Adaptation transculturelle de tests et échelles de mesure psychologiques : guide pratique basé sur les Recommandations de la Commission Internationale des Tests et les Standards de pratique du testing de l'APA

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

Adaptation transculturelle de tests et échelles de mesure psychologiques: Guide pratique basé sur les Recommandations de la Commission Internationale des Tests et les Standards de pratique du testing de l'APA

Transcultural Adaptation of Psychological Tests and Scales: A Practical Guide Based on the ITC Guidelines for Translating and Adapting Tests and the Standards for Educational and Psychological Testing

Kamel Gana[1], Nedjem Eddine Boudouda[2], Samia Ben Youssef[3], Nicolas Calcagni[1] & Guillaume Broc[4]

[1] université de Bordeaux, France

[2] Université 8 mai 1945, Guelma, Algérie

[3] Université de Tunis, Tunisie

[4] Université de Montpellier, France

Auteur pour la correspondance

Kamel Gana, Professeur de psychologie

Kamel.gana@u-bordeaux.fr

Adaptation transculturelle de tests et échelles de mesure psychologiques: Guide pratique basé sur les *Recommandations* de la Commission Internationale des Tests et les *Standards* de pratique du testing de l'APA

Transcultural Adaptation of Psychological Tests and Scales: A Practical Guide Based on the ITC Guidelines for Translating and Adapting Tests and the Standards for Educational and Psychological Testing

Article pour le Numéro Spécial: «**Cadres méthodo-épistémologiques actuels pour la construction et l'adaptation transculturelle de tests et échelles de mesure psychologiques**»

Résumé

Il y a 30 ans, Vallerand (1989) écrivit un article précurseur, intitulé «Vers une méthodologie de validation trans-culturelle de questionnaires psychologiques: Implications pour la recherche en langue française», dont le succès ne s'est jamais démenti, ainsi que son utilité d'ailleurs. Toutefois, des avancées considérables ont été réalisées dans le domaine de construction et d'adaptation de tests et échelles de mesure depuis la parution de cet article, notamment avec les publications (et leurs mises à jour) du *Guidelines for Translating and Adapting* Tests par la Commission Internationale des Tests (ITC), et des *Standards for Educational and Psychological Testing* par, conjointement, l'American Educational Research Association (AERA), l'American Psychological Association (APA), et le National Council on Measurement in Education (NCME). Le présent article se propose de mettre à jour la procédure proposée par Vallerand en se fondant sur les dernières *Recommandations* de l'ITC (2017) et les *Standards* (AERA, APA, NCME, 2014). Sans prétendre les remplacer, bien au contraire, notre objectif ici étant de proposer une procédure d'adaptation de tests et échelles de mesure psychologiques - développés à l'origine pour d'autres cultures et/ou langues- en 10 étapes, sorte de canevas guidant pas-à-pas l'adaptation d'un test et la validation de ses scores.

Mots-clés : Adaptation de tests, Recommandations de l'ITC, Standards de l'APA, testing, adaptation transculturelle de tests

Abstract

Thirty years ago, Vallerand (1989) wrote a pioneering article entitled "Toward a methodology for the transcultural validation of psychological questionnaires: Implications for research in the French language", whose success has never failed, as well as its usefulness. However, considerable progress has been made in the field of test development and test adaptation since the publication of this article, notably with the publications (and their updates) of the *ITC Guidelines for Translating and Adapting Tests* (2017) by the International Test Commission, and of the *Standards for Educational and Psychological Testing* (2014) by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). Based on these two key references, the present article aimed to update the procedure proposed by Vallerand. Without claiming to replace these major references, on the contrary, our objective here is to propose a 10-step procedure for adapting tests and psychological measures originally developed for other cultures and/or languages, a sort of framework guiding in step-by-step manner the adaptation of tests and the validation of their scores.

Key words: Test adaptation, ITC Guidelines, Standards, testing, cross-cultural test adaptation

Henri Piéron (1881-1964)

«Until the phenomena of any branch of knowledge have been

subjected to measurement and number,

it cannot assume the status and dignity of a science »

(Galton, 1822-1911)

«The history of science is the history of measurement»

(Cattell, 1860-1944)

Dans les sociétés multilingues, multiculturelles, multirégionales, développer des tests a-culturels (*culture free*)[1] ou plutôt, culturellement équitables (*culture-fair*)- c'est à dire une évaluation culturellement impartiale, débarrassée de toute forme d'influences ethnico-culturelles sans lien direct au construit et pouvant en biaiser la mesure- est une ancienne exigence d'équité et d'éthique (Cattell, 1940). Exporter un test ou un instrument de mesure conçu dans une culture et l'adapter pour s'en servir dans une autre culture obéit aux mêmes exigences d'équité et d'éthique. Et ceci est d'autant plus vrai et nécessaire que la culture source et la culture cible sont sensiblement différentes. Comparer, classer, distinguer, différencier, et dégager les invariants universels dans le fonctionnement psychique humain exigent des méthodes communes et des outils de mesure équivalents. La pire situation pour la psychologie scientifique est de se retrouver avec autant d'outils de mesure que de chercheurs, autant d'instruments de mesure que de groupes socio-ethniques, autant de tests que de psychologues.

Cependant, il est nécessaire de fixer les idées, d'emblée. D'abord, la position que nous défendons dans le présent article n'incite pas à gommer les différences ethnico-culturelles, mais au contraire à en tenir compte, pour mieux les contrôler, dans toute procédure d'adaptation de tests. Peut-on, par exemple, comparer correctement les performances

[1] Un test a-culturel au sens strict du terme est inconcevable (voir Frijda & Jahoda, 1966).

cognitives (e.g., résolution de problème) d'enfants issus de pays différents si l'un des items comprenait des énoncés exotiques voire irréels pour certains enfants dans certains pays (e.g., administrer un item de complètement d'images présentant un «paysage enneigé» à des populations d'Afrique subsaharienne). Ensuite, par « guide», nous entendons un canevas souple, sorte de cadre général structurant et orientant la pratique d'adaptation transculturelle d'outils de mesure psychologiques. Enfin, ce canevas, qui se propose de mettre à jour le travail pionnier de Vallerand (1989), se fonde, sans prétendre les remplacer, sur le *Guidelines for Translating and Adapting Tests* de l'International Test Commission (ITC), et les *Standards for Educational and Psychological Testing* de l'American Educational Research Association (AERA), l'American Psychological Association (APA), et le National Council on Measurement in Education (NCME).

Équivalence et Biais

Adapter un test ne se réduit pas simplement à le traduire et le rendre en une langue différente. Adapter un test exige de le rendre équivalent-au niveau psychologique- à la version originale. Quand bien même nécessaire, cette équivalence ne peut être uniquement linguistique, elle est: (a) d'abord pragmatique qui, comme la sémantique, prend en compte la signification de l'item (Armengaud, 2007), c'est-à-dire qu'elle n'est pas simplement dénotative elle doit être surtout connotative; (b) cette équivalence est aussi conceptuelle (équivalence de construit): le construit mesuré par le test a-t-il son équivalent dans la culture cible? Comment adapter les éventuelles différences relatives au construit en présence? (c) cette équivalence est aussi culturelle: le test source est-il saturé de références culturelles? Et comment les transposer dans la version adaptée et les ajuster aux spécificités culturelles du groupe cible? Dans ce sens, l'équivalence culturelle plaide en faveur de l'universalité du contenu du test tandis que la variance culturelle plaide au contraire pour l'existence de spécificités culturelles d'un groupe à l'autre; et enfin, (d) cette équivalence est métrologique

ayant trait aux procédures de mesurage et au format et au matériel du test. Le format du test source est-il approprié à la culture cible? Par exemple, le format papier-crayon, l'administration en face-à-face, ou l'échelle de réponse de type Likert sont-ils appropriés à la culture cible? (van de Vijver, 2016).

Ainsi, l'adaptation optimale d'un test à des langues et cultures différentes vise à réduire les multiples biais tout en cherchant à augmenter l'équivalence avec la version source. Il y a trois principaux biais. Les biais de construit (*construct bias*) d'abord. Ceux-ci trahissent l'existence de signification différente du construit d'une culture à l'autre allant à l'encontre d'une équivalence-nécessaire et recherchée- de construit. Les différences culturelles relatives aux liens entre le construit et ses manifestations –comportementales- y sont pour beaucoup dans ces biais. La notion « bon époux» ou «bonne épouse» pour mesurer le construit «satisfaction conjugale» est potentiellement saturée culturellement. Les biais de méthode ensuite (*method bias*). On y trouve, entre autres, les biais générés par un échantillonnage inadéquat, un format de réponse ou une modalité d'administration inappropriés à la culture cible. Les biais d'item (*item bias*), enfin, qui renvoient aux distorsions au niveau de chaque item. Un item est considéré comme étant biaisé lorsqu'il suscite une signification différente d'un groupe à l'autre (Byrne, 2016). Le lecteur trouvera dans Van de Vijver et Poortinga (1997) des détails précis concernant les différentes sources de ces biais.

Biais et équivalence sont des préoccupations majeures dans la procédure d'adaptation transculturelle de tests: comment en détecter la présence, en éviter les effets tout en augmentant l'équivalence entre les différentes versions du test?

Les procédures d'adaptation de tests et objectifs du présent travail

Plusieurs auteurs ont proposé des procédures d'adaptation d'outils de mesure opérationnalisées en étapes (Bracken & Barona, 1991; Geisinger, 1994; Gudmundsson,

2009; Krach, McCreery, & Guerard, 2017; Ljungberg, Fossum, Fürst, & Hagelin, 2015; Sidani et al., 2010; Sousa & Rojjanasrirat, 2011; Vallerand, 1989). Au demeurant, la proposition de Vallerand (1989) est de loin la plus connue et la plus utilisée en francophonie[2]. Il convient de reconnaître que Vallerand (1989) fut un vrai précurseur et sa proposition tombait à point nommé dans un contexte de forte demande d'outils de mesure pour conduire des recherches scientifiques, et où les traductions à la sauvette de mesures psychologiques, essentiellement américaines, pullulaient.

La procédure de validation transculturelle d'instruments de mesure proposée par Vallerand (1989) se décline en 7 étapes : (1) la préparation de la version préliminaire; (2) l'évaluation et la modification de la version préliminaire; (3) l'évaluation de la clarté des questions par des membres de la population cible dans un pré-test; (4) l'évaluation de la validité concomitante et de contenu du questionnaire; (5) l'évaluation de la fidélité test-retest et de consistance interne de l'instrument; (6) l'étude de la validité de construit du questionnaire; (7) l'établissement de normes.

Nous proposons dans le présent article de mettre à jour la procédure suggérée par Vallerand (1989) en se basant sur les deux cadres méthodo-épistémologiques de la psychométrie et du testing actuels: les *Recommandations* de la Commission Internationale des Tests (International Test Commission, ITC, 2017) et les *Standards* (AERA, APA, & NCME, 2014) (concernant ces deux cadres voir Boudouda et ses collaborateurs dans ce Numéro Spécial). Tout le long de cet article, ces deux cadres seront désignés par les raccourcis suivants: les *Recommandations* de l'ITC et les *Standards* (ou *Standards* de l'APA)[3].

[2] Selon Google Scholar, l'article de Vallerand (1989) a été cité 1067 fois en trente ans (depuis sa parution en 1989 jusqu'à fin décembre 2019), soit en moyenne 35 fois par an. Il est l'un des articles le plus cité parmi les travaux de Vallerand.

[3] Le lecteur trouvera dans ce numéro spécial la traduction française des *Recommandations* de l'ITC.

On notera avec regret que les *Standards* ainsi que les *Recommandations* de l'ITC sont rarement cités comme références dans les travaux de construction et d'adaptation de tests (Rios & Sireci, 2014). Notre objectif ici étant de proposer une procédure d'adaptation en langue française[4] de tests et de mesures psychologiques -validés en langues étrangères- en 10 étapes, sorte de canevas méthodologique structurant et guidant pas-à-pas l'adaptation d'un test et la validation de ses scores. Arrimé aux *Recommandations* de l'ITC (voir la version française de ces *Recommandations* dans ce Numéro spécial) et aux *Standards* de l'APA, ce canevas synthétise la démarche d'adaptation et validation de tests, sans prétendre les remplacer pour autant. Bien au contraire, à chaque étape, ces deux cadres pourraient fournir des détails techniques précieux voire indispensables aux chercheurs et praticiens de tous bords. La Figure 1 présente les différentes étapes de la procédure d'adaptation et validation de tests. Comme le suggérait, à juste titre, Vallerand (1989), le passage d'une étape à l'autre exige d'avoir satisfait aux exigences requises de l'étape précédente. Le retour à l'étape précédente appelle une réévaluation du test. Le retour à l'étape initiale de traduction/adaptation pourrait parfois se révéler nécessaire.

Il convient de préciser clairement et avec force que ce canevas doit rester souple et s'ajuster au type de test à adapter: tests cognitifs (aptitude, performance…) versus tests non-cognitifs (personnalité, attitude…). Le vocable « test » renvoie ici à tout instrument de mesure psychologique qu'il soit étalonné ou non, destiné à la pratique ou à la recherche. Il ne s'agit pas ici de corseter une pratique mais plutôt lui offrir un cadre général souple et structurant. La pire situation serait, nous semble-t-il, de se retrouver avec autant de procédures d'adaptation de tests que d'auteurs d'adaptation.

[4] Nous nous adressons ici aux lecteurs francophones. Toutefois, notre procédure s'applique à toute adaptation de tests en une autre langue que le français (e.g., l'arabe, l'italien, le mandarin, le serbe…).

Il est un autre point qui mérite d'être rappelé et souligné ici. Il concerne le choix du test qui fera l'objet d'adaptation. Un tel choix pourrait être dicté par la popularité du test (WAIS, Beck Depression Inventory….) ou peut-être par l'absence d'alternatives possibles. Dans tous les cas, il faut s'assurer que le test d'origine possède des qualités psychométriques solides. D'ailleurs, le choix du test à adapter mériterait de constituer une étape préalable à part entière dans un protocole d'adaptation. En effet, en présence de plusieurs instruments mesurant le même construit, il est nécessaire de justifier et bien argumenter le choix de celui retenu pour l'adaptation.

Insérer Figure 1 ici

ÉTAPE 1. Demander l'autorisation auprès du titulaire du droit d'auteur (copyright)

Il est inutile ici de reprendre les recommandations assez précises de l'ITC concernant une telle autorisation (voir la traduction française de ces *Recommandations* dans ce Numéro Spécial). Rappelons seulement que celle-ci concerne aussi bien les tests/échelles de mesure protégés que ceux qui tombent dans le domaine public. Contacter, si possible, le(s) auteur(s) pourrait se révéler utile à plus d'un titre : (1) ces auteurs pourraient être au courant d'une éventuelle adaptation, soit en cours soit déjà aboutie, de leur test ; (2) ils pourraient être sollicités au cours de la procédure d'adaptation afin d'apporter des précisions utiles relatives au sens de certains items; (3) leur avis pourrait importer en cas de décision d'abréger le test/échelle de mesure.

Les auteurs de la version originale sont des véritables atouts dans une procédure d'adaptation de tests. Les informer que la procédure d'adaptation de leur test adopte le cadre des *Recommandations* de l'ITC et celui des *Standards* de l'APA pourrait favoriser leur consentement. Les remercier dans la publication des résultats de l'adaptation va de soi et ce, quelle que soit leur contribution à cette adaptation.

ÉTAPE 2. Traduire le test

Bien qu'une bonne traduction ne signifie pas une adaptation transculturelle de qualité, une mauvaise traduction compromet, quant à elle, toute adaptation de qualité. Il y a d'abord l'importance qu'il faut accorder à l'oral (parlé) plutôt qu'à l'écrit d'une langue lors d'une transposition d'un texte. Le traducteur doit d'abord se demander « Que signifie cet item dans le test source », et traduire ensuite cette compréhension en mots accessibles dans la langue cible. Il doit produire une traduction qui ne réduit ni n'élargit les informations au point de perdre le sens de l'item source. Il est primordial que les items traduits déclenchent le même stimulus que les items sources. Ceci correspond à l'approche appelée «poser la même question/même item». Cependant, il peut parfois s'avérer impossible d'assurer une traduction totalement équivalente, en particulier dans le cas où les deux langues ne disposent pas du tout de termes ayant la même connotation sémantique ou de concepts équivalents. Dans ces cas précis, il convient de rechercher la meilleure approximation possible du sens original. Il ne s'agit pas ici de traduction littérale, mot-à-mot. Adapter un item ne signifie pas en créer un nouvel item ni en proposer un autre différent. Il s'agit plutôt d'ajuster la signification de l'item source (e.g., «to get the blues» ou «to have goosebumps») aux spécificités culturelles et linguistiques de la langue cible («avoir le cafard» ou «avoir la chair de poule»)

Il convient ici de prévoir au moins deux traductions parallèles réalisées par deux traducteurs dont les compétences requises sont bien définies dans les *Recommandations* de l'ITC. Le choix des traducteurs est crucial ici pour parvenir à une traduction adéquate. Maitriser (parler et écrire) deux langues est une condition nécessaire mais parfois insuffisante pour être un bon traducteur de tests pour ces langues. La traduction de tests exige non seulement des qualifications en science de la traduction (traductologie), mais aussi des connaissances en psychométrie (e.g., une certaine familiarité avec la rédaction d'item, format d'item, échelle de réponse), et une appropriation suffisante des codes socio-culturels de la

culture cible. Les traducteurs de tests ne doivent pas être obligatoirement des traducteurs professionnels, ils peuvent être des psychologues autochtones bilingues possédant une expérience suffisante dans la traduction de tests et mesures psychologiques. Toutefois, l'adaptation de tests se doit de mobiliser les compétences de toute une équipe (Harkness, Villar, & Edwards, 2010).

ÉTAPE 3. Approuver la version traduite/adaptée du test

Cette phase incombe à l'équipe de chercheurs/praticiens engagée dans l'adaptation du test à laquelle doivent se joindre les traducteurs. En l'absence des traducteurs du test, une ou plusieurs personnes possédant de solides compétences bilingues pourraient participer au processus d'examen et évaluation des traductions. Toutefois, souvent les adaptations sont motivées moins par les caractéristiques de la langue cible que par l'exigence de répondre et coller aux sensibilités sociales, culturelles et contextuelles du nouveau groupe linguistique. Il s'agit de bien rendre en français (ou toute autre langue cible du test à adapter) le contenu psychologique des items.

Le fait d'impliquer les traducteurs dans le processus d'examen et d'évaluation des traductions non seulement améliore l'évaluation mais peut également l'accélérer. L'équipe doit passer en revue chaque item, échelle de réponse, consigne et tout autre aspect du matériel du test. Elle doit procéder à la comparaison et la confrontation des traductions en présence tout en envisageant d'autres alternatives. Les membres de l'équipe sont invités à identifier les points faibles et les points forts des traductions proposées, et à soulever tous les problèmes qui se posent, telles que la comparabilité avec le texte source, les adaptations nécessaires, et les difficultés dans le texte source. Ici chaque élément du test (items, échelle de réponse, consigne et tout autre matériel) sera passé en revue en confrontant ses traductions disponibles de manière à aboutir à une adaptation qui fasse consensus et obtienne l'approbation de toute

l'équipe. La procédure d'évaluation se termine lorsque l'ensemble des éléments du test adapté obtient l'approbation de tous les membres de l'équipe.

Dès lors que le but de cette procédure d'examen est non seulement d'améliorer la traduction mais d'en faire une adaptation chaque fois que cela est nécessaire, la discussion et l'évaluation sont au cœur du processus. Ainsi, en cas de litige sérieux sur le sens d'un item, d'un mot ou d'une expression, il est conseillé de solliciter, si possible, l'avis de l'auteur(s) de la version originale.

Ce travail évaluatif permettra éventuellement de passer de la traduction parfois littérale (recherchée parfois involontairement par les traducteurs) à une réelle adaptation c'est-à-dire une transposition sémantique du texte d'une langue à l'autre. C'est la raison pour laquelle, nous déconseillerions la procédure des traductions à rebours (inversées, bidirectionnelles) car celles-ci sous-tendent la recherche d'une équivalence plutôt linguistique et littérale que sémantique et pragmatique[5]. Bien qu'une traduction purement littérale puisse être une bonne traduction, elle peut échouer à être une bonne adaptation quand elle enclot le risque de ne pas pouvoir capter les subtilités sémantiques de la langue cible et ne pas bien rendre le contenu psychologique des items.

Enfin, il est utile de rendre compte de tous les détails concernant cette procédure d'évaluation et d'approbation de la version adaptée du test. Il s'agit de mettre en exergue la transparence et la qualité du travail d'adaptation.

ÉTAPE 4. Pré-tester la version approuvée par l'équipe

[5]Rappelons qu'une traduction à rebours (back-translation) se déploie en trois étapes : a) transposer un texte d'une langue source à une langue cible, b) transposer ensuite cette version traduite à sa langue source (e.g. retraduction inversée par un autre traducteur dans la langue source la version traduite), c) comparer les versions source et inversée. La traduction est jugée satisfaisante lorsque la traduction inversée correspond à la version source.

Le pré-test est essentiel pour améliorer la version traduite/adaptée d'un test. Il permet de s'assurer non seulement de la clarté et la compréhensibilité de la consigne et des items mais aussi de leur équivalence sémantique d'avec la version source. Plusieurs techniques peuvent être engagées dans le cadre du pré-test.

a) Le recours aux experts d'abord: il s'agit de faire examiner la version approuvée par l'équipe auprès d'un panel de méthodologistes, de spécialistes du domaine du test, et des traducteurs expérimentés.

b) Le focus groupe (Ryan, Gandha, Culbertson, & Carlson, 2014; Vogt, King, & King, 2004), qui est un type particulier d'entretien de groupe où l'animateur/modérateur pose une série de questions ciblées conçues pour susciter des opinions collectives sur un sujet précis. il s'agit ici de réunir un petit groupe de personnes (6-10) appartenant à la population cible du test pour discuter de sujets bien déterminés (e.g., le sens d'un item, d'un mot, d'un énoncé, de la consigne) de manière relativement peu structurée, dirigé par un modérateur qui veille à ce que la conversation ne s'écarte pas du sujet en question, et à ce que tous les participants y prennent part effectivement.

c) L'entretien cognitif (Beatty & Willis, 2007; Daouk-Öyry & McDowal, 2013; García, 2011; Willis, 2005) : à la lecture d'un item (considéré comme stimulus), un répondant s'engage dans un processus de traitement de l'information multiple avant d'y répondre. Il procède d'une manière bien structurée: (a) il interprète l'item, (b) récupère de l'information de sa mémoire, (c) décide comment répondre, et enfin (d) choisit une réponse parmi les catégories de réponses qui correspond à son état d'esprit. La manière de comprendre l'item et d'en déchiffrer le sens est cruciale ici. Pour s'assurer que le répondant ait compris le sens de l'item ainsi que la signification des échelles de réponse, on peut recourir ici à deux techniques au moins : (a) technique de la réflexion à haute voix qui permet de saisir les pensées

qu’entretiennent les répondants pendant qu’ils répondent à un item, (b) technique de la réflexion à haute voix rétrospective : il s’agit de s’entretenir avec les répondants après qu'ils aient complété le test sur la façon dont ils ont trouvé des réponses à chaque item (psychologie de réponse à l’item).

d) L'oculométrie (eye tracking) peut être utilisée pour examiner la compréhension de l'item/texte. Elle permet de montrer: (a) les parties spécifiques du texte sur lesquelles les participants se concentrent (ou ne se concentrent pas), et (b) combien de temps cette focalisation dure. Le temps de réponse à chaque item peut aussi être un indicateur intéressant à utiliser aussi (Wood, Harms, Lowman, & DeSimone, 2017).

Ces techniques ainsi que l’analyse des stratégies de réponse permettent de prouver que les réponses aux items sont plus intentionnelles (elles sont le fruit de l'intention du participant) qu'accidentelles.

e) L’entretien ethnographique (Bauman & Adair, 1992) : pour tester la qualité d’une traduction/adaptation d'un test issu d’une culture source fort différente des observations spécifiques à la culture ou à la langue cible peuvent être nécessaires. Il s’agit de sonder et fureter au fond des schémas de pensées culturels (manière d’être, de paraître et d’agir au sein d’une culture donnée) afin de comprendre la signification sans ambiguïté de certains énoncés et s’assurer de leur équivalence d’avec la culture source. Les techniques ethnologiques mettent l'accent sur les variables culturelles, telles que les systèmes de croyances, de représentations et les pratiques quotidiennes qui déterminent si un item a un sens ou non au sein de la culture cible. Dans ce sens, les entretiens ethnographiques posent des questions plus larges que les entretiens cognitifs. Ils peuvent être utilisés pour identifier des schémas culturels pour parvenir à trouver les mots/énoncés adéquats d’un item. Il s'agit d'entretiens non structurés et non directifs qui visent à comprendre le background culturel de la personne

interrogée afin de pouvoir lui proposer les items les plus appropriés à ses schémas culturels tout en conservant le sens original de ces items (Willis, 2005).

f) L'évaluation par questionnaire: il s'agit ici de demander à un petit échantillon «représentatif» de la population cible d'évaluer la clarté de la consigne et de chaque item sur une échelle Likert allant de «pas du tout clair» à « tout à fait clair ». Lorsque le test est conçu pour une population générale, Il est parfois recommandé de réaliser ce type de test auprès d'un échantillon possédant un niveau éducatif moyen voire modeste. Des formulations compréhensibles aux gens d'un niveau socio-éducatif modeste le seront aussi et probablement aux autres.

On notera que certaines de ces stratégies sont à même d'apporter des indices de validité basés sur les processus de réponse (sous-tendant les réponses) ainsi que des indices de validité basés sur le contenu permettant d'étayer les interprétations susceptibles d'être tirées des scores au test (voir *Standards*, 2014).

Les résultats de ce pré-test pourraient amener à des modifications nécessaires de la version approuvée précédemment. Il est primordial ici de revenir à l'étape 3 et de procéder à l'approbation collective argumentée de toute modification éventuelle. La version adoptée *Mutatis mutandis* constitue la version adaptée quasi-définitive du test. C'est sur cette version que l'on procédera à l'analyse d'items décrite dans l'étape suivante.

ÉTAPE 5. Réaliser une analyse d'items

Cette étape nécessite l'administration de la version adoptée, adaptée et approuvée à l'issue du pré-test (Étape 4) à un échantillon (N > 300) «représentatif» (autant que faire se peut) de la population cible et ce, afin d'analyser les items. Cette analyse vise à examiner les caractéristiques essentielles des items: moyenne, médiane, variance, normalité de leur distribution, leur degré de difficulté et leur potentiel de discrimination. Les items sont-ils

potentiellement entachés de désirabilité sociale? Sont-ils potentiellement biaisés (e.g., sensibles au sexe du participant, à son statut social…)? (voir les *Recommandations* de l'ITC). La normalité univariée et multivariée des items déterminera le choix de certaines analyses et méthodes statistiques à utiliser ultérieurement (Gana & Broc, 2018). Bien qu'elle puisse faire partie de l'analyse d'items, la contribution d'un item à la fiabilité générale des scores trouve mieux sa place à l'étape suivante car cette l'analyse dépend du coefficient de fiabilité choisi : alpha de Cronbach (donc «*alpha if item deleted*») ou bien un coefficient basé sur un modèle de mesure (Raykov, 2007, 2008). Nous y reviendrons.

ÉTAPE 6. Apporter/fournir des preuves de validité basées sur la structure interne du test

La structure interne pose le modèle de mesure du test, renvoyant ainsi à sa représentation factorielle (Gana & Broc, 2018). Tout test est un modèle de mesure dont on peut tester la vraisemblance. En principe, la structure factorielle d'un test est posée dès le début par l'auteur du test dès lors qu'elle renvoie aux composantes (dimensions) constitutives du test. Elle a partie liée avec le contenu du test et ses fondements théoriques et définitoires, sauf si l'auteur s'en remet, peu vraisemblable, aux analyses factorielles pour déterminer, de façon a-théorique, la structure de son test. C'est la raison pour laquelle, elle est la première à être soumise à l'épreuve des faits dans une procédure de validation. Interroger la structure interne d'un test c'est questionner en partie sa validité de contenu. Nous avançons cinq suggestions ici :

a) Envisager une étude empirique portant en même temps sur l'analyse d'items (Étape 5) et la structure interne du test (Étape 6) sur un échantillon dont la taille devrait dépendre de la complexité du modèle et de la nature des données (Gana & Broc, 2018). Bien que certains auteurs considèrent qu'à partir de 300 répondants l'échantillon est suffisant pour ces analyses

(Worthington & Whittaker, 2006), il convient de souligner que plus l'échantillon est large plus les résultats en seront robustes et stables.

b) Après l'analyse d'items, il faut réaliser d'abord une analyse factorielle confirmatoire -en tenant impérativement compte du format d'items (dichotomique, polytomique, continu) et de leur distribution multivariée- et, ce faisant, détecter les sources d'une éventuelle inadéquation du modèle aux données (Gana & Broc, 2018).

c) Dans le cas d'inadéquation, il convient de réaliser une analyse factorielle exploratoire ou une *Exploratory Structural Equation Modeling* (ESEM ; Asparouhov & Muthén, 2009). On notera ici que c'est la méthode des axes principaux (ou du maximum de vraisemblance) qu'il faut choisir et non l'analyse en composantes principales qui n'est pas, à proprement parler, une analyse factorielle (Fabrigar, Wegener, MacCallum, & Strahan, 1999)[6].

d) Considérée comme un coefficient de validité, la valeur seuil d'acceptabilité d'une saturation factorielle est capitale. Sa significativité statistique n'est pas suffisante (McNeish, An, & Hancock, 2018). Élevé au carré, le coefficient de saturation exprime le % de variance de l'item explicable par le facteur dont il dépend. Selon Guadagnoli et Velicer (1988), une saturation de .40 représente une limite inférieure typique, une saturation de .60 est une saturation modérée alors qu'une saturation de .80 est une excellente saturation. Une saturation de .40 signifie que 16% de la variance de l'item sont imputables au facteur (construit) dont il est le représentant. Une saturation factorielle de .80 signifie que 64% de la variance de l'item dépendent du facteur dont il est l'indicateur (le représentant). La validité de l'item est d'autant meilleure que ce pourcentage est proche de cent. Même si l'item parfait n'existe pas. En effet, que vaut un item quelconque dont la saturation par le facteur/construit (e.g., Humeur dépressive, Capacités mnésiques) dont il est censé être l'indicateur/représentant est de .30 ou

[6] Fabrigar et al. (1999) écrivent "Thus, although principal component analysis (PCA) is often referred to and used as a method of factor analysis, it is not factor analysis at all" (p. 275).

.40 ? Que vaut un item dont la part de variance imputable au facteur (construit) dont il est l'indicateur ne dépasse pas les 20% ? Sachant que le reste (80%) est imputable à autres choses et, principalement, aux erreurs de mesure. Est-il, alors, un bon indicateur, ou un bon représentant de ce construit ?

e) Ne pas retrouver la structure interne d'origine avec la version adaptée du test et/ou se retrouver avec des saturations factorielles problématiques (faibles saturations factorielles, saturations croisées, saturations par un autre facteur que celui dont devait dépendre initialement l'item…) invitent à en comprendre les raisons afin de parvenir à les améliorer en reprenant généralement le travail initial de traduction et d'adaptation (Étapes 3 voire Étape 2). A ce stade, il est déconseillé de supprimer des items ou de tenter d'abréger le test. Il convient plutôt de réviser la version adaptée (en revenant à l'Étape 3) et de la resoumettre ensuite à l'épreuve des faits en l'administrant à un nouvel échantillon de participants.

ÉTAPE 7. Apporter/fournir des preuves basées sur la fiabilité/précisions des scores

Trois cas de figure peuvent se présenter à ce stade.

1) L'étape précédente a été franchie avec succès, c'est à dire la structure factorielle confirmée et les saturations factorielles satisfaisantes, les mêmes données peuvent alors servir à l'analyse de fiabilité/fidélité pour apporter les preuves plaidant en faveur du degré de précision des scores au test (voir les *Standards* de l'APA pour la définition de la fiabilité des scores et les *Recommandations* de l'ITC pour les analyses statistiques appropriées). Ainsi, connaître la structure interne du test (unidimensionnelle, multidimensionnelle, hiérarchique, bifactorielle) et la confirmer sont deux prérequis pour le calcul de certains coefficients de fiabilité (Gerbing & Anderson, 1988; Green & Yang, 2015). Le choix de l'indice de fiabilité dépend du modèle mesure en présence. D'ailleurs, il convient de bien justifier ce choix. Les coefficients de fiabilité basés sur les modèles nécessitent de tester au préalable le modèle de

mesure (via une analyse factorielle confirmatoire) dont les résultats serviront ensuite à calculer ces coefficients (e.g., le coefficient oméga).

Le test-retest (i.e., stabilité temporelle)[7] et le coefficient alpha de Cronbach (Cronbach, 1951)[8], désormais assorti de son intervalle de confiance, sont les deux estimations de la fiabilité les plus couramment utilisées (Dunn, Baguley, & Brunsden, 2013).

Le coefficient alpha mérite qu'on s'y attarde tant il est l'indice de fiabilité le plus utilisé mais le plus galvaudé aussi (Béland & Cousineau, 2018; Laveault, 2012; McNeish, 2018; Raykov & Marcoulides, 2019; Sijtsma, 2009). Le coefficient alpha « est très connu mais très mal compris par les chercheurs» affirment Deng et Chan (2017). Premièrement, il n'est pas le seul coefficient de fiabilité disponible, loin de là (e.g, le coefficient oméga; voir Revelle & Condon, 2019). Deuxièmement, il n'est pas exempt de limites dont sa sensibilité au nombre d'items dans le test et surtout il n'est pas approprié aux mesures multidimensionnelles car il requiert l'hypothèse de la tau-équivalence entre items (i.e., égalité de la contribution de chaque item au score total au test qui se traduit par des saturations factorielles des items quasiment identiques)[9]. Troisièmement, bien qu'il soit considéré, malgré le débat concernant cette terminologie, comme un indicateur de la consistance/cohérence interne d'un test, le coefficient alpha ne peut pour autant être utilisé comme un indicateur de l'unidimensionnalité. La consistance interne a trait aux intercorrélations d'un ensemble d'items alors que l'unidimensionnalité a trait au degré auquel les items dépendent d'un seul facteur latent (i.e., construit). C'est la raison pour laquelle il faut d'abord confirmer l'unidimensionnalité du test avant d'estimer son coefficient alpha, et non pas l'inverse (Green & Yang, 2015). En présence

[7] Il convient de préciser deux points ici : (a) le test-retest est pertinent lorsque le test mesure un aspect stable (e.g., trait), (b) le coefficient de corrélation intraclasse pourrait être utilisé pour estimer la fiabilité test-retest.
[8] L'article de Cronbach (1951) a été cité 42906 fois (soit en moyenne 631 citations par an, et une cinquantaine de citations par mois); «cronbach's alpha» génère 433000 résultats dans Google Scholar (consultation réalisée le 5 janvier 2020).
[9] Un modèle congénérique est moins strict que le modèle tau-équivalent dans le sens où il ne fait pas l'hypothèse d'égalité de variance et covariance entre les items ni l'hypothèse d'égalité de leurs saturations factorielles (i.e., le modèle le plus communément utilisé). Un tel assouplissement viole l'hypothèse qui fonde le coefficient alpha aboutissant à une sous-estimation de la valeur alpha (c'est-à-dire une valeur plus faible).

de tests multidimensionnels, le coefficient oméga est plus approprié (Cho, 2016; Dunn, Baguley, & Brunsden, 2013). Quatrièmement, la valeur seuil indiquant une fiabilité acceptable fait toujours débat. La fameuse valeur seuil d'acceptabilité .70 est erronément attribuée à Nunnally (1978) qui, au contraire, préconise une valeur seuil de .80 pour les mesures utilisées en recherche scientifique (Lance, Butts, & Michels, 2006). Cette valeur seuil d'acceptabilité grimpe à .90 pour les tests utilisés en pratiques institutionnelles et professionnelles (Nunnally, 1978). Nous pensons que le niveau d'exigence à l'égard des mesures psychologiques devrait être le même quel que soit l'usage auquel elles sont destinées. Toutefois, un coefficient alpha dépassant .95 pourrait être problématique car symptomatique de la présence d'items redondants dans un test. Or, on le sait, introduire des items redondants dans une mesure fait augmenter sa fiabilité mais engendre par là même le risque de sacrifier sa validité de contenu. Nunnally et Bernstein (1994) conseillent de «ne jamais passer à une mesure moins valide simplement parce qu'elle est plus fiable» (p. 265).

2) L'étape précédente (Étape 6) a échoué à confirmer la structure interne (i.e., le modèle de mesure), nécessitant ainsi une révision voire une modification de la formulation de certains items. L'étape 7 doit alors impérativement reposer sur un nouvel échantillon différent de celui ayant répondu à la première version du test (Anderson, & Gerbing, 1991). Les données recueillies auprès du deuxième échantillon (si possible avec N > 300) permettra de re-tester la structure interne mais aussi d'analyser la fiabilité des scores. L'analyse de fiabilité des scores doit obligatoirement succéder à l'analyse de la structure interne car celle-ci détermine la manière dont les scores sont utilisés et leur fiabilité évaluée (e.g., structure hiérarchique, structure bifactorielle, voir Gana & Broc, 2018). Doit-on seulement utiliser le score total au test, ou bien peut-on aussi, s'il y a lieu, utiliser les sous-scores aux différentes dimensions/facettes du test ? Ainsi, faut-il analyser la fiabilité aussi bien du score total que

celle des sous-scores aux différentes dimensions du test (e.g. structure interne multidimensionnelle)?

3) Dans le cas où il s'avère impossible de confirmer la structure interne, l'élimination des items problématiques pourrait se poser, et l'idée de proposer une version abrégée du test pourrait s'imposer. Il convient dans ce cas précis d'en discuter, si possible, avec l'auteur du test original avant de prendre une décision définitive car la validité de contenu pourrait en pâtir. Une telle décision n'est pas anodine. Abréger un test adapté pourrait parfois améliorer ses qualités psychométriques (Clark & Watson, 2019). Toutefois, une telle démarche pourrait lui être préjudiciable aussi. En effet, Smith, McCarthy et Anderson (2000) fournissent un inventaire assez bien argumenté des nombreux défis liés à l'élaboration et à la validation d'une version abrégée d'un test.

ÉTAPE 8. Apporter/fournir des preuves de validité basées sur les relations avec d'autres construits du réseau nomologique

Pour qu'un score à un test soit valide, il doit également être fiable. Cependant, l'inverse n'est pas vrai. Un score fiable n'est pas forcément valide. Après les preuves étayant la fiabilité des scores il faut s'atteler à en apporter d'autres plaidant en faveur de leur validité. Établir la validité d'un test nécessite la mise en place d'un processus de validation permettant d'accumuler progressivement les indices ("preuves") de validité des scores obtenus par l'entremise de ce test (*Standards*, 2014, voir Boudouda et ses collaborateurs dans ce numéro spécial).

Parmi ces preuves, il y a celles basées sur les relations entre les scores à notre test et les scores à des construits qui lui sont proches et reliés. La validité prédictive (les scores obtenus à notre test permettent-ils de prédire dans les faits, comme attendu, des comportements/attributs futurs?), la méthode des groupes connus (un groupe présentant

certaines caractéristiques que l'on souhaite évaluer à l'aide du test obtient-il des scores significativement plus élevés qu'un groupe de la population cible?), la validité convergente (des scores à des tests différents mais proches corrèlent-ils entre eux ?) ainsi que la validité divergente (i.e. absence de corrélation entre construits différents) permettent d'apporter des indices allant dans ce sens.

Il s'agit d'un processus permanent permettant d'accumuler un corpus d'évidences plaidant en faveur du test, de ses scores, et de leur utilisation. Selon les *Standards* "la validité indique dans quelle mesure des indices/preuves probants et des éléments théoriques viennent étayer l'interprétation des résultats des tests pour l'usage auquel ces tests sont destinés" (p. 11, traduction libre).

Cette étape nécessite la mise en place d'un protocole de validation où le choix des mesures des construits proches ainsi que leurs propriétés psychométriques est crucial. Une bonne connaissance du réseau nomologique du construit mesuré par le test est ici indispensable. Il peut être judicieux de ne pas se contenter de répliquer l'étude princeps de validation. Un échantillon encore plus large que celui de l'étape précédente est requis ici (N> 500).

ÉTAPE 9. Contre-valider/Répliquer

Il s'agit de réaliser une nouvelle étude avec un nouvel échantillon assez large (N >500) visant à répliquer l'étude réalisée à l'étape précédente. La contre-validation suppose: (a) d'administrer le même protocole à un nouvel échantillon "représentatif" de la population cible; (b) d'obtenir les mêmes résultats (structure interne, fiabilité et validité des scores…) que ceux obtenus précédemment. Le pire serait de se retrouver avec autant de mesures que d'échantillons issus de la même population (e.g., instabilité de la structure factorielle d'un échantillon à l'autre). D'ailleurs, l'adaptation d'un test qui n'a pas été contre-validé auprès de

la population pour laquelle il a été développé est une entreprise téméraire fortement risquée. Une approche multi-traits-multi-méthodes (Schmitt & Stults, 1986) pourrait, si possible, constituer une option à envisager dans le cadre d'une étude de réplication. En effet, une analyse factorielle appliquée à une matrice de corrélations entre plusieurs traits (multi-traits) mesurés par l'entremise de méthodes différentes (multi-méthodes) offre l'avantage d'évaluer simultanément la validité convergente, la validité divergente ainsi que les effets de méthode (voir par exemple K'Delant & Gana, 2009).

ÉTAPE 10. Établir des Normes ou des scores cliniquement significatifs

Si l'étude de contre-validation arrive à tenir toutes ses promesses, ses données peuvent servir à établir un étalonnage, voire, si envisagé dans le protocole, définir des scores seuils/scores cliniquement significatifs permettant de "diagnostiquer", "dépister" le phénomène étudié. Toutefois, il convient de rappeler que l'établissement des normes exige que le test soit administré auprès d'un large échantillon représentatif de la population cible. Les normes serviront à cette population et à elle seule. Il existe plusieurs échelles destinées à exprimer les normes d'un test. Par exemple, la courbe *Receiver Operating Characteristic* (ROC) est souvent utilisée pour déterminer les scores seuils cliniquement significatifs d'un test[10].

Conclusion

L'adaptation transculturelle de tests ne constitue pas, à notre avis, un domaine unifié de la psychométrie. Plusieurs modèles et pratiques ont essaimé ici et là depuis quelques décennies. L'objectif de cet article est de fournir un canevas synthétisant, pour mieux la

[10] En cas d'utilisation d'un échantillon clinique dûment diagnostiqué (DSM, CIM) pour déterminer les scores cliniquement significatifs avec une courbe ROC, il convient de prêter attention à la taille de cet échantillon. Certains logiciels tels que pROC dans l'environnement R peuvent aider à déterminer la taille appropriée d'un tel échantillon. Le lecteur trouvera dans Youngstrom (2014) une introduction facile à la manière de réaliser une courbe ROC.

guider, la procédure d'adaptation de tests et échelles de mesure psychologiques. Ce canevas, sorte de cadre général, se fonde, rappelons-le, sur deux cadres méthodo-épistémologiques de la psychométrie actuelle : la seconde édition des *Recommandations* de l'ITC pour la traduction et l'adaptation de tests, et la dernière édition des *Standards* de pratique du testing en éducation et en psychologie de l'APA (voir Boudouda et ses collaborateurs dans ce numéro spécial).

Il est loisible de croire que ce canevas, eu égard à l'évolution des exigences concernant l'adaptation de tests, pourrait être utile à plusieurs acteurs impliqués dans l'adaptation de tests et leurs diffusions (e.g., auteurs, éditeurs, experts). Pour ce faire, nous avons élaboré une liste de vérification (checklist) permettant de faciliter les tâches des uns et des autres. Soyons clairs ici: loin de nous l'idée d'imposer, et encore moins corseter une activité scientifique qui dépend, *in fine*, des auteurs qui la réalisent. Car à trop normaliser on court le risque de tuer toute créativité. Ainsi, ce guide se doit de rester souple et ouvert.

Insérer la liste de vérification ici

Aussi, il est loisible de croire que ce canevas pourrait attirer l'attention sur certaines pratiques problématiques. Est révolu, osons-nous l'espérer, le temps où une seule étude avec un échantillon de convenance (des étudiants généralement) suffise à valider l'adaptation d'une échelle de mesure psychologique, et à en publier les résultats. Nous l'avions écrit à maintes reprises et nous sommes prêts à le répéter à satiété, s'il le faut : les résultats d'une recherche scientifique en psychologie -quantitative et corrélationnelle- valent ce que valent les méthodes et les instruments de mesure utilisés pour les obtenir. Méthodes bancales, résultats douteux; instruments de mesure insatisfaisants, résultats hasardeux[11]. Inéluctablement. *In fine*, c'est la qualité et la crédibilité des connaissances publiées par les revues dont il est question ici, mais

[11] On reconnaitra ici le principe «garbage in, garbage out» (GIGO) ou «quality in, quality out» voulant dire ceci : «si ce qui rentre est de qualité, ce qui sortira le sera aussi».

également de leur applicabilité. Investir dans l'élaboration et l'adaptation de tests c'est investir dans la construction du savoir psychologique. Examinant les artéfacts et les effets de méthodes causés par la formulation négative des items de l'échelle d'estime de soi de Rosenberg (RSE ; 1965), Marsh, Scalas et Nagengast (2010) de conclure que « la seule preuve claire est que le modèle utilisé implicitement comme base de presque toutes les recherches appliquées était clairement inapproprié, remettant en cause la vaste littérature basée sur l'échelle RSE » (p. 379). En analysant les effets de méthodes entachant la version française de l'échelle de Rosenberg, Gana et ses collaborateurs (2013) confirment les réserves de Marsh et ses collaborateurs (2010) quant à la qualité d'un pan entier de la littérature portant sur la psychologie de l'estime de soi générée par l'entremise de cette échelle. Et Gana et ses collaborateurs (2013) de conclure que « comme l'a noté à juste titre Marsh et al. (2010), nos solides résultats remettent légitimement en cause la vaste littérature qui avait utilisé le modèle de mesure de l'échelle RSE» (p. 141).

Dans le cadre de leur recherche doctorale, les doctorants peuvent être freinés voire bridés dans la mise en place de protocoles de recherche du fait de l'absence d'instruments de mesure en français évaluant les construits étudiés. Si le bricolage à la sauvette (pour reprendre l'expression de Vallerand, 1989) de versions françaises d'instruments généralement américains tend, fort heureusement, à diminuer, il n'en reste pas moins vrai que l'on est encore loin du niveau minimal d'exigence quant à la qualité des adaptations. Le présent canevas pourrait servir de cadre général structurant, permettant de valoriser l'adaptation d'un instrument de mesure réalisée dans un travail doctoral. En effet, il nous paraît regrettable qu'une adaptation d'un instrument de mesure réalisée dans le cadre d'un travail doctoral ne puisse donner lieu à une mise à disposition des chercheurs et praticiens via une publication. Il est loisible de croire que ce canevas pourrait y contribuer.

Une procédure d'adaptation convoque plusieurs expertises dans différents domaines notamment psychométriques et statistiques (e.g., modélisation par équations structurales, modèles de réponse à l'item) (Rios & Hambleton, 2016). Elle nécessite d'être informé non seulement des récentes mises à jours des *Standards* de l'APA et des *Recommandations* de l'ITC mais aussi des avancées scientifiques fréquentes en psychométrie. Elle est chronophage et couteuse car elle nécessite la mise en place d'au moins trois à quatre études empiriques dont une pour la contre-validation. Celle-ci sert à répliquer, pour les confirmer, les qualités métrologiques du test adapté. Elle est plus que nécessaire, incontournable même. Il est regrettable de noter que la psychologie détient le triste titre de la discipline scientifique la moins «répliquante» des résultats de ses recherches (Yong, 2012). La psychologie vit donc une crise de la reproductibilité qui est à l'origine d'une crise de confiance à l'égard du savoir psychologique (Martin & Clarke, 2017). Il est de notre devoir de la dissiper (Chopik, Bremner, Defever, & Keller, 2018).

Pour finir, il convient de suggérer que dans le titre d'un papier rendant compte d'un travail d'adaptation de test puisse figurer le mot «adaptation française» plutôt que «traduction française », et «preuves de validité des scores » et non «validation d'une version française»[12]. Car il n'y a pas à proprement parler de tests valides mais de preuves de validité de leurs scores et des interprétations qu'on en tire.

[12] Par exemple : «Preuves de validité des scores de l'adaptation française de l'échelle XXX » ou «Preuves des qualités psychométriques de l'adaptation française de l'échelle XXX ».

Références

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

Anderson, J. C., & Gerbing, D. W. (1991). Predicting the performance of measures in a confirmatory factor analysis with a pretest assessment of their substantive validities. *Journal of Applied Psychology*, 76, 732–740. doi:10.1037/0021-9010.76.5.732.

Armengaud, F. (2007). *La pragmatique*. Paris: Presses Universitaires de France.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, 16, 397–438. doi:10.1080/10705510903008204

Bauman, L. J., & Adair, E. G. (1992). The use of ethnographic interviewing to inform questionnaire construction. *Health Education Quarterly, 19*, 9–23. doi:10.1177/109019819201900102

Beatty, P. C., & Willis, G. B. (2007). Research Synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly, 71*, 287– 311.

Béland, S., & Cousineau, D. (2018). Adieu coefficient alpha de Cronbach! J'ai trouvé plus fidèle que toi. *Revue de Psychoéducation,* 47, 449–460. doi:10.7202/1054068ar.

Bourgoz Froidevaux, A. (2017). Les écrits de Jean Cardinet. Citations commentées.

Evaluer. Journal international de Recherche en Education et Formation, 3, 97-104.

Bracken, B. A., & Barona, A. (1991). State of the art procedures for translating, validating

and using psychoeducational tests in cross-cultural assessment. *School Psychology*

International, 12, 119–132. doi: 10.1177/0143034391121010.

Byrne, B. M. (2016). Adaptation of assessment scales in cross-national research: Issues, guidelines, and caveats. *International Perspectives in Psychology: Research, Practice, Consultation*, 5, 51–65. doi:10.1037/ipp0000042.

Cardinet, J. (1995). A prehistory of the International Test Commission. *European Journal of Psychological Assessment*, 11(2), 128–132. doi.org/10.1027/1015-5759.11.2.128.

Cattell, R. B. (1940). A culture-free intelligence test I. *Journal of Educational Psychology*, 31, 161–179. doi:10.1037/h0059043.

Cho, E. (2016). Making Reliability Reliable: A Systematic approach to reliability coefficients. *Organizational Research Methods*, 19, 651–682. doi:10.1177/1094428116656239

Chopik, W. J., Bremner, R. H., Defever, A. M., & Keller, V. N. (2018). How (and whether) to teach undergraduates about the replication crisis in psychological science. *Teaching of Psychology*, 45, 158–163. doi:10.1177/0098628318762900.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, 68(3), 397–412. doi:10.1177/0013164407310130.

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, 31, 1412–1427. doi:10.1037/pas0000626.

Commission Suisse des Tests. (1971). Règlement de la commission suisse des tests visant à promouvoir la qualité des tests psychologiques et à prévenir leur utilisation abusive. *Revue Suisse de Psychologie,* 30, 340–349.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334. doi:10.1007/BF02310555.

Daouk-Öyry, L., & McDowal, A. (2013). Using cognitive interviewing for the semantic enhancement of multilingual versions of personality questionnaires. *Journal of Personality Assessment*, 95, 407–416. doi:10.1080/00223891.2012.735300.

Deng, L., & Chan, W. (2017). Testing the difference between reliability coefficients alpha and omega. *Educational and Psychological Measurement*, 77, 185–203. doi:10.1177/0013164416658325

Dunn, T. J., Baguley, T., & Brunsden, V. (2013). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, 105, 399-412.doi: 10.1111/bjop.12046

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4, 272-299.

Frijda, N., & Jahoda, G. (1966). On the scope and methods of cross-cultural research. *International Journal of Psychology*, 1, 109–127. doi:10.1080/00207596608247118.

Gana, K. & Broc, G. (2018). *Introduction à la modélisation par équations structurales. Manuel Pratique avec lavaan.* Londres: ISTE editions.

Gana, K., Saada, Y., Bailly, N., Joulain, M., Hervé, C., & Alaphilippe, D. (2013). Longitudinal factorial invariance of the Rosenberg Self-Esteem Scale: Determining the nature of method effects due to item wording. *Journal of Research in Personality*, 47, 406–416. doi:10.1016/j.jrp.2013.03.011.

García, A. A. (2011). Cognitive interviews to test and refine questionnaires. Public Health Nursing, 28(5), 444–450. doi: 10.1111/j.1525-1446.2010.00938.x

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment*, 6, 304–312. doi:10.1037/1040-3590.6.4.304.

Gerbing, D., & Anderson, J. (1988). An Updated Paradigm for Scale Development Incorporating Unidimensionality and Its Assessment. *Journal of Marketing Research*, 25, 186-192. doi:10.2307/3172650

Green, S. B., & Yang, Y. (2015). Evaluation of dimensionality in the assessment of internal consistency reliability: coefficient alpha and omega coefficients. *Educational Measurement: Issues and Practice*, 34, 14–20. doi:10.1111/emip.12100

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103*, 265–275.

Gudmundsson, E. (2009). Guidelines for translating and adapting psychological instru-ments. *Nordic Psychology*, 61, 29–45. doi: 10.1027/1901-2276.61.2.29.

Harkness, J. A., Villar, A., & Edwards, B. (2010). Translation, adaptation and design. In J. A. Harkness., M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B.-E. Pennell, & T. W. Smith (Eds.), *Survey methods in multinational, multicultural and multiregional contexts* (pp.117-139). Hoboken, NJ: John Wiley & Sons.

International Test Commission (2001). International Guidelines for Test Use. *International Journal of Testing,* 1, 93-114.

International Test Commission (2005). *International Guidelines on Test Adaptation*. [www.intestcom.org]

International Test Commission (2005). *International Guidelines on Computer-Based and Internet Delivered Testing*. [www.intestcom.org]

International Test Commission (2012). *International Guidelines on Quality Control in Scoring, Test Analysis, and Reporting of Test Scores.* [www.intestcom.org]

International Test Commission (2014). *International Guidelines on the Security of Tests, Examinations, and Other Assessments.* [www.intestcom.org]

International Test Commission (2015). *International Guidelines for Practitioner Use of Test Revisions, Obsolete Tests, and Test Disposal*. [www.InTestCom.org]

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

International Test Commission. (2018). *ITC Guidelines for the Large-Scale Assessment of Linguistically and Culturally Diverse Populations*. [www.InTestCom.org]

K'Delant, P. & Gana, K. (2009). Analyse Multitraits-Multiméthodes des scores au Questionnaire d'Attributs Personnels auprès d'un échantillon féminin. *Psychologie Française*, 54, 323-336.

Krach, S. K., McCreery, M. P., & Guerard, J. (2017). Cultural-linguistic test adaptations: Guidelines for selection, alteration, use, and review. *School Psychology International*, 38, 3–21. doi:10.1177/0143034316684672

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The Sources of Four Commonly Reported Cutoff Criteria: What Did They Really Say? *Organizational Research Methods*, 9, 202–220. doi:10.1177/1094428105284919.

Laveault, D. (2012). Soixante ans de bons et mauvais usages du alpha de Cronbach. *Mesure et évaluation en éducation 35*, 1–7. doi: 10.7202/1024716ar

Ljungberg, A. K., Fossum, B., Fürst, C. J., & Hagelin, C. L. (2015). Translation and cultural adaptation of research instruments—Guidelines and challenges: An example in FAMCARE-2 for use in Sweden. *Informatics for Health & Social Care*, 40, 67–78. doi:10.3109/17538157.2013.872111.

Lyons-Thomas, J., Liu, Y., & Zumbo, B. D. (2014). Validation practices in the social, behavioral, and health sciences: A synthesis of syntheses. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences*. (Vol. 54, pp. 313–319). Cham: Springer International Publishing. doi:10.1007/978-3-319-07794-9_18.

Marsh, H. W., Scalas, L. F., & Nagengast, B. (2010). Longitudinal tests of competing factor structures for the Rosenberg Self-Esteem Scale: Traits, ephemeral artifacts, and stable response styles. *Psychological Assessment*, 22, 366–381. doi: 10.1037/a0019225.supp (Supplemental)

Martin, G. N., & Clarke, R. M. (2017). Are psychology journals anti-replication? A snapshot of editorial practices. *Frontiers in Psychology*, 8.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, 23, 412–433. doi:10.1037/met0000144.

McNeish, D., An, J., & Hancock, G. R. (2018). The thorny relation between measurement quality and fit index cutoffs in latent variable models. *Journal of Personality Assessment, 100*, 43–52. https://doi.org/10.1080/00223891.2017.1281286

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Oakland, T., Poortinga, YH., Schlegel, J., & Hambleton, RK. (2001). International

Test Commission: Its History, Current Status, and Future Directions. *International Journal of Testing, 1*, 3 — 32. DOI: 10.1207/S15327574IJT0101_2.

To link to this Article: DOI: 10.1207/S15327574IJT0101_2

Pichot, P. (1949). *Les tests mentaux en psychiatrie (T1). Instruments et méthodes*.

Paris: PUF.

Plake, B. S., & Wise, L. L. (2014). What is the role and importance of the revised AERA, APA, NCME Standards for Educational and Psychological Testing? *Educational Measurement: Issues and Practice*, 33(4), 4–12. doi:10.1111/emip.12045.

Raykov, T. (2007). Reliability if deleted, not 'alpha if deleted': Evaluation of scale reliability following component deletion. *British Journal of Mathematical and Statistical Psychology*, 60, 201–216.doi: 10.1348/000711006X115954

Raykov, T. (2008). Alpha if item deleted: a note on loss of criterion validity in scale development if maximizing coefficient alpha. *British Journal of Mathematical and Statistical Psychology,* 61, 275–285. doi: 10.1348/000711007X188520

Raykov, T., & Marcoulides, G. A. (2019). Thanks coefficient alpha, we still need you! *Educational and Psychological Measurement*, 79, 200–210. doi:10.1177/0013164417725127.

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment,* 31, 1395–1411. doi:10.1037/pas0000754.supp (Supplemental).

Rios, J. A., & Hambleton, R. K. (2016). Statistical methods for validating test adaptations used in cross-cultural research. In N. Zane, G. Bernal, & F. T. L. Leong (Eds.), *Evidence-based psychological practice with ethnic minorities: Culturally informed research and clinical strategies.* (pp. 103–124). Washington, DC: American Psychological Association. doi:10.1037/14940-006.

Rios, J. A., & Sireci, S. G. (2014). Guidelines versus practices in cross-lingual assessment: A disconcerting disconnect. *International Journal of Testing*, 14, 289–312. doi:10.1080/15305058.2014.924006.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Ryan, K. E., Gandha, T., Culbertson, M. J., & Carlson, C. (2014). Focus group evidence: Implications for design and analysis. American Journal of Evaluation, 35(3), 328–345. doi:10.1177/1098214013508300.

Sackett, P. R. (2014). An employment testing and credentialing perspective on the Standards for Educational and Psychological Testing. *Educational Measurement: Issues and Practice*, 33(4), 22–24. doi:10.1111/emip.12049.

Sarrazin, G. (2003). *Normes de pratique du testing en psychologie et en éducation*. Montréal : Institut de recherches psychologiques.

Schmitt, N., & Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement*, 10, 1–22. doi:10.1177/014662168601000101.

Sidani, S., Guruge, S., Miranda, J., Ford-Gilboe, M., & Varcoe, C. (2010). Cultural adaptation and translation of measures: An integrated method. *Research in Nursing & Health*, 33, 133–143.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107–120. doi:10.1007/s11336-008-9101-0.

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment*, 12, 102–111. doi:10.1037/1040-3590.12.1.102.

Sousa, VD., & Rojjanasrirat, W. (2011). Translation, adaptation and validation of instrumentsor scales for use in cross-cultural health care research: a clear and user-friendly guideline. Journal of Evaluation in Clinical Practice, 17, 268–74. doi: 10.1111/j.1365-2753.2010.01434.x.

Thorndike, E.L. (1904). *An introduction to the theory of mental and social measurements*. NY: The Scientific Press

Vallerand, R. J. (1989). Vers une méthodologie de validation trans-culturelle de questionnaires psychologiques: Implications pour la recherche en langue française. *Psychologie Canadienne*, 30, 662–680. doi:10.1037/h0079856.

van de Vijver, P. J. R. (2016). Test adaptations. In F. T. L. Leong, D. Bartram, F. M. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC international handbook of testing and assessment*. (pp. 364–376). New York, NY: Oxford University Press.

Van de Vijver, F. J., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in

cross-cultural assessment. *European Journal of Psychological Assessment*, 13, 29–37. doi: 10.1027/1015-5759.13.1.29.

Vogt, D. S., King, D. W., & King, L. A. (2004). Focus Groups in Psychological Assessment: Enhancing Content Validity by Consulting Members of the Target Population. *Psychological Assessment*, 16, 231–243. doi:10.1037/1040-3590.16.3.231.

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, 34, 806-838. doi:10.1177/0011000006288127

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.

Wood, D., Harms, P. D., Lowman, G. H., & DeSimone, J. A. (2017). Response speed and response consistency as mutually validating indicators of data quality in online samples. *Social Psychological and Personality Science*, 8, 454 – 464. doi.org/10.1177/1948550617703168

Yong, E. (2012). Bad copy. *Nature 485*, 298–300. doi:10.1038/485298a

Youngstrom, E. A. (2014). A primer on receiver operating characteristic analysis and diagnostic efficiency statistics for pediatric psychology: We are ready to ROC. *Journal of Pediatric Psychology*, 39, 204–221. doi:10.1093/jpepsy/jst062

Zumbo, B. D. (2014). What role does, and should, the test Standards play outside of the United States of America? *Educational Measurement: Issues and Practice*, 33, 31–33. doi:10.1111/emip.12052.

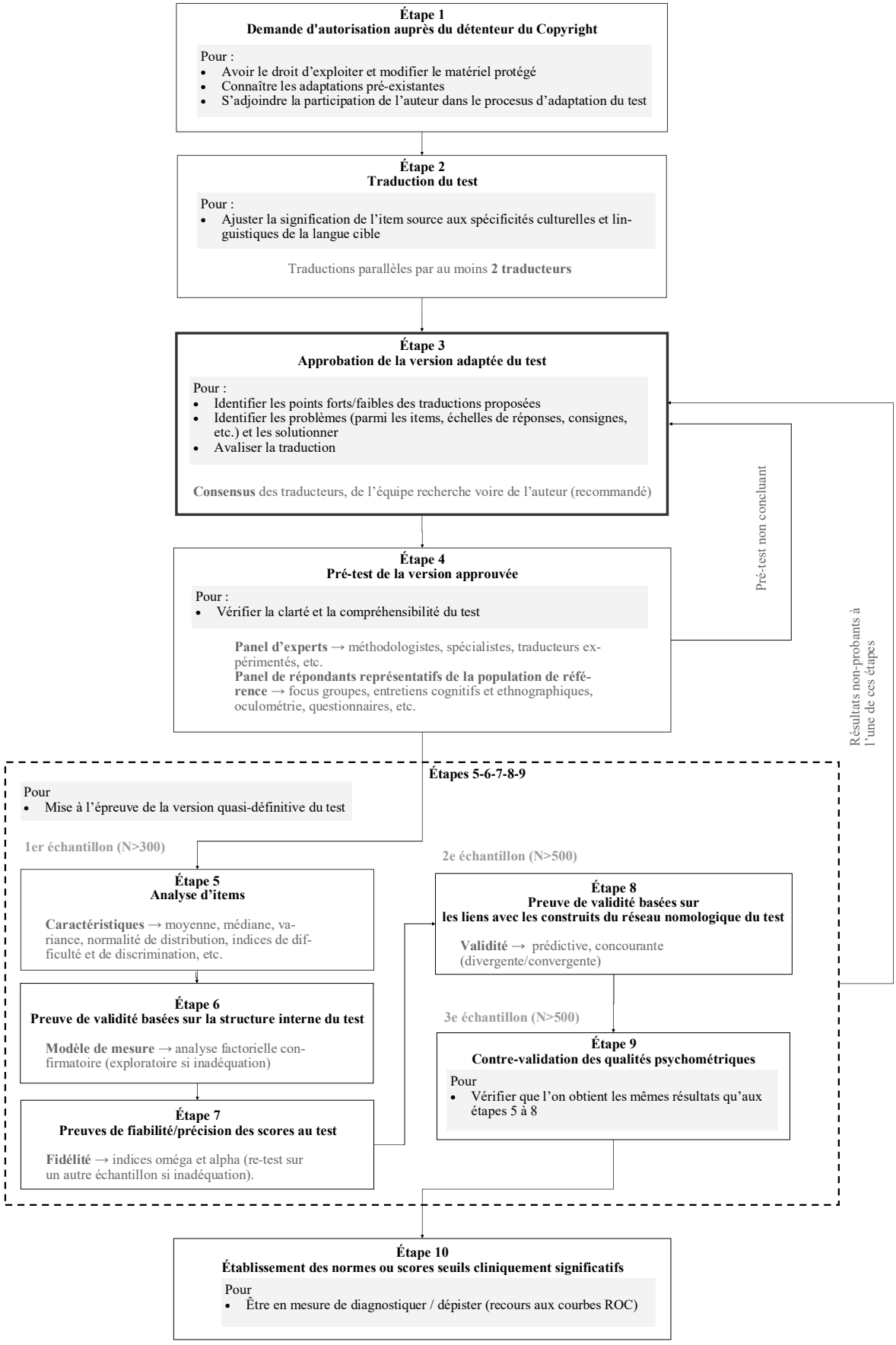

Figure 1. Étapes pour l'adaptation transculturelle de tests/échelles de mesure

Tableau 1. Liste de vérification (checklist) à l'usage des auteurs d'adaptation de tests/échelles de mesure psychologiques

Section	#	Item	Numéro de page
TITRE			
Titre	1	Indiquer dans le titre de l'article « adaptation d'un test/échelle » et non pas « traduction d'un test/échelle »	
RESUME			
Résumé	2	Fournir un résumé structuré incluant, si possible : le rationnel de l'étude ; ses objectifs ; les méthodes de traduction/adaptation, pré-test, validation, contre-validations employées ; une description des échantillons auxquels ont été administré le test/échelle; les preuves de validité psychométrique du test/échelle; les limites de l'étude; une conclusion et les implications tant cliniques que scientifiques des résultats trouvés.	
INTRODUCTION			
Rational	3	Décrire le test/échelle concerné à la lumière du contexte de la littérature scientifique et justifier les enjeux de sa traduction et de son adaptation culturelle et linguistique par rapport aux autres tests existants dans la langue cible. S'il s'agit d'une contre-validation, présenter les précédents travaux de développement, pré-test et de validation initiale du test.	
Objectifs	4	Fournir les objectifs de l'étude et préciser les différentes étapes entreprises (traduction/adaptation et/ou pré-test et/ou validation et/ou contre-validation), ainsi que le nombre d'échantillons nécessaire à la réalisation de l'étude.	
Analyses statistiques	5	Présenter les différentes analyses statistiques prévues : analyses d'item, analyse factorielle, analyse de fiabilité, analyse de validité. Préciser le ou les logiciels utilisés pour réaliser ces analyses. S'inscrire explicitement dans les cadres méthodo-épistémologiques : *Recommandations* de l'ITC et les *Standards* de l'APA	
MÉTHODES			
Obtention des autorisations	6	Décrire comment ont été obtenues les autorisations nécessaires auprès des titulaires des droits d'auteur (copyright) du test/échelle de mesure.	
Traduction /adaptation du test/échelle	7	Décrire de façon exhaustive la méthode de traduction/adaptation vers la langue cible du test/échelle de mesure (unidirectionnelle, à rebours, double traduction), les vérifications nécessaires au bon déroulé du processus (degré de concordance entre la définition et le contenu du construit ; réduction de l'influence des différences culturelles et linguistiques, prise en compte des différences linguistiques, psychologiques et culturelles des populations visées; preuves du sens équivalent de la consigne du test et du contenu des items dans les langues visées) ainsi que l'ensemble des acteurs ayant été impliqués dans ces processus (experts, traducteurs, etc..).	
Approbation de la version adaptée du test	8	Décrire comment a été obtenu le consensus sur la version approuvée de l'adaptation du test/échelle de mesure.	

Pré-test de la version approuvée	9	Fournir les éléments de méthode employée pour vérifier la clarté et la compréhensibilité du test/échelle approuvé, des processus de réponses et de l'équivalence sémantique du test/échelle cible avec la version source.	
Participants	10	Spécifier les caractéristiques de l'échantillon retenu pour effectuer le pré-test/ la validation/ la contre-validation du test adapté dans la langue cible (population générale, enfants, adolescents, etc.) et préciser les critères d'inclusion et d'exclusion dans l'étude, et ce en argumentant vos choix à chacune des étapes.	
Procédure : Conditions de passation du test	11	Préciser dans quelles conditions s'est effectuée l'administration du test/échelle auprès de la population cible (qui a administré, le lieu, les conditions, etc.).	
RÉSULTATS			
Aléas et difficultés rencontrés lors de la procédure de traduction/adaptation	12	Décrivez avec transparence toutes les difficultés, désaccords et résultats non-probants rencontrés dans la phase de traduction/adaptation et/ou de pré-test et/ou de validation et/ou de contre-validation du test adapté, et expliquer quelles ont été les démarches entreprises pour y remédier	
Statistiques descriptives	13	Fournir les effectifs et les indices descriptifs de la population étudiée, ainsi que le taux d'inclusion effectif à la fin de l'étude	
Analyse d'items	14	Décrire les résultats de l'analyse d'items du test/échelle adapté, en précisant les indices statistiques utilisés et leur significativité.	
Analyse de la structure du test/échelle	15	Décrire les résultats de l'analyse factorielle du test adapté, en précisant les indices statistiques utilisés et leur significativité.	
Analyse de la fiabilité des scores	16	Décrire les résultats de la fiabilité du test adapté, en précisant les indices statistiques utilisés et leur significativité.	
Analyses de validité	17	Décrire les résultats des analyses de validité prévues et annoncées dans la section « Analyses statistiques » en précisant les indices statistiques utilisés et leur significativité.	
Normes et scores-seuils	18	Dans le cadre d'une contre-validation, et si applicable, décrire toute procédure d'établissement des normes ou de scores seuils (courbes ROC).	
DISCUSSION			
Résumé des preuves	19	Résumer les preuves d'une traduction/adaptation suffisante du test dans la langue et culture cible, ainsi que les validités structurales et psychométriques de la version adaptée.	
Limites	20	Discuter des limites inhérentes au processus de traduction/adaptation et de validation psychométrique employé. Déclarer impérativement chaque impair et résultat non-concluants	
Conclusions	21	Fournir une interprétation générale des preuves apportées, et les recontextualiser auprès des autres études sur l'outil adapté. Discuter des implications tant cliniques que scientifiques des résultats de l'adaptation du test/échelle, et proposer des pistes de recherches ultérieures.	
REMERCIEMENTS			
Remercier les collaborations externes	22	Remercier le/les auteur(s) de la version originale du test; les traducteurs extérieurs et toute personne en dehors des co-auteurs ayant contribué à l'adaptation du test/échelle de mesure.	

FINANCEMENT			
Financement et conflits d'intérêt	23	Décrire les sources de financement pour l'adaptation et la validation du test employé, ainsi que toute autre forme de soutien matériel ou logistique.	
ACCÈS AU TEST/ÉCHELLE			
Comment accéder au test/échelle et son matériel	24	Préciser si le test/échelle est en accès libre aux chercheurs et praticiens et sous quelles modalités. Exposer dans le cas contraire les raisons de préserver la confidentialité du test/échelle de mesure.	