plus fréquemment utilisée pour évaluer l'équivalence d'un construit à travers différentes langues et cultures. Cette affirmation de 1991 est toujours d'actualité, bien que les approches de modélisation statistique aient considérablement progressé (voir, par exemple, Hambleton & Lee, 2013, Byrne, 2008). Étant donné qu'avec l'AFE il est difficile de comparer différentes structures factorielles, et qu'il n'existe pas de règles communes pour pouvoir juger de l'équivalence de ces structures, des approches statistiques telles que l'AFC (voir, par exemple, Byrne, 2001, 2003, 2006, 2008) et le EMDP (Échelonnement multidimensionnel pondéré) sont préférables car elles peuvent évaluer simultanément plusieurs groupes à la fois (Sireci, Harter, Yang, & Bhola, 2003).

L'AFC a été utilisée dans nombreuses études pour vérifier si la structure factorielle d'une version originale d'un test était identique dans toutes ses versions adaptées (p. ex., Byrne et van de Vijver, 2014). L'AFC est intéressante pour l'évaluation de l'équivalence factorielle concernant les tests adaptés, car d'une part elle peut traiter plusieurs groupes simultanément, et d'autre part elle fournit des tests statistiques et des indices

descriptifs d'adéquation du modèle aux données (Sireci, Patsula, et Hambleton, 2005). La possibilité de traiter simultanément plusieurs groupes est d'autant plus importante qu'il est de plus en plus courant d'adapter les tests dans de nombreuses langues (par exemple, certains tests d'intelligence sont désormais traduits/adaptés dans plus de cent langues et, dans les études TIMSS et OCDE/PISA, les tests sont adaptés dans plus de 30 langues). Cependant, étant donné que la stricte exigence relative à l'absence de saturations croisées dans l'AFC n'est pas tenable pour les instruments multidimensionnels complexes, la modélisation exploratoire par équations structurales (MEES) gagne en popularité, en particulier pour les données de personnalité ou les variables plus complexes et interreliées (Asparouhov & Muthén, 2009).

L'échelonnement multidimensionnel pondéré est une autre approche séduisante pour évaluer l'équivalence du construit à travers les différentes versions linguistiques d'un test. Comme pour l'AFE, l'échelonnement multidimensionnel pondéré n'exige pas la spécification a priori de la structure factorielle des tests et, comme l'AFC, il permet l'analyse multigroupes (p. ex. Sireci et al., 2003).

Van de Vijver et Tanzer (1997) ont suggéré que les chercheurs en interculturalité devraient examiner la fiabilité de chaque version culturelle du test en question et rechercher des preuves de sa validité convergente et discriminante dans chaque groupe culturel. Ces études peuvent souvent être plus pratiques que les études sur la structure factorielle du test qui, elles, exigent des échantillons de très grande taille.

Il faut cependant reconnaître que la comparaison des performances des candidats dans deux versions linguistiques d'un test n'a pas toujours été le but visé par la traduction/adaptation d'un test. Peut-être, entre autres, l'objectif est-il simplement de pouvoir évaluer les candidats d'un groupe linguistique différent de celui pour lequel a été élaboré le test. Dans ce cas, il est essentiel d'examiner sérieusement la validité du test dans le groupe de la seconde langue, mais tenter de prouver l'équivalence des performances dans les deux versions n'est pas nécessaire. L'importance de cette recommandation dépendra de l'objectif ou des objectifs du test dans la seconde langue (c.-à-d. le groupe linguistique cible). Les tests comme ceux utilisés dans l'étude PISA ou l'étude TIMSS exigent la preuve d'une grande similitude de contenu d'une version à l'autre, car leurs résultats sont utilisés pour comparer les performances des élèves dans de nombreux pays. L'utilisation d'une échelle de dépression traduite de l'anglais vers le chinois pour permettre aux chercheurs d'étudier la dépression ou aux thérapeutes/praticiens d'évaluer la dépression de leurs clients ne nécessiterait pas une forte similitude de contenu entre les deux versions. En revanche, l'utilisation de cette échelle de dépression en Chine passe d'abord par les preuves de sa validité.

Cette recommandation peut également être abordée avec des méthodes statistiques une fois que le test aura été adapté. Par exemple, si l'on pense que les groupes culturels diffèrent sur des variables importantes, mais sans rapport aucun avec le construit mesuré, on peut utiliser des plans et des analyses statistiques pour contrôler ces variables "parasites". L'analyse de covariance, les plans expérimentaux en blocs randomisés et d'autres techniques statistiques (analyse de régression, corrélation partielle, etc.) peuvent être utilisés pour contrôler les effets des sources de variation indésirables à travers les groupes.

Suggestions pour la pratique. Il s'agit d'une recommandation très importante appelant de nombreuses analyses. Concernant les analyses d'équivalence, nous faisons les suggestions suivantes pour la pratique :

- Si la taille des échantillons est suffisante, procéder à une étude comparative de l'équivalence du construit des versions linguistiques source et cible du test. Il existe de nombreux logiciels pour faciliter la réalisation de ces analyses (voir Byrne, 2006).
- Afin de déterminer l'équivalence de la structure du test à travers les groupes linguistiques et/ou culturels, il convient d'effectuer une analyse factorielle exploratoire (avec, de préférence, une "rotation cible") ou une analyse factorielle confirmatoire, et/ou une analyse d'échelonnement multidimensionnel pondéré. La nécessité d'avoir des d'échantillons assez larges (10 participants par variable) rend ces études difficiles à réaliser dans de nombreuses études interculturelles. Une excellente illustration pour ce type d'étude est disponible dans Byrne et van de Vijver (2014).
- Rechercher des preuves de validité convergente et discriminante (essentiellement, rechercher des preuves corrélationnelles parmi un ensemble de construits et vérifier la constance/similarité de ces corrélations à travers les groupes linguistiques et/ou culturels) (voir van de Vijver & Tanzer, 1997).

En ce qui concerne le fonctionnement différentiel des items (FDI), certaines suggestions sont présentées ci-dessous. Pour des approches plus sophistiquées, les chercheurs sont encouragés à se référer à la littérature spécialisée sur le FDI:

Effectuer une analyse FDI en utilisant l'une des procédures standard (si les items sont dichotomiques, la procédure de Mantel-Haenszel peut être la plus simple; si les items sont polytomiques, la procédure généralisée de Mantel-Haenszel constitue une option possible). D'autres solutions plus sophistiquées existent y compris les approches basées sur la théorie de réponse à l'item. Si la taille de l'échantillon est plus modeste, un "diagramme delta" peut révéler les items potentiellement biaisés. Les comparaisons conditionnelles sont une autre possibilité (pour les méthodes de comparaison en présence d'échantillons

modestes, voir, par exemple, Muñiz, Hambleton, & Xing, 2001).

C-3 (11) Fournir des preuves étayant les normes, la fiabilité et la validité de la version adaptée du test dans les populations visées.

Explication. Les normes, les preuves de validité et de fiabilité d'un test dans sa version linguistique d'origine ne s'appliquent pas automatiquement aux autres adaptations possibles du test à différentes cultures et langues. Par conséquent, les preuves empiriques de validité et de fiabilité de toute nouvelle version d'un test doivent également être présentées. Toutes sortes de preuves empiriques venant étayer les interprétations susceptibles d'être tirées à partir des scores au test devraient être incluses dans le manuel du test. Une attention particulière devrait être accordée aux cinq sources de preuves de validité basées sur : le contenu du test, les processus de réponse, la structure interne du test, les relations avec les autres construits et les conséquences liées à l'utilisation du test (AERA, APA, NCME, 2014). L'analyse factorielle exploratoire et confirmatoire, la modélisation par équations structurales et l'analyse multitraits-multiméthodes sont quelques-unes des techniques statistiques qui peuvent être utilisées pour obtenir des preuves basées sur la validité de la structure interne d'un test.

Suggestions pour la pratique. Les suggestions sont identiques à celles requises pour tout test dont l'utilisation est envisagée :

- Si l'on suggère d'utiliser les normes élaborées pour la version originale du test avec la version adaptée, il faut fournir la preuve que cette utilisation est statistiquement appropriée, et garantit l'équité. Si aucune preuve ne peut être fournie pour l'utilisation des normes originales, des normes spécifiques devraient être élaborées pour la version adaptée, conformément aux standards en vigueur en matière d'élaboration des normes.
- Compiler un nombre suffisant de preuves de fiabilité pour justifier l'utilisation de la version linguistique cible du test. Ces preuves pourraient normalement comprendre une estimation de la cohérence interne (p. ex., KR-20 ou coefficient alpha ou oméga).
- Compiler autant de preuves de validité que nécessaire pour déterminer si la version linguistique cible du test doit être utilisée. Le type de preuves compilées dépendrait de l'utilisation prévue des scores (p. ex. validité du contenu pour les tests d'acquisitions, validité prédictive pour les tests d'aptitude, etc.)

C-4 (12) Utiliser une procédure de mise en équivalence et des procédures d'analyse des données

appropriées pour parvenir à jumeler des scores provenant de différentes versions linguistiques d'un test.

Explication. Lorsque l'on souhaite relier deux versions linguistiques d'un test à une échelle commune de notation, plusieurs options sont possibles. Si un ensemble commun d'items est utilisé, le fonctionnement de ces items communs dans ces deux groupes linguistiques doit être évalué et si le fonctionnement différentiel de certains items est observé, il faut envisager de les retirer des données utilisées pour établir ce lien. Les diagrammes delta (Angoff et Modu, 1973) sont utiles à cette fin, et Cook et Schmitt-Cascallar (2005) ont fourni une bonne illustration de la façon de les utiliser pour identifier les items qui ont une signification différente pour les deux groupes de personnes testées. Les types d'items n'ont pas tous le même potentiel de lien entre les versions linguistiques. Les estimations de la difficulté de l'item et de son potentiel de discrimination dans le cadre de la théorie de réponse à l'item, peuvent être restituées sur un diagramme pour aider à identifier les items communs dont la performance est inappropriée (voir Hambleton, Swaminathan, & Rogers, 1991).

Mais jumeler (c.-à-d. "mettre en équivalence ") les scores provenant de deux versions linguistiques d'un test sera toujours problématique parce qu'il faut avancer des présupposés forts au sujet des données. Parfois, il est supposé de manière audacieuse que les différentes versions linguistiques du test sont équivalentes, et donc que les scores des deux versions du test peuvent être utilisés de façon interchangeable. Une telle hypothèse peut être jugée valable avec les tests/épreuves de mathématiques car leur traduction/adaptation est généralement simple. On peut également lui accorder du crédit si les deux versions du test ont été construites avec soin, et si l'on peut donc supposer que la version en langue source du test fonctionne auprès de la population source d'une manière équivalente à celle de la version en langue cible du test auprès de la population cible. Cette hypothèse peut être fondée si toutes les autres preuves à l'appui suggèrent que les deux versions linguistiques du test sont équivalentes et qu'il n'y a pas de biais de méthode influençant les scores de la version linguistique cible du test.

Deux autres solutions existent, mais aucune des deux n'est parfaite. Premièrement, le jumelage des deux versions pourrait être fait avec un sous-échantillon d'items qui sont jugés essentiellement équivalents dans les deux versions linguistiques du test. Par exemple, ces items peuvent être ceux qui ont été jugés très faciles à traduire/adapter. En principe, la solution pourrait fonctionner, mais il faut que les items

équivalents ainsi que le reste des items mesurent le même construit. Une deuxième solution consiste en la mise en équivalence avec un échantillon de personnes bilingues. Cet échantillon répondant aux deux versions du test, il serait possible d'établir une table de conversion des scores. Dans l'idéal, il ne faut pas que l'échantillon soit trop petit et, que dans le protocole, l'ordre de présentation des versions du test soit contrebalancé. Le grand présupposé de cette approche est que les répondants sont vraiment bilingues, et donc, à part les difficultés relatives des tests, les répondants devraient réussir de façon égale sur les deux versions du test. Toute différence est utilisée pour ajuster les scores lors de leur conversion d'une version du test à l'autre.

Suggestions pour la pratique. Le jumelage des scores provenant des versions adaptées d'un test va être problématique dans le meilleur des cas, car tous les procédés de mise en équivalence présentent au moins une limite majeure. La meilleure stratégie consiste probablement à respecter complètement toutes les étapes de l'établissement de l'équivalence des scores. Si les évidences relatives aux trois questions ci-dessous sont solides, les scores des deux versions du test peuvent être alors traités de façon interchangeable :

- Y a-t-il des preuves montrant que le même construit est bel et bien mesuré dans les versions en langue source et en langue cible du test ? Le construit a-t-il les mêmes relations avec d'autres variables externes dans la nouvelle culture ?

- Y a-t-il des preuves solides montrant que les sources de biais de méthode ont été éliminées (par exemple, pas de problèmes de durée de passation, les formats utilisés dans le test sont également familiers aux répondants, pas de malentendu dans les consignes, pas de fausse représentation systématique dans un groupe ou l'autre, consignes standardisées, absence de styles de réponse (scores extrêmes, motivations différentes...) ?

- Le test est-il exempt d'items potentiellement biaisés ? Ici, vous pouvez vous aider d'un diagramme des valeurs *p* (c.a.d, la proportion de réponses correctes à l'item) ou, mieux, des valeurs delta, à partir des items des deux versions du test. Les points qui ne tombent pas le long de la ligne de mise en équivalence linéaire doivent être étudiés pour déterminer si les items associés sont également appropriés dans les deux langues. Les analyses FDI fournissent des preuves encore plus solides de l'équivalence des items auprès des groupes linguistiques et culturels.

- Si vous tentez de jumeler des scores provenant des différentes versions linguistiques du test, il faut alors choisir et mettre en œuvre un procédé de mise en équivalence approprié. Des preuves concernant la validité de ce procédé doivent être fournies...

Recommandations relatives à/sur l'administration de test

A-1 (13). Préparer le matériel et les consignes d'administration afin de minimiser tout problème en lien avec la culture et la langue causé par les procédures d'administration et les modes de réponse, et pouvant affecter la validité des interprétations tirées à partir des scores.

Explication. La mise en œuvre des recommandations relatives à l'administration devrait commencer par une analyse de tous les facteurs susceptibles de menacer la validité des scores au test dans un contexte culturel et linguistique spécifique. L'expérience dans l'administration de tests dans un contexte monolingue ou monoculturel peut d'ores et déjà être utile pour pouvoir anticiper les problèmes auxquels on peut s'attendre dans un contexte multilingue ou multiculturel. Par exemple, ceux qui ont une expérience dans l'administration de tests savent souvent quels aspects de la consigne peuvent être difficiles pour les répondants. Ces aspects peuvent demeurer difficiles après la traduction ou l'adaptation. L'utilisation des tests dans un nouveau contexte linguistique ou culturel pourrait également soulever des problèmes que l'on ne trouvait pas auparavant dans les utilisations monoculturelles.

Suggestions pour la pratique. Il est important d'anticiper, à travers cette recommandation, les facteurs susceptibles de générer des problèmes lors de l'administration de tests. Voici quelques-uns des facteurs qui doivent être étudiés pour garantir l'équité dans l'administration de tests :
La clarté des consignes du test (y compris leur version traduite), le procédé de réponse (p. ex., la feuille de réponse), durée (une source d'erreur courante est le fait de ne pas accorder suffisamment de temps aux candidats pour terminer le test), la motivation des candidats à terminer le test, leur connaissance du but du test et la façon dont il sera coté.

A-2 (14) Préciser les conditions de testing qui doivent être identiques pour toutes les populations d'intérêt.

Explication. La présente recommandation a pour but d'encourager les auteurs de tests à établir des procédures de testing et d'administration (p. ex. conditions d'administration, durée, etc.) qui doivent être

respectées auprès de toutes les populations d'intérêt. Cette recommandation vise principalement à encourager ceux qui administrent les tests à s'en tenir aux instructions standardisées. En même temps, des accommodements pourraient être précisés pour répondre aux besoins de sous-groupes particuliers de personnes au sein de chaque population, comme du temps supplémentaire, document imprimé en plus gros caractères, des conditions d'administration de tests plus calmes, etc. Dans le domaine du testing, ces mesures sont connues aujourd'hui sous le nom d "accommodements de la procédure de testing". Le but de ces accommodements n'est pas de gonfler les scores des candidats, mais plutôt de créer un environnement spécifique de testing leur permettant d'exprimer ce qu'ils ressentent, ou ce qu'ils savent et peuvent faire.

Les écarts par rapport aux conditions standardisées de testing doivent être signalés tant est si bien que leur incidence sur les généralisations et les interprétations puissent être pris en compte plus tard dans le processus.

Suggestions pour la pratique. Cette recommandation peut en partie recouper la recommandation A-1 (13), mais elle est reformulée ici pour souligner l'importance pour les répondants de passer les tests dans des conditions aussi semblables que possible. Cela est essentiel si les scores des deux versions linguistiques doivent être utilisés de façon interchangeable. Voici quelques suggestions :

- Les consignes du test et les procédures connexes doivent être adaptées et réécrites d'une manière standardisée, qui convient à la nouvelle langue et culture.

- Si les consignes du test et les procédures connexes sont modifiées en fonction des nouvelles cultures, ceux qui sont chargés d'administrer le test doivent recevoir une formation sur les nouvelles procédures; ils doivent être informés de l'obligation de respecter celles-ci et non les procédures originales.

Recommandations relatives à la cotation et à l'interprétation des scores (CIS)

CIS-1 (15). Interpréter toute différence de score entre les groupes en tenant compte de toutes les informations pertinentes disponibles.

Explication. Même si un test a été adapté grâce à des procédures techniquement sûres, et que la validité des scores a déjà été dans une certaine mesure établie, il faut garder à l'esprit que le sens des différences entre groupes peut être interprété de nombreuses façons en raison des différences culturelles ou autres

entre les pays et/ou cultures en présence. Dans son article, Sireci (2005) a passé en revue la méthode d'évaluation de l'équivalence de deux versions linguistiques différentes d'un test, en administrant les versions linguistiques distinctes du test à un groupe de répondants maîtrisant les deux langues (bilingues) et provenant du même groupe culturel ou linguistique. Il a tout à la fois décrit certaines options du protocole de recherche pour les études d'équivalence utilisant des répondants bilingues, listé les éventuelles variables parasites à contrôler, et offert de précieuses suggestions pour l'interprétation des résultats.

Suggestions pour la pratique. Ci-dessous une suggestion pour améliorer la pratique :

- Selon l'objet de la recherche (ou le contexte pour lequel les comparaisons entre groupes sont réalisées), un certain nombre d'interprétations possibles peuvent être envisagées, avant d'en retenir une en particulier. Par exemple, il est important de tenir compte des différences quant à la motivation de bien réussir au test avant de conclure qu'un groupe a obtenu un meilleur résultat qu'un autre. Il peut aussi y avoir des effets de contexte qui ont eu un impact significatif sur la performance au test. Par exemple, un groupe de personnes peut simplement faire partie d'un système d'éducation moins efficace, ce qui aurait un impact significatif sur la performance au test.

CIS-2 (16) Ne comparer les scores entre les populations que lorsque le niveau d'invariance a été établi sur l'échelle sur laquelle les scores sont rapportés.

Explication. Lorsque les études comparatives entre groupes linguistiques et culturels sont au cœur de l'initiative de traduction et d'adaptation, les versions multilingues d'un test doivent être placées sur une échelle de mesure commune, et cela se fait par un processus appelé "jumelage" ou "mise en équivalence". Cela exige des échantillons de taille importante et des preuves montrant que la version adaptée du test ne comporte pas de biais de construit, de biais de méthode et de biais d'item.

Van de Vijver et Poortinga (2005) ont délimité plusieurs niveaux d'équivalence des tests entre groupes linguistiques et culturels, et leurs travaux sont particulièrement utiles pour comprendre ce concept dont ils se partagent la paternité Par exemple, ils ont souligné que l'équivalence des unités de mesure/évaluation exige que les échelles de notation dans chaque groupe aient le même système métrique, ce qui garantit que les différences entre les personnes au sein des groupes ont la même signification. (Par exemple, les différences entre les hommes et les femmes d'un échantillon chinois peuvent être comparées à celles d'un

échantillon français). Cependant, les comparaisons directes valides des scores ne peuvent être faites que lorsque les scores présentent le plus haut niveau d'équivalence, appelé équivalence scalaire ou équivalence totale des scores, ce qui exige que les échelles de chaque groupe aient la même unité de mesure/évaluation et la même origine d'un groupe à l'autre.

De nombreuses méthodes (tant dans le cadre de la théorie classique des tests, que dans celui de la théorie de réponse à l'item) ont été proposées pour jumeler ou mettre en équivalence les scores de deux groupes (ou versions linguistiques d'un test). Les lecteurs intéressés peuvent se référer à Angoff (1984) et à Kolen et Brennan (2004) pour mieux comprendre ce sujet. Cook et Schmitt-Cascallar (2005) proposent une base pour comprendre les méthodes statistiques qui sont actuellement disponibles pour la mise en équivalence et l'échelonnage des tests éducatifs et psychologiques. Les auteurs décrivent et critiquent les procédures de jumelage d'échelles spécifiques utilisées dans les études d'adaptation de tests. Ils illustrent également certaines procédures et problèmes du jumelage d'échelles en décrivant et en critiquant trois études qui ont été menées au cours des vingt dernières années pour la mise en équivalence des scores du *Scholastic Assessment Test* avec sa version espagnole, la *Prueba te Aptitude Academica.*

Suggestions pour la pratique. Ici, l'élément crucial est que les scores aux tests ne doivent pas être sur-interprétés :

- Interpréter les scores en fonction du niveau de preuve de validité disponible. Par exemple, ne pas faire de constats comparatifs sur les niveaux de performance des répondants dans deux groupes linguistiques, à moins que l'invariance métrique ait été établie pour les scores aux tests comparés...

Recommandations relatives à la documentation

Doc-1 (17). Fournir la documentation technique de tout changement, y compris les éléments de preuves obtenues pour plaider en faveur de l'équivalence, lorsqu'un test est adapté pour être utilisé dans une autre population.

Explication. L'importance de cette recommandation a été réalisée et soulignée par de nombreux chercheurs (voir, par exemple, Grisay, 2003). L'étude TIMSS et l'étude PISA ont très bien réussi à respecter cette recommandation en documentant soigneusement les changements tout au long des travaux d'adaptation. Grâce à ces informations, il est possible d'apprécier la pertinence des changements qui ont

été apportés.

La documentation technique devrait également contenir suffisamment de détails sur la méthodologie pour que de futurs chercheurs puissent reproduire les procédures utilisées sur la même population ou sur d'autres. Elle doit contenir suffisamment d'informations sur les preuves relatives à l'équivalence de construit et de l'équivalence de l'échelonnage (si elle est réalisée) pour justifier l'utilisation de l'instrument dans la nouvelle population. Lorsque des comparaisons inter-populations sont envisagées, la documentation devrait faire état des preuves convoquées pour déterminer l'équivalence des scores entre les populations.

Parfois, la question se pose de savoir à qui s'adresse la documentation technique. La documentation doit être rédigée à l'intention des experts et des personnes qui devront évaluer l'utilité de l'utilisation du test dans la nouvelle population ou dans d'autres populations. (Un document simplifié supplémentaire pourrait être ajouté au profit des non-experts).

Suggestions pour la pratique. Les tests adaptés devraient être assortis d'un manuel technique qui documente toutes les preuves qualitatives et quantitatives jalonnant le processus d'adaptation. Il est particulièrement utile de documenter tout changement apporté pour accommoder le test à une autre langue et autre culture. Essentiellement, les experts et les éditeurs de revues voudront obtenir de la documentation sur le processus qui a été suivi pour développer et valider la version du test dans la langue cible. Bien sûr, ils voudront aussi voir les résultats de toutes les analyses réalisées. Voici les types de questions qui doivent être abordées :

- Quelles sont les preuves disponibles étayant l'utilité du construit et du test adapté dans la nouvelle population ?

- Quelles données sur les items ont été recueillies et à partir de quels échantillons ?

- Quelles autres données ont été obtenues pour évaluer la validité de contenu, critérielle et de construit ?

- Comment les différents jeux de données ont-ils été analysés ?

- Quels en ont été les résultats ?

Doc-2 (18). Fournir de la documentation pour les utilisateurs du test qui confortera les bonnes pratiques de l'administration d'un test adapté auprès des personnes dans le contexte de la nouvelle population.

Explication. La documentation doit être rédigée à l'intention des personnes qui utiliseront le test dans le cadre d'une évaluation pratique. Elle doit être conforme aux bonnes pratiques définies dans les Recommandations sur l'utilisation des tests de la Commission Internationale des Tests (voir www.InTestCom.org).

Suggestions pour la pratique. L'auteur du test doit fournir des renseignements précis sur la façon dont les contextes socioculturels et écologiques des populations pourraient influer sur la performance au test. Le manuel d'utilisation devrait alors :

- Décrire le ou les construits mesurés par le test ainsi que la procédure d'adaptation.

- Résumer les preuves étayant l'adaptation, y compris les preuves des bien-fondés culturels du contenu des items, du caractère approprié des consignes du test, du format de réponse, etc.

- Définir les bien-fondés de l'utilisation du test avec divers sous-groupes de la population et toute autre restriction d'utilisation.

- Expliquer les problèmes/enjeux qui doivent être pris en compte en ce qui concerne les bonnes pratiques de l'administration des tests.

- Préciser s'il est possible d'effectuer des comparaisons entre les populations et, le cas échéant, expliquer comment le faire.

- Fournir les renseignements nécessaires à la cotation et à l'étalonnage (p. ex., tableaux de consultation des normes pertinentes) ou décrire comment les utilisateurs peuvent accéder aux procédures de cotation (p. ex., lorsque celles-ci sont informatisées).

- Fournir des recommandations pour l'interprétation des scores, y compris de l'information sur les implications des données de validité et de fiabilité sur les interprétations qui peuvent être tirées à partir des scores aux tests.

EN CONCLUSION

Nous avons fait de notre mieux pour aboutir à un ensemble de recommandations de nature à aider les auteurs/créateurs et les utilisateurs de tests dans leur travail. Toutefois, pour que celles-ci et les autres efforts visant à modifier les mauvaises pratiques aient un effet, il faut que de bons mécanismes de diffusion soient mis en place. Une étude systématique récente de Rios et Sireci (2014) a démontré que la majorité des projets d'adaptation de tests dans la littérature publiée ne suivaient pas, en fait, les recommandations de l'ITC qui sont disponibles depuis environ 20 ans maintenant. Nous encourageons donc les lecteurs à faire tout leur possible pour sensibiliser leurs collègues à cette deuxième édition en tant que source principale de meilleures pratiques à laquelle ont contribué de nombreux professionnels du monde entier.

En même temps, nous avons conscience que tout comme la première édition de ces recommandations est en train d'être remplacée, il en sera de même pour cette deuxième édition. Les bien-connus standards pour une évaluation en éducation et psychologie de l'AERA, de l'APA et du NCME en sont maintenant à leur sixième édition (AERA, APA et NCME, 2014). Nous nous attendons à ce que les recommandations de l'ITC pour l'adaptation des tests fassent également l'objet d'autres révisions dans les années à venir. Si vous connaissez de nouvelles études qui devraient être citées, ou qui pourrait influencer la troisième édition, ou si vous souhaitez proposer de nouvelles recommandations ou révisions aux 18 recommandations présentées ici, nous vous prions d'en informer l'ITC. Vous pouvez contacter le président actuel du comité de recherche et des recommandations qui a produit la deuxième édition et/ou le secrétaire de l'ITC à l'adresse électronique que vous trouverez sur www.InTestCom.org.

RÉFÉRENCES

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). Mahwah, NJ: Lawrence Erlbaum associates.

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). Hillsdale, NJ: Lawrence Erlbaum Associates.

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: Taylor and Francis Group.

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*. Mannheim: Zentrum fürUmfragen, Methoden und Analysen -ZUMA-

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

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

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

Jeanrie, C., & Bertrand, R. (1999). Translating tests with the International Test Commission Guidelines:

Keeping validity in mind. *European Journal of Psychological Assessment, 15* (3), 277-283.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression

procedures. *Journal of Educational Measurement, 27*, 361-370.

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

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

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

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

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

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

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

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

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

van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodical issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39-64). Mahwah, NJ: Lawrence Erlbaum Publishers.

van de Vijver, F. J. R., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47* (4), 263-279.

Wolf, E.J., Harrington, K.M., Clark, S.L., & Miller, M.W. (2013). Sample size requirements for structural

equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement, 73* (6), 913–934.

APPENDICE A. Checklist des recommandations de l'ITC pour la traduction et l'adaptation de tests.

Voici une checklist pour vous rappeler les dix-huit recommandations de l'ITC. Nous vous invitons à vérifier celles que vous estimez avoir traitées de manière satisfaisante dans votre projet de

traduction/adaptation de tests, et de traiter ensuite celles qui sont en souffrance.

Recommandations concernant les Conditions Préalables (CP)

- CP-1 (1) Obtenir l'autorisation nécessaire auprès du titulaire des droits de propriété intellectuelle du test avant d'entreprendre son adaptation.

- CP-2 (2) Évaluer auprès de la population cible que le degré de concordance entre la définition et le contenu du construit mesuré par le test original et le contenu de chaque item soit suffisant en vue de l'utilisation prévue (ou les utilisations prévues) des scores au test.

- CP-3 (3) Réduire drastiquement l'influence des différences culturelles et linguistiques préjudiciables/non désirables/inutiles à l'utilisation prévue du test dans les populations cibles.

Recommandations pour l'élaboration de tests (ET)

- ET-1 (4) Veiller à ce que la procédure de traduction et d'adaptation tienne compte des différences linguistiques, psychologiques et culturelles des populations visées en choisissant des experts possédant les compétences nécessaires.

- ET-2 (5) Utiliser des designs et procédures de traduction appropriés pour maximiser l'adéquation/convenance de l'adaptation du test aux populations cibles.

- ET-3 (6) Fournir les preuves que la consigne du test et le contenu des items ont la même signification pour toutes les populations visées.

- ET-4 (7) Démontrer que les formats d'items, les échelles de réponse/cotation, les catégories de cotation, les conventions relatives aux tests, les modes d'administration et toute autre procédure conviennent à toutes les populations visées.

- ET-5 (8) Collecter des données d'études pilotes sur le test adapté afin de réaliser l'analyse d'items, l'évaluation de la fiabilité et la validité permettant d'effectuer les révisions jugées nécessaires.

Recommandations pour la Validation/confirmation

- C-1 (9) Choisir un échantillon dont les caractéristiques sont appropriées à l'utilisation prévue du test et dont la taille et la pertinence sont suffisantes pour les analyses empiriques.

- C-2 (10) Fournir des preuves statistiques satisfaisantes/crédibles sur les équivalences relatives au construit, aux méthodes et aux items à travers les populations visées.

- C-3 (11) Fournir des preuves étayant les normes, la fiabilité et la validité de la version adaptée du test dans les populations visées**.**

- C-4 (12) Utiliser une procédure de mise en équivalence et des procédures d'analyse des données appropriées pour pouvoir jumeler des scores provenant de différentes versions linguistiques d'un test.

Recommandations relatives à l'administration de test

- A-1 (13). Préparer le matériel et les consignes d'administration afin de minimiser tout problème en lien avec la culture et la langue causé par les procédures d'administration et les modes de réponse, et pouvant affecter la validité des interprétations tirées à partir des scores.

- A-2 (14) Préciser les conditions de testing qui doivent être identiques pour toutes les populations d'intérêt.

Recommandations relatives à la cotation et à l'interprétation des scores

- CIS-1 (15). Interpréter toute différence de scores entre les groupes en se référant à toutes les informations pertinentes disponibles.

- CIS-2 (16) Ne comparer les scores entre populations que lorsque le niveau d'invariance a été établi sur l'échelle sur laquelle les scores sont rapportés.

Recommandations relatives à la documentation

- Doc-1 (17). Fournir la documentation technique de tout changement, y compris un compte rendu des preuves obtenues pour appuyer l'équivalence, lorsqu'un test est adapté pour être utilisé dans une autre population.

- Doc-2 (18). Fournir aux utilisateurs de tests de la documentation qui confortera les bonnes pratiques dans l'administration d'un test adapté auprès des personnes dans le contexte de la nouvelle population.

APPENDICE B. Glossaire des notions

Alpha (ou parfois appelé "Coefficient Alpha" ou "Alpha de Cronbach"). Coefficient de fiabilité d'un test dont les items sont supposés mesurer un attribut commun et possédant tous un potentiel discriminant identique (c'est donc un cas particulier d'Omega - voir ci-dessous). Dans des conditions plus générales, il

s'agit d'une exigence minimale de la fiabilité.

Analyse Factorielle Confirmatoire (AFC). Une hypothèse sur la structure d'un test est formulée au préalable, puis des analyses sont effectuées pour évaluer cette structure à partir de la matrice de corrélation des items du test. Un test statistique est effectué pour voir si la structure hypothétique et celle estimée sont suffisamment proche pour que l'hypothèse nulle selon laquelle les deux structures sont équivalentes ne puisse pas être rejetée.

Analyse Factorielle Exploratoire (AFE). L'analyse factorielle est une procédure statistique qui est appliquée, par exemple, avec la matrice de corrélation produite par les inter-corrélations entre un ensemble d'items dans un test (ou un ensemble de tests). L'objectif est d'essayer d'expliquer les inter-corrélations entre les items du test (ou des tests) en fonction d'un petit nombre de facteurs que l'on croit être mesurés par le test (ou les tests). Par exemple, dans le cas d'un test de mathématiques, une analyse factorielle pourrait permettre de déterminer que les items se répartissent en trois catégories : les items de calcul, les concepts et la résolution de problèmes. On pourrait donc dire que le test de mathématiques mesure trois facteurs - le calcul, les concepts mathématiques et la résolution de problèmes mathématiques.

Développement simultané de tests. Développement de questionnaires en langue source et en langue cible simultanément, en utilisant des procédures standardisées de contrôle de la qualité des traductions. Les projets internationaux de grande envergure utilisent de plus en plus le développement simultané afin d'éviter que la version développée dans une langue ne puisse être traduite/adaptée à toutes les langues de l'étude.

Double Traduction et Recoupement. Dans ce protocole de traduction, un traducteur indépendant ou un groupe d'experts identifie et résout les divergences entre les différentes traductions unidirectionnelles afin de les réconcilier/harmoniser en une version unique.

EMDP. Acronyme d'Échelonnement Multidimensionnel Pondéré. Il s'agit d'une autre procédure statistique permettant d'investiguer la dimensionnalité des tests.

Fonctionnement Différentiel des Items (FDI). Il existe une classe de procédures statistiques qui peuvent déterminer si un item fonctionne plus ou moins de la même façon dans deux groupes différents. Les comparaisons de performance sont faites en appariant d'abord les répondants sur le caractère mesuré par le test. Lorsque des différences sont observées, on dit que l'item est potentiellement biaisé. On s'efforce alors, d'expliquer les différences conditionnelles de performance pour les candidats des deux groupes appariés sur le caractère mesuré par l'item.

Formule de Kuder-Richardson 20. (Ou, parfois simplement appelé "KR-20"). Le coefficient de fiabilité d'un test formé à partir d'items binaires, qui sont supposés mesurer un attribut commun et possédant tous un potentiel discriminant identique.

Localisation. Il s'agit d'un terme populaire dans le domaine du testing qui est utilisé pour décrire le processus permettant de rendre acceptable un test préparé dans une langue et une culture pour une utilisation dans une autre. Un terme équivalent serait traduction/adaptation.

Mise en Équivalence des scores au test. Une procédure statistique pour jumeler les scores à deux tests mesurant le même construit mais qui ne sont pas strictement parallèles.

Modélisation par Équations Structurales. Un ensemble de modèles statistiques complexes qui sont utilisés pour identifier la structure sous-jacente d'un test ou d'un ensemble de tests. Souvent, ces modèles sont utilisés pour étudier les inférences causales des relations entre un ensemble de variables.

Oméga (ou parfois appelé "Coefficient Oméga" ou "Oméga de McDonald"). Le coefficient de fiabilité d'un test dont les items sont supposés mesurer un facteur commun (s'applique au modèle d'analyse factorielle à un facteur commun ou général). Plus généralement applicable que le coefficient Alpha.

PISA. Acronyme du "Programme for International Student Achievement" (Prgramme International pour le suivi des acquis des élèves). Il s'agit de l'évaluation internationale de performance scolaire qui est parrainée par l'Organisation de Coopération et de Développement Économiques (OCDE) avec plus de 40 pays participants.

Procédure de Mantel-Haenszel (MH) pour l'Identification du Fonctionnement Différentiel des Items (FDI). Une procédure statistique pour comparer la performance de deux groupes de répondants sur un item de test. Les comparaisons sont faites pour les répondants de chaque groupe qui sont appariés sur le trait ou le construit mesuré par le test.

Régression Logistique (RL) pour l'Identification du Fonctionnement Différentiel des Items(FDI). Cette procédure statistique est une façon supplémentaire d'exécuter les analyses FDI. Une courbe logistique est ajustée aux données de performance de chaque groupe, puis les deux courbes logistiques, une pour chaque groupe de langues, sont comparées statistiquement.

Répondants. Terme utilisé de façon interchangeable dans le domaine du testing avec " personnes examinées ", " candidats ", " personnes testées " et " élèves " (s'il s'agit de tests académiques).

Test de dimensionnalité. Il s'agit du nombre de dimensions ou de facteurs mesurés par un test. Souvent, cette analyse est effectuée statistiquement à l'aide d'une des nombreuses procédures, y compris les

diagrammes de valeurs propres ou la modélisation par équations structurales.

Théorie de réponse à l'item (TRI). Une classe de modèles statistiques pour relier les réponses des items à un trait ou à un ensemble de traits qui sont mesurés par les items du test. Des modèles spécifiques de TRI peuvent traiter des données de réponse dichotomiques/binaires et polytomiques. Les données binaires peuvent provenir de la notation d'items à choix multiples ou d'items vrai-faux d'une échelle de personnalité. Les données de réponse polytomiques peuvent provenir de la notation des tâches de performance et ou des essais/tâches pratiques d'un test de rendement, ou d'échelles de réponse telles que " Likert".

TIMSS. Acronyme de "Trends in International Mathematic and Science Studies" (Tendances internationales dans l'enseignement des mathématiques et des sciences). Il s'agit d'une évaluation internationale des élèves de 4e (CM1), 8e (Quatrième) et 12e (Terminale) de différents pays dans les domaines des mathématiques et des sciences, parrainée par l' International Association for the Evaluation of Educational Achievement (IEA).

Traduction unidirectionnelle. Avec ce protocole, un test est transposé dans la langue cible d'intérêt par un traducteur ou, plus souvent, un groupe de traducteurs, puis un autre traducteur ou groupe de traducteurs juge de l'équivalence des versions du test en langue source et en langue cible.

Traductions bidirectionnelles/à Rebours/Inversées. Avec ce protocole, un test est traduit de la version de langue source à la version de langue cible par un groupe de traducteurs, puis la version de langue cible est à son tour traduite dans la langue source, par un deuxième traducteur ou groupe de traducteurs. La version source originale et la version source issue de la traduction à rebours/inversée sont comparées et un jugement est porté sur la compatibilité de la version du test dans la langue source. Si les deux versions en langue source sont très proches, on suppose que la version en langue cible du test est acceptable.

Valeurs/Indices Delta. Les valeurs/indices delta sont simplement des valeurs p (c.a.d, la proportion de réponses correctes à l'item) transformées de manière non linéaire et appliquées à des items binaires. Une valeur delta d'item est l'écart normal correspondant à l'aire sous une distribution normale (moyenne= 0,0; *SD* = 1,0) où l'aire sous la distribution normale est égale à la proportion de candidats répondant correctement à l'item. Ainsi, si p = 0,84, la valeur delta de l'item serait de 1,0. Cette transformation est effectuée sous le postulat que les valeurs/indices delta sont plus susceptibles d'être sur une échelle d'intervalles égales que les valeurs p.

Version de la langue cible. La langue dans laquelle un test est traduit/adapté. Ainsi, par exemple, si un test est traduit de l'anglais à l'espagnol, la version anglaise est souvent appelée " version de langue source

" et la version espagnole est appelée " version de langue cible ".

Version de la langue source. La langue dans laquelle un test est écrit à l'origine.