

Sources d'invalidité et d'erreur dans la traduction ou l'adaptation de tests : un état de la question

Julie Grondin, Éric Dionne, Carole Fleuret et Nancy Boiteau

Volume 46, numéro 1-2, 2015

URI : <https://id.erudit.org/iderudit/1039041ar>

DOI : <https://doi.org/10.7202/1039041ar>

[Aller au sommaire du numéro](#)

Éditeur(s)

Revue de l'Université de Moncton

ISSN

1712-2139 (numérique)

[Découvrir la revue](#)

Citer cet article

Grondin, J., Dionne, É., Fleuret, C. & Boiteau, N. (2015). Sources d'invalidité et d'erreur dans la traduction ou l'adaptation de tests : un état de la question. *Revue de l'Université de Moncton*, 46(1-2), 291–323.
<https://doi.org/10.7202/1039041ar>

Résumé de l'article

La traduction ou l'adaptation d'épreuves pour tenir compte de la diversité culturelle et linguistique est de plus en plus fréquente. Différents facteurs expliquent cet engouement comme les impératifs économiques ou encore la complexité de bâtir une épreuve originale. Il est nécessaire de s'assurer que ces opérations de traduction ou d'adaptation se réalisent selon une méthode rigoureuse qui tient compte des principales menaces à la validité. En effet, les concepteurs d'épreuves doivent s'assurer de réduire le plus possible les biais qui pourraient affecter les différents sous-groupes ciblés par l'épreuve à administrer. Il existe peu de ressources francophones visant à présenter les menaces à la validité dans un tel contexte. L'objectif de ce texte est d'apporter une contribution afin de faire le point sur les pièges à éviter dans un contexte de traduction ou d'adaptation.

SYNTHÈSE DE RECHERCHES

SOURCES D'INVALIDITÉ ET D'ERREUR DANS LA TRADUCTION OU L'ADAPTATION DE TESTS : UN ÉTAT DE LA QUESTION

Julie Grondin

Université du Québec à Rimouski, Campus de Lévis,

Éric Dionne

et

Carole Fleuret

Université d'Ottawa

Nancy Boiteau

Université du Québec à Rimouski, Campus de Lévis

Résumé

La traduction ou l'adaptation d'épreuves pour tenir compte de la diversité culturelle et linguistique est de plus en plus fréquente. Différents facteurs expliquent cet engouement comme les impératifs économiques ou encore la complexité de bâtir une épreuve originale. Il est nécessaire de s'assurer que ces opérations de traduction ou d'adaptation se réalisent selon une méthode rigoureuse qui tient compte des principales menaces à la validité. En effet, les concepteurs d'épreuves doivent s'assurer de réduire le plus possible les biais qui pourraient affecter les différents sous-groupes ciblés par l'épreuve à administrer. Il existe peu de ressources francophones visant à présenter les menaces à la validité dans un tel contexte. L'objectif de ce texte est d'apporter une contribution afin de faire le point sur les pièges à éviter dans un contexte de traduction ou d'adaptation.

Mots-clés : validité, traduction, adaptation, test, culture.

Abstract

The translation or the adaptation of tests, which take into account cultural and linguistic diversity, is more and more frequent. Different factors explain this situation, such as economic imperatives or the complexity of creating a new test. It is necessary to make sure that these translation or adaptation procedures take place according to a rigorous method which takes into account main threats to validity. Indeed, the test developers must make sure to reduce as much as possible the biases which could affect various subgroups of the population. This paper presents some pitfalls that can affect test validity when test developers have to translate or adapt items.

Keywords : validity, translation, adaptation, test, culture.

Introduction et problématique

De plus en plus de systèmes éducatifs ont recours à des enquêtes à grande échelle dans le cadre des évaluations du rendement scolaire des élèves. Ces enquêtes permettent de recueillir les informations nécessaires à la planification des politiques éducatives, de même qu'à dresser un portrait de la réussite des élèves dans une perspective de reddition des comptes. Or, bien souvent, il existe une diversité linguistique ou culturelle¹ dans la population évaluée qui astreint les responsables des systèmes éducatifs à offrir une version adaptée de leurs instruments de cueillette de données. En effet, un instrument qui serait seulement administré dans la langue du groupe majoritaire d'un système éducatif ne permettrait pas aux divers groupes linguistiques ou culturels minoritaires, composant aussi ce système d'offrir une pleine performance, particulièrement si leur niveau de connaissance de cette langue est restreint (Sireci et Khaliq, 2002).

La traduction d'épreuves constitue donc une opération fréquente en éducation (Elosua et López-Jauregui, 2008; Tomás, Hernández et González-Romá, 2006). Jusqu'à tout récemment, les évaluations multilinguistiques impliquaient surtout une comparaison entre deux groupes de langues tels que l'anglais et l'espagnol aux États-Unis ou l'anglais et le français au Canada (Sireci, Xing, et Fitzgerald, 1999). Au

cours des 15 dernières années, l'intérêt pour la traduction et l'adaptation² d'épreuves a augmenté de façon marquante (Elosua et López-Jauregui, 2007, 2008; Hambleton, 2006). Plusieurs éléments laissent croire que le nombre d'épreuves traduites et qui seront utilisées dans une variété de langues différentes subiront une forte augmentation (Sireci *et al.*, 1999) : 1) les échanges d'épreuves entre différents pays ou différentes juridictions sont de plus en plus communs, 2) les épreuves certificatives sont adaptées dans une multitude de langues, 3) l'intérêt pour la recherche comparative entre groupes linguistiques et culturels différents (Hambleton et Patsula, 2000). Mentionnons aussi que les coûts importants associés à la rédaction et à la validation de nouveaux items³ ajoutent de l'intérêt à traduire des épreuves plutôt que d'en créer de nouvelles. De nos jours, les évaluations réalisées en milieu pluriethnique portent sur des comparaisons entre des individus ayant une variété de répertoires langagiers et culturels. À titre d'illustration, les épreuves à grande échelle pour l'évaluation internationale du rendement scolaire des élèves telles que produites par le *Program for International Student Assessment* (PISA) ou le *Trends in International Mathematics and Science Study* (TIMSS) sont administrées dans plus de 30 langues (Hambleton, 2006).

Toutefois, la demande croissante pour des épreuves adaptées à la réalité des milieux scolaires offre deux défis importants aux éducatriciens et aux évaluateurs; il leur faut produire différentes versions linguistiques et valider ces nouvelles versions par rapport aux buts de l'évaluation (Sireci *et al.*, 1999). Or, il semble que, bien souvent, ces validations ne soient pas suffisamment considérées (Hambleton et Patsula, 2000; Simões Forte et Dionne, 2013). En effet, différents biais de mesure, liés à des différences linguistiques ou culturelles ou encore associés aux formats choisis pour l'instrument de mesure, continuent d'être mis à jour dans diverses évaluations effectuées à l'échelle internationale, et ce, autant en éducation qu'en psychologie (Elosua, Hambleton, et Zenisky, 2006).

Il est vrai que concevoir un instrument de mesure pour un groupe ethnolinguistique⁴ différent du groupe pour lequel l'instrument a été originalement conçu constitue une entreprise majeure qui implique plusieurs étapes et la prise en compte d'éléments importants (Auchter et Stansfield, 1997; Rogers, Gierl, Tardif, et Lin, 2003; Solano-Flores, 2012; van Widenfelt, Treffers, de Beurs, Siedelink, et Koudijs, 2005). Cela

pourrait même impliquer d'apporter quelques modifications à la version originale de l'instrument. En effet, l'hypothèse selon laquelle les versions adaptées d'un instrument sont automatiquement équivalentes à la version originale de l'épreuve est erronée, selon Sireci et Bastari (1998), car adapter un instrument a presque toujours pour conséquence de modifier, avec plus ou moins d'intensité, les propriétés métriques des items; on note d'ailleurs quasi systématiquement une perte d'équivalence, et ce, même si l'adaptation est de bonne qualité (Grisay et Monseur, 2007). Ainsi le but premier de la présente synthèse est-il de discuter de la complexité des adaptations relatives aux différentes épreuves d'évaluation.

Quant à notre deuxième objectif, il découle du fait que la littérature anglo-saxonne offre de nombreuses propositions pour assurer le contrôle de la qualité quand il s'agit de produire des épreuves comparables dans des contextes linguistiques ou culturels variés. Parmi les guides qui reviennent le plus souvent, citons par exemple celui de l'*European Federation of Psychologists' Associations* (EFPA) et de l'*European Association of Work and Organizational Psychologists* (EAWOP) (2005), celui de l'*International Test Commission* (ITC) (2010), ou le *Principles for Fair Student Assessment Practices for Education in Canada* (1993). Par contre, en ce qui concerne la littérature francophone, notre recherche nous a seulement permis de trouver quatre références. Trois d'entre elles sont des adaptations francophones de guides anglophones existants : la première est la traduction effectuée par la Société Française de Psychologie (SFP, 2000) du guide de l'ITC portant sur les recommandations internationales sur l'utilisation des tests; la seconde, plus récemment (SFP, 2007), est également une traduction effectuée par la SFP du guide de l'ITC, mais qui porte cette fois sur les recommandations internationales sur les tests informatisés ou distribués par Internet. La troisième est un guide élaboré par un comité consultatif mixte à Edmonton (Alberta) qui porte sur les principes d'équité relatifs aux pratiques d'évaluation des apprentissages scolaires au Canada (1993). Enfin, nous avons également trouvé le texte de Leplège (2001) qui porte sur l'adaptation transculturelle⁵ d'instruments existants liés à la mesure de la qualité de la vie. Il ne s'agit donc pas tant de recommandations en soi, mais on y trouve tout de même un survol de quelques méthodes de développement « transculturel ». Il nous apparaissait donc pertinent d'offrir aux chercheurs francophones un texte en français leur permettant

de déterminer les contrôles de qualité à réaliser pour pouvoir produire des versions de tests comparables dans des contextes ethnolinguistiques hétérogènes. Ce texte à caractère théorique vise donc également à présenter, en français, sous un angle pédagogique, une revue de la littérature sur les aspects à prendre en considération afin de maximiser la validité de construit lors de la traduction ou de l'adaptation d'épreuves.

Pour ce faire, une recension des écrits a été effectuée sur la base de données ERIC, ainsi qu'à l'aide de recherches sur le moteur Google. Les mots-clés *translation*, *adaptation*, *cross-cultural* jumelés aux mots *test*, *testing* ou *assessment* ont été utilisés. Les recherches ont été effectuées en français et en anglais. Aucune restriction sur les années n'a été appliquée. Les titres et résumés des textes trouvés étaient lus afin de juger de leur pertinence par rapport aux objectifs de l'article. Une fois les articles retenus, les problèmes rencontrés lors de la traduction ou l'adaptation, les méthodes, les biais et les exemples présentés étaient notés et classés suivant leurs ressemblances. Le classement ainsi effectué a été vérifié par un pair afin d'assurer une bonne validité dans le portrait que nous souhaitons dresser relativement à la complexité inhérente à tout processus d'adaptation d'épreuves. Ainsi, dans le présent texte, nous proposons un état de la question en divisant le document de la façon suivante : la question de l'équivalence entre deux versions d'une même épreuve est d'abord abordée, suivie des enjeux associés à la comparabilité des groupes culturels. La comparabilité des méthodes d'évaluation est ensuite traitée en discutant particulièrement des propriétés des épreuves, de la sélection et de la formation des traducteurs et des processus de traduction. Enfin, des modèles d'évaluation de la traduction basés sur l'approche qualitative sont présentés.

Cadre théorique

Comme nous l'avons mentionné, les versions adaptées d'un instrument ne sont pas automatiquement équivalentes même si ses différentes versions visent à se rapprocher, dans leur traduction, d'un contexte sémantique comparable, pour lesquelles les catégories de réponses offertes sont équivalentes ou répondent à certaines normes d'adaptation (van Widenfelt, Treffers, de Beurs, Siedelink, et Koudijs, 2005). En effet, les différentes langues sont souvent loin d'être isomorphes. Chacune est

intrinsèquement liée à des référents qui prennent appui sur un contexte socioculturel dans lequel elle prend essor, se construit et se développe. Par conséquent, dès lors qu'un instrument est adapté, il y a un risque quasi inévitable d'introduire des dissemblances. Par exemple, certains mots clés dans un item pourraient apparaître plus ou moins souvent dans la version traduite que dans la version d'origine. Cela pourrait s'avérer nécessaire pour respecter les conventions linguistiques ou la syntaxe de la langue cible, mais pourrait aussi augmenter ou réduire l'information disponible pour répondre à cet item (Solano-Flores, Backhoff, et Contreras-Niño, 2009).

La clarté de l'instrument ou son niveau de difficulté (Rogers, Gierl, Tardif, et Lin, 2003) est également une source de difficulté potentielle. En effet, la version adaptée d'un instrument est susceptible de ne pas rendre entièrement les idées sous-jacentes associées à une langue ou à ses référents culturels. Par exemple, si une traduction est faite de façon trop littérale, en perdant de vue le sens, la version adaptée risque de perdre autant de sa clarté que de sa crédibilité. Une syntaxe boîteuse introduira une distorsion et rendra les items de piètre qualité, ce qui signifie généralement qu'ils seront alors plus difficiles à comprendre par le groupe visé (Auchter et Stansfield, 1997). De la même façon, si la traduction met l'accent sur une forme linguistique trop simplifiée des items, dans le but de limiter des bris de compréhension, les items risquent alors d'être plus faciles que dans la version d'origine. L'équilibre précaire entre les mots utilisés dans les énoncés et les contenus évalués risque donc souvent d'être perdu lors d'une adaptation d'épreuve (Solano-Flores, Trumbull, et Nelson-Barber, 2002).

D'un point de vue pratique, l'équivalence entre différentes versions d'un instrument ne peut être présumée et doit plutôt être démontrée de façon empirique (ITC, 2010; Sireci et Bastari, 1998; Sireci, Foster, Robin, et Olsen, 1997; Sireci et Khaliq, 2002; Solano-Flores, Trumbull, et Nelson-Barber, 2002; Van de Vijver et Poortinga, 1997). Pour ce faire, Hambleton et Patsula (2000) proposent d'examiner les trois éléments suivants : 1) la comparabilité des groupes ethnolinguistiques visés par l'épreuve par rapport à l'objet d'évaluation; 2) la comparabilité des méthodes d'évaluation utilisées pour chacun de ces groupes; 3) la comparabilité dans l'interprétation des résultats. Plus récemment, Martin

et Blais (2011) réitéraient l'importance de ces trois éléments : « ... il y a ensuite toute la question de la précision des résultats, donc du contrôle des sources d'erreurs comme celles reliées à l'échantillonnage, aux instruments de mesure et aux techniques d'appariement des résultats... » (p. 90). Ainsi l'impossibilité de démontrer l'équivalence entre les différentes versions d'un instrument à l'égard de ces trois catégories risque-t-elle d'introduire des biais dans les données et de fausser les interprétations et les conclusions qui découleront de l'analyse de ces données. Les prochaines sections présenteront donc différentes sources d'invalidité et d'erreur relatives aux deux premières catégories proposées par Hambleton et Patsula (2000). En effet, la troisième catégorie étant liée aux résultats (ce qui implique une cueillette de données), dépasse le cadre des objectifs que nous nous sommes fixés ici.

1. Comparabilité des groupes ethnolinguistiques

Les différents groupes ethnolinguistiques impliqués dans une enquête à grande échelle ne sont pas comparables. Plusieurs caractéristiques diffèrent d'un groupe à l'autre et influencent la façon de répondre des individus. Certains auront une plus grande propension pour l'acquiescence, d'autres auront davantage tendance à deviner ou seront plus sensibles à la désirabilité sociale (Hambleton et Patsula, 2000; Kappelhof, 2014).

De nombreux articles font état d'exemples de différences culturelles trouvées lors du processus d'adaptation d'un instrument ou de comparaisons entre deux groupes (par exemple : Bravo, Canino, Rubio-Stipec, et Woodbury-Farina, 1991; Chavez, Matias-Carrello, Barrio, et Canino, 2007; Gregorio, 2006; Van de Vijver et Hambleton, 1996; Van de Vijver et Tanzer, 2004; van Widenfelt *et al*, 2005). Hambleton et Patsula (2000), par exemple, rapportent les résultats d'une recherche réalisée en Floride qui soulignent que les adultes japonais seraient plus dépressifs que leurs homologues américains. Cependant, une autre étude plus approfondie aurait révélé qu'en fait les Japonais feraient davantage preuve d'inhibition que les Américains dans l'expression de sentiments positifs tels que : *Je suis heureux*. En conséquence, les Japonais auraient obtenu un score moindre sur l'échelle du bonheur, ce qui les faisait paraître plus dépressifs. Quant à Hess et Azuma (1991), ils rapportent que les

différences associées aux différents systèmes d'éducation peuvent avoir une influence sur le développement des jeunes enfants. À titre d'exemple, ils mentionnent que les enfants des États-Unis sont invités à développer la curiosité, la motivation intrinsèque, l'accomplissement de tâches spécifiques et la créativité. En contrepartie, les enfants japonais sont invités à développer la patience, la persévérance et le sens de la minutie dans l'exécution d'une tâche. À cet égard, il ne serait pas étonnant de constater que ces enfants répondent très différemment aux mêmes questions étant donné les différences liées à leur contexte culturel respectif (He et Wolfe, 2010).

D'autres référents culturels propres à chaque groupe ethnolinguistique peuvent influencer la façon de répondre des individus et introduire un biais dans l'interprétation des résultats (Hambleton, 2006; Hambleton et Patsula, 2000). Pour n'en nommer que quelques-uns, nous retrouvons : les programmes scolaires, les politiques éducationnelles, la santé, les habitudes de vies, les valeurs, les structures familiales, la religion, la motivation, etc.

Ainsi les normes⁶ et les référents culturels sont-ils inhérents aux comportements des individus; ils permettent de définir des codes dans les relations qu'ils entretiennent avec le reste du groupe et les autres personnes ou encore de déterminer ce qui semble acceptable ou non (van Widenfelt *et al*, 2005). Dans toute comparaison entre deux groupes, il importe donc de tenir compte de ces différences avant de tirer des conclusions trop hâtives. Dans la mesure du possible, par rapport à l'objet d'évaluation, il importe de comparer des groupes qui sont les plus proches possible sur le plan langagier (par exemple : code alphabétique, langue de tradition écrite, etc.) et culturel même si des différences existent inévitablement (van Widenfelt *et al*, 2005). Si les groupes ne sont pas comparables, il serait nécessaire de le mentionner dans le rapport technique accompagnant l'épreuve.

2. Comparabilité des méthodes d'évaluation utilisées

Il ne suffit pas cependant de travailler avec des groupes comparables pour pouvoir effectuer des comparaisons entre différents groupes ethnolinguistiques. Les méthodes d'évaluation ou les instruments de mesure utilisés doivent également être comparables. Dans tout processus

d'évaluation, il est de commune pratique de s'assurer de la fidélité et de la validité de l'instrument utilisé par rapport aux buts de l'évaluation, de même que pour chacune des populations visées par le test. Les différentes propositions émanant des organismes faisant autorité en la matière, comme celle de l'*American Educational Research Association* (AERA), *American Psychological Association* (APA) et *National Council on Measurement in Education* (NCME) (1999) ou celle de l'ITC (2010), sont assez limpides à ce sujet. Il est également recommandé d'établir la comparabilité des différentes formes d'un instrument avant d'effectuer des comparaisons.

Pour établir l'équivalence entre les différentes versions d'un instrument, plusieurs aspects du processus d'évaluation doivent être étudiés. Selon Hambleton et Patsula (2000), les sources d'invalidité et d'erreur liées à la méthode d'évaluation utilisée peuvent être regroupées selon les cinq éléments suivants : 1) le test lui-même; 2) la sélection et la formation des traducteurs; 3) le processus de traduction; 4) les modèles d'évaluation de la traduction basés sur l'approche qualitative (*judgmental designs for adapting tests*); et 5) les modèles d'évaluation de la traduction basés sur l'approche quantitative et qui sont utilisés afin de démontrer l'équivalence entre les différentes versions du test (analyses empiriques sur les données recueillies). Encore une fois, étant donné l'objectif que nous nous sommes fixé, soit celui de faire le point sur les pièges à éviter lors de l'adaptation d'épreuves, nous examinerons uniquement, dans le cadre de cette synthèse, les quatre premiers éléments que nous venons de citer.

2.1. *Le test lui-même*

Dans un premier temps, étudier la comparabilité des méthodes d'évaluation utilisées pour les différents groupes ethnolinguistiques visés consiste à poser un regard critique sur le test lui-même. Or, cela implique qu'il faut s'assurer de l'équivalence de différents aspects (Hambleton et Patsula, 2000; Martin et Blais, 2011; Sireci, Fitzgerald, et Xing, 1998; Sireci *et al*, 1997; Solano-Flores *et al*, 2002) : les construits mesurés, les niveaux de difficulté (ou les niveaux de connaissances et d'habiletés requis), les modes d'administration utilisés (la logistique de l'opération, l'organisation des séances, etc.), les formats utilisés (autant pour

l'instrument que pour les items), le matériel fourni aux individus, les contextes utilisés pour poser une question, les illustrations incluses dans l'instrument, la correction, la diffusion des résultats, etc. Cela implique également de s'assurer que le niveau de familiarité des sujets évalués avec les formats choisis est le même entre les différents groupes ou que la vitesse de réponse est comparable. Dans les paragraphes qui suivent, examinons ces différents aspects à considérer.

2.1.1. L'équivalence des construits

Vérifier l'équivalence des construits mesurés, par les différentes versions d'un instrument, consiste à effectuer une étude complète des réseaux de variables associés au construit que l'on souhaite évaluer, et ce, dans les différents groupes visés par l'évaluation (Sireci *et al*, 1997). Toutes les dimensions associées aux construits ciblés devraient être les mêmes et avoir la même signification pour tous les groupes de sujets (Grisay et Monseur, 2007). Les items d'un instrument ne sont effectivement utiles que si les différentes versions offertes sont comparables et servent à mesurer une même chose (Van de Vijver et Poortinga, 2005). Même s'il s'agit d'un souhait théorique louable, il est probablement plus réaliste de penser diminuer au maximum l'écart pouvant exister entre les construits appréhendés dans les deux versions linguistiques d'un même item. À cet égard, il importe de distinguer la comparabilité des items sur les plans linguistique et psychologique (Bornman, Sevcik, Ronski, et Pae, 2010; Gregorio, 2006) ou encore sur les plans cognitif, affectif ou psychomoteur. Une traduction littérale qui permet de revenir au texte d'origine par une retraduction vers la langue de départ (voir section 2.4 pour plus de détails sur les modèles d'évaluation de la traduction) permet d'obtenir des items comparables du point de vue linguistique. Mais est-ce que les items ainsi produits sont comparables sur les plans psychologique, cognitif, affectif ou même psychomoteur? Autrement dit, est-ce que la signification pour les sujets est suffisamment semblable pour que ces items puissent être considérés utiles et valides?

À titre d'illustration, l'item « *Who is the president of the United States?* » peut très bien être comparé à l'item « Qui est le président des États-Unis? » sur le plan linguistique. Mais sur le plan psychologique est-il réellement comparable (Van De Vijver et Poortinga, 2005)? Est-ce que

la signification est réellement la même pour des élèves vivant aux États-Unis et au Canada? Un item comparable sur le plan psychologique pour les élèves canadiens devrait plutôt demander le nom du premier ministre du Canada, mais il y aurait alors une perte d'équivalence sur le plan linguistique. Sur le plan cognitif, les items d'une épreuve visant à mesurer l'orthographe de certains mots anglais, que les élèves trouvent parfois difficiles à distinguer, seraient inutiles et invalides une fois traduits en espagnol (Auchter et Stansfield, 1997). En effet, les mots *there* et *their* ne présentent aucune similitude une fois traduits par leurs équivalents espagnols *hay*, *su* ou *allí*. Il en est de même pour les mots *would* et *wood* que les élèves espagnols ne confondent généralement pas avec : *madera* et *hubiera*. D'autres traductions pourraient également être problématiques. C'est le cas pour des concepts tels que *raton laveur*, *borne-fontaine*, *chaton* (*kity*) ou *chiot* (*puppy*), de même que des onomatopées comme « aïe » (*ouch*) ou « miam-miam » (*yum-yum*) qui ne constituent pas des références familières pour les enfants espagnols (Gregorio, 2006), tout comme le concept de bonhomme de neige n'est pas une référence familière en Afrique (Bornman, Sevcik, Ronski et Pae, 2010). Sur le plan affectif, la section 2.1.6 présente quelques exemples où des items portant sur des problèmes de santé, comparables sur le plan linguistique, ne sont pas perçus de la même façon par les Américains et les Hispaniques, ces derniers étant peu habitués à classer leurs problèmes de santé suivant une échelle de type Likert. Enfin, sur le plan psychomoteur, des enfants de Grande-Bretagne et du Zambie se sont vus proposer une tâche comparable : reproduire des modèles en les dessinant sur papier à l'aide d'un crayon et en les modélisant à l'aide d'un fil de fer. Les enfants de Grande-Bretagne ont mieux réussi la tâche en dessinant les modèles, alors que les enfants de la Zambie l'ont mieux réussi en modélisant à l'aide du fil de fer, un passe-temps populaire pour ces derniers (Serpell, 1979). Dans tous ces exemples, c'est une adaptation complète des items qui devient nécessaire (plutôt qu'une simple traduction) afin d'offrir aux différents groupes ethno-linguistiques des items qui sont plus compréhensibles, intelligibles et de même niveau de difficulté pour chacun d'eux.

L'exemple suivant est celui tiré d'une étude de Weisz, Suwanlert, Chaiyasit, Weiss, Achenbach et Walter (1987) cité dans van Widenfelt *et al.*, (2005). Dans cet exemple, un des items du *Child Behavior Checklist*

(CBCL) (Achenbach et Edelbrock, 1983), un instrument qui permet aux parents, ou à toute personne proche d'un enfant, d'évaluer les problèmes de comportement de ce dernier a permis de constater que deux fois plus de parents thaïlandais ont rapporté une utilisation de jurons ou de langage obscène par leur enfant que les parents états-uniens. En fait, cette différence importante est liée au sens accordé au verbe *jurer*. Pour les Thaïlandais, l'utilisation d'un langage qui, pour les Américains, n'est considéré que comme une légère impolitesse, fait partie de la définition même de *jurer*. Autrement dit, la traduction thaïlandaise du mot engloberait un construit plus vaste que le terme original. Quant à eux, Swanson et Watson (1982), cité dans Hambleton (1993) rapportent que le mot espagnol *paloma* est équivalent aux mots anglais *dove* et *pigeon*. Dans ce cas précis, il serait donc impossible pour des hispanophones de noter une distinction ou une nuance si on procédait à une traduction de l'anglais vers l'espagnol. Ainsi, dans certains cas, bien que la traduction puisse être effectuée avec soin, certaines nuances sémantiques peuvent être difficiles à corriger, car elles font appel à des référents culturels (van Widenfelt *et al*, 2005).

Enfin, à titre de dernier exemple, Rogers *et al* (2003) présentent un item visant à mesurer l'habileté des élèves en mathématiques qui porte sur la durée d'une journée de tournage. Dans la version anglaise, l'information était offerte de la façon suivante : « *they arrived at 5:20 am and left at 8:15 pm. How long...?* ». L'équivalent français de cet item était le suivant : « ils sont arrivés à 5h20 du matin et sont repartis à 20h15. Combien de temps...? ». De prime abord, les deux énoncés semblent assez comparables. En y regardant de plus près, il est possible de constater que le construit de ces deux items n'est pas exactement le même. Dans la version anglophone, l'énoncé réfère au système de notation des heures sur une période de 12 heures alors qu'il réfère au système sur 24 heures dans la version francophone. Cette différence, mineure en apparence, a cependant eu un impact important sur les élèves. En effet, les résultats ont révélé que les élèves francophones avaient mieux réussi que les élèves anglophones, et ce, de façon assez marquée. Ce léger changement dans le construit a effectivement eu pour effet de changer le niveau de difficulté de l'item, autre aspect important à vérifier pour assurer l'équivalence entre les différentes versions d'un test.

2.1.2. L'équivalence des niveaux de difficulté

Pour assurer la comparabilité des différentes versions d'un instrument, le deuxième aspect qu'il importe d'étudier par rapport au test lui-même est son niveau de difficulté dans les différentes versions offertes ainsi que celui de chacun des items qu'il contient. En définitive, il faut vérifier que les niveaux de connaissances et d'habiletés requis pour chacun des groupes ethno linguistiques impliqués dans l'évaluation sont comparables (Solano-Flores *et al*, 2002). Dans l'exemple précédent, portant sur la durée de tournage, le simple changement de référentiel pour le système de notation des heures a rendu le niveau d'habileté requis plus facile pour les élèves francophones. En effet, contrairement aux élèves francophones qui devaient effectuer une simple soustraction entre l'heure de départ et l'heure d'arrivée pour résoudre ce problème, les élèves anglophones devaient procéder en au moins deux étapes : 1) soit débiter en déterminant la durée du tournage effectuée le matin et celle effectuée en après-midi puis additionner les deux; 2) soit convertir les heures selon le système de notation en 24 heures, puis procéder en soustrayant l'heure d'arrivée à l'heure de départ. De tels changements dans le niveau de difficulté des différentes versions d'un instrument risquent d'introduire des biais importants dans les données et de fausser les interprétations et les conclusions qui découleront de l'analyse de ces données s'ils ne sont pas étudiés, corrigés ou, à tout le moins, pris en compte lors de l'analyse des données.

L'exemple suivant permet également de constater l'impact important d'une traduction rendant ainsi une des versions de l'épreuve plus facile. Dans une étude comparative portant sur la lecture, les étudiants devaient considérer des paires de mots et indiquer si la signification de ces deux mots était similaire ou différente. Chez les étudiants qui ont fait l'épreuve aux États-Unis, la paire de mots *pessimistic-sanguine* est l'une de celle où les étudiants ont à peine mieux réussi que si leur réussite était le fruit du hasard (54 %). Une analyse de l'instrument présenté au pays qui s'est classé premier à cette épreuve, où 98 % des étudiants ont réussi cet item, a permis de découvrir que le mot *sanguine* n'avait pas son équivalent dans cette langue et que la traduction la plus proche de ce mot correspondait au mot *optimistic*. En conséquence, il était beaucoup plus facile pour les répondants de cette version de l'épreuve de déterminer que les

mots *pessimiste* et *optimiste* sont en opposition (Hambleton et Patsula, 2000).

2.1.3. L'équivalence des modes d'administration

Le mode d'administration utilisé pour présenter l'épreuve aux répondants est un autre élément lié au test lui-même qu'il importe d'étudier. Plusieurs aspects relèvent de cet élément et, afin d'assurer la comparabilité des données recueillies et la validité des conclusions qui en découleront, il est pertinent d'étudier leur équivalence dans les différentes versions de l'épreuve. Voici quelques exemples : les locaux (taille, capacité), les circonstances (heure de l'épreuve, jour, instructions, surveillance), le matériel requis (format de l'épreuve, épreuve uniforme, dictionnaire), le temps de passation, la correction⁷, etc.

Pour compléter ces quelques exemples au regard de la comparabilité des modes d'administration, il importe de vérifier deux autres aspects : le format de l'instrument qui sera utilisé, de même que l'échelle de réponse qui sera proposée aux sujets évalués. Étant donné l'importance relative de ces deux aspects, ils seront abordés séparément dans les sections suivantes.

2.1.4. L'équivalence du format de l'instrument

Le format de l'instrument choisi pour administrer l'épreuve est un autre élément relatif au test lui-même qu'il importe d'étudier pour assurer la comparabilité des méthodes d'évaluation utilisées pour les différents groupes ethnolinguistiques par une évaluation. Dans un premier temps, le format de l'instrument réfère au type de support qui sera utilisé : papier-crayon, en ligne, etc. Il est facile d'imaginer l'impact sur les données recueillies si l'un des groupes devait réaliser une version informatisée d'un test à l'aide d'ordinateurs récents et des connexions Internet à haute vitesse par rapport à un autre groupe où les ordinateurs seraient moins puissants, moins rapides et où les connexions Internet seraient à basse vitesse.

Cela réfère également au type de tâche qui sera demandé aux sujets évalués (manipulations en laboratoire, essais, exposés, questions à choix multiples, à réponse courte, etc.), aux outils qui seront présentés dans ces tâches (illustrations, textes, formules, graphiques, etc.), au format de

présentation qui sera utilisé (regroupement des items, ordre de présentation, choix des mots, ponctuation, mise en page, etc.), aux instructions, etc.

Certaines études ont montré que les données fournies dans le texte d'introduction utilisé pour présenter un instrument de mesure, préciser son contexte ou ses objectifs, exerçaient une influence sur les répondants et, par conséquent, sur les réponses recueillies (Norenzayan et Schwarz, 1999; Tourangeau et Rasinski, 1988; Wänke, et Schwarz, 1997). De la même façon, le choix des mots pourrait avoir une influence sur les données recueillies. Deux items, utilisant des formulations comparables, peuvent effectivement induire une perception différente du sens ou de l'intention véhiculée par l'item ou son échelle, et ainsi modifier les réponses des répondants (Blais et Grondin, 2011). Par exemple, en réponse à des items auxquels les finissants d'un baccalauréat en éducation préscolaire et enseignement primaire devaient indiquer à quel point leur programme de formation les avait préparés à « Corriger les productions écrites des élèves » ou à « Corriger la langue écrite des élèves », des différences d'environ 20 % ont pu être observées entre le niveau d'endossement de ces deux items. Une différence assez marquée pour conclure qu'il s'agit de deux items différents, même s'ils portent sur un contenu très proche (qui se voulait comparable). Une différence de 13,2 % a également pu être observée entre les items « Adapter mes activités d'enseignement aux caractéristiques des élèves » et une formulation plus générale comme « Adapter mes activités d'enseignement ». Mais l'exemple le plus marquant est sûrement celui lié à une étude américaine où environ 20 % des répondants étaient généralement en faveur de “ne pas permettre” comparativement à « interdire » des déclarations politiques contre la démocratie (Rugg, 1941).

Enfin, le format d'item choisi peut également influencer les répondants : les questions de type choix multiples, bien connues et largement utilisées en Amérique du Nord, pourraient défavoriser des répondants issus de systèmes d'éducation d'inspiration britannique, qui utilisent davantage les questions à réponse courte ou les essais (Hambleton et Patsula, 2000). Ce format d'items pourrait également désavantager des étudiants chinois qui ont plutôt l'habitude de faire des calculs, des démonstrations ou des preuves et qui trouvent les questions à

choix multiples intéressantes, mais un peu longues et fatigantes (nombre d'items plus élevé qu'à l'habitude) (Hambleton, Yu, et Slater, 1999).

La mise en forme des items ou de l'instrument pourrait elle aussi avoir son importance (Sanchez, 1992). En effet, la disposition du texte, la police de caractère utilisée ou la longueur du texte peuvent également introduire des biais dans les données (Cyr et Trevor-Smith, 2004; Stern, Smyth, et Mendez, 2012). Par exemple, certaines langues nécessitent, des structures phrastiques plus longues que d'autres (ex. français *versus* anglais). Ainsi, un item peut s'afficher à l'écran sur deux pages par rapport à son équivalent, plus court, qui s'affiche sur une seule page, et, ainsi, prendre plus de temps à être lu et répondu. L'affichage sur deux pages peut même augmenter le niveau de difficulté de l'item parce qu'il doit faire appel à la mémoire à court terme du sujet évalué en plus du construit visé par l'item (Lessler, 1995). Il importe donc d'étudier soigneusement le format de l'instrument pour éviter d'introduire des sources d'invalidité ou d'erreur dans la comparaison des résultats entre divers groupes linguistiques ou culturels.

2.1.5. L'équivalence de la longueur du texte ou de la vitesse de réponse

Parmi les aspects relatifs au format de l'instrument, l'équivalence de la longueur du texte est de nature secondaire par rapport à l'équivalence de construit (objet d'évaluation), mais cela mérite tout de même réflexion. L'expansion de texte entre les différentes versions d'un instrument est fréquente. Dans certains cas, elle peut être liée à la présence de certains mots ou concepts qui n'existent pas pour certains groupes ethnolinguistiques. Par exemple, la notion de « voyageurs de la liberté » (*freedom rider*), ne possède aucun équivalent espagnol. Par conséquent, lorsque cette expression doit être traduite, le traducteur se doit de paraphraser ou définir l'expression, ce qui peut rendre la traduction beaucoup plus longue à lire ou plus difficile à comprendre (Auchter et Stansfield, 1997).

Mais même en l'absence de tels concepts à définir ou à paraphraser, la longueur du texte peut varier de façon relativement importante d'une langue à l'autre. C'est le cas par exemple de l'allemand qui requiert généralement plus de texte que son homologue anglais (Sireci *et al*, 1997). L'expansion de texte entre ces deux langues peut être de l'ordre de 10 à

35 %. Pour le français, l'expansion peut être de l'ordre de 15 à 20 %, et pour le japonais, les variations de longueur peuvent atteindre jusqu'à 60 % (Wooten, 2011). De plus, les variations entre deux langues peuvent être asymétriques. Par exemple, le passage de l'anglais à l'espagnol requiert par exemple de 15 % à 30 % plus de mots alors que l'inverse nécessite de 5 % à 15 % moins de mots.

Aussi, si le texte est plus long dans une des versions traduites de l'épreuve, il est possible de penser que cette version pourrait requérir plus de temps à ses répondants pour lire les items et y répondre (Sireci *et al*, 1997). Le temps de réponse constituerait donc un autre aspect à considérer pour déterminer l'équivalence entre deux versions d'un instrument.

2.1.6. L'équivalence des échelles de réponses

Le dernier aspect qu'il importe d'étudier par rapport au test lui-même est l'équivalence des échelles de réponses. En effet, les échelles de réponses ne sont pas automatiquement équivalentes parce qu'elles proposent des choix de mots équivalents. Par exemple, dans le *Peabody Picture Vocabulary Test – III* (PPVT-III) (Dunn et Dunn, 1997), les enfants doivent identifier l'image représentant une échelle. Or, la traduction littérale espagnole d'une échelle est « *escalera* ». En conséquence, la version espagnole de cet item contiendrait alors deux bonnes réponses : l'échelle et l'escalier (Gregorio, 2006).

Aussi, même si une échelle de réponse est traduite ou adaptée avec soin, des changements de sens très subtils peuvent tout de même être introduits et avoir une influence importante sur les caractéristiques d'un item et, par la suite, sur les données recueillies (van Widenfelt *et al*, 2005). Par exemple, un des items du questionnaire CARES (*Cancer Rehabilitation Evaluation System*), qui vise à comprendre, documenter et évaluer les besoins de patients atteints du cancer, portait sur la prise de poids. Cet item est formulé comme suit « *I cannot gain weight* ». A cet item, une patiente de 70 livres qui aurait probablement dû répondre que cette situation correspondait à la sienne *very much*, a plutôt répondu *not at all*, parce qu'effectivement, elle n'arrivait pas du tout à prendre du poids (Canales, Ganz, et Coscarelli, 1995).

Enfin, il importe de s'assurer que la traduction ou l'adaptation des choix de réponses dans les différentes versions d'une épreuve n'introduise

pas d'écarts ou de changements dans la distribution des réponses d'un groupe culturel ou linguistique à l'autre (van Widenfelt, et coll., 2005). En effet, les changements dans les distributions de réponses entre les différents groupes évalués devraient être liés aux variations existant dans l'objet d'évaluation et non pas à la façon dont l'item ou ses choix de réponses ont été traduits ou adaptés. Par exemple, Canales, Ganz et Ciscarelli (1995) ont pu constater que les patients hispaniques, peu habitués à réfléchir à leurs problèmes de santé de cette façon, avaient de la difficulté à associer les valeurs numériques proposées dans leur questionnaire aux étiquettes (p. ex. : 0 = pas du tout). De la même façon, pour certains patients, il était ardu de décider à quel point leurs problèmes les affectaient *much* ou *very much*. Selon Hui et Triandis (1989), les hispaniques auraient une plus grande facilité à utiliser les catégories extrêmes d'une échelle de Likert en cinq points que les non-hispaniques, mais cette différence serait moindre sur une échelle en 10 points.

2.2. *La sélection et la formation des traducteurs*

Dans un deuxième temps, pour étudier la comparabilité des méthodes d'évaluation utilisées, Hambleton et Patsula (2000) suggèrent de poser un regard critique sur la sélection et la formation des traducteurs. En effet, à leur avis, les chercheurs ou les administrateurs d'épreuve ont trop souvent tenté de demander à un seul traducteur d'effectuer l'ensemble de la traduction (adaptation) pour des raisons de coûts, de disponibilités, etc. Mais l'utilisation d'un seul traducteur, sans égard à son niveau de compétence, ne permet pas le recul suffisant pour détecter les problèmes potentiels liés à l'adaptation d'une épreuve ni les échanges entre traducteurs permettant de les résoudre.

En conséquence, il semble que le recours à une équipe de traducteurs compétents soit requis, ce qui sous-entend une familiarité et des compétences linguistiques avec la langue ciblée pour la traduction (Hambleton, 2006; Hambleton et Patsula, 2000). De plus, de bonnes connaissances des cultures visées par l'épreuve, et particulièrement de celle vers laquelle la traduction sera faite, sont nécessaires, car c'est là l'essence même d'une adaptation efficace. Les personnes retenues devraient également posséder une bonne connaissance du domaine évalué afin que les nuances ou subtilités propres à ce domaine ne soient pas

perdues au cours de la traduction. En effet, lorsqu'un traducteur manque de connaissances relativement au domaine évalué, il en découle généralement une traduction plus littérale, ce qui peut nuire à la compréhension des répondants et à la validité du test.

Enfin, il pourrait être important de dispenser aux traducteurs une formation sur la construction de tests (développement de l'épreuve, écriture d'items, construction d'une échelle de réponses, comptabilisation de scores, etc.) (Hambleton, 2006; Hambleton et Patsula, 2000). En effet, traduire ou adapter une épreuve sans connaître certains fondements psychométriques pourrait augmenter ou diminuer la difficulté de l'épreuve et, par la suite, en affecter la validité pour le groupe ethnolinguistique visé. Par exemple, de façon inconsciente, les traducteurs pourraient introduire des biais liés à des choix de mots basés sur des associations sonores plutôt que sur une base sémantique, ce qui pourrait fournir des indices aux répondants quant à la bonne réponse.

2.3. Le processus de traduction

En troisième lieu, pour déterminer la comparabilité des méthodes d'évaluation utilisées, Hambleton et Patsula (2000) suggèrent d'étudier le processus de traduction qui sera ou qui aura été utilisé. En effet, afin de rendre la traduction (ou l'adaptation) la plus juste ou la plus appropriée possible, il importe de s'intéresser à la façon dont celle-ci sera ou aura été produite. Idéalement, il est préférable de commencer à s'intéresser au processus de traduction (adaptation) au moment même de la construction de la version originale de l'épreuve. Comme l'indiquent Rogers *et al* (2003), les développeurs de tests ont généralement les compétences nécessaires pour développer l'épreuve dans leur langue d'origine relativement aux connaissances de langue, de la culture, du domaine, etc. Mais, les développeurs possèdent rarement les connaissances d'autres langues ou d'autres cultures visées par une adaptation d'épreuve. Ce manque de connaissances peut conduire à la construction d'une version plutôt ethnocentrique de l'épreuve de départ qui sera ensuite difficile à adapter, et qui forcera les traducteurs à en faire une traduction plus littérale. Par ricochet, cette adaptation manquera de naturel et certaines connotations auront disparu.

Enfin, étudier le processus de traduction consiste à poser un regard critique sur le modèle de traduction qui sera ou qui aura été utilisé. Deux catégories de modèles d'évaluation de la traduction sont principalement utilisées : les modèles basés sur une approche qualitative et ceux basés sur une approche quantitative. Ces modèles constituent les deux derniers éléments proposés par Hambleton et Patsula (2000) pour déterminer l'équivalence des méthodes d'évaluation utilisées. Au regard des objectifs de ce texte, seuls les modèles d'évaluation basés sur une approche qualitative seront examinés dans la prochaine section.

2.4. *Modèles d'évaluation de la traduction basés sur l'approche qualitative (judgmental designs for adapting tests)*

Étudier de façon qualitative le modèle de traduction qui sera utilisé lors du processus de traduction constitue donc le quatrième élément proposé par Hambleton et Patsula (2000) pour assurer la comparabilité des méthodes d'évaluation utilisées pour les différents groupes ethnolinguistiques visés par une épreuve. Les modèles de traduction basés sur une approche qualitative sont des modèles utilisant le jugement d'un ou plusieurs experts pour déterminer le degré d'équivalence entre les différentes versions d'une épreuve. Bien que basés sur le jugement, ces modèles permettent de détecter, autant que possible, les items qui pourraient s'avérer problématiques, et ce, avant même leur utilisation dans une analyse quantitative ou une mise à l'essai (Elosua, Hambleton, et Zenisky, 2006). Dans certains cas, cela pourrait permettre de sauver temps et argent en corrigeant les problèmes liés à ces items avant leur utilisation. Dans d'autres cas, il pourrait être utile de les conserver tels quels en attendant les résultats de l'analyse statistique ou de la mise à l'essai pour prendre une décision (van Widenfelt *et al*, 2005).

Les deux modèles de traduction les plus souvent utilisés sont la traduction « vers l'avant » (*forward translation*, ou traduction de la langue de départ « vers la langue cible ») et la traduction « vers l'arrière » (*backward translation*, ou traduction de la langue de départ vers la langue cible et « retraduction⁸ de la langue cible vers la langue de départ ») (Hambleton et Patsula, 2000). Dans le premier cas, la version traduite peut être évaluée par des experts bilingues, qui compareront la version d'origine et la version traduite pour en déterminer l'équivalence, ou par

des sujets issus du groupe ethnolinguistique ciblé par la version traduite et par des experts qui interrogeront ces derniers sur leurs réponses. L'avantage de cette méthode est qu'elle permet d'étudier les deux versions (départ et cible) de près et, par conséquent, d'établir avec plus de certitude l'équivalence des deux versions.

Tassé et Craig (1999), par exemple, ont proposé une méthode par comité d'experts en sept étapes : 1) traduction et adaptation de l'épreuve par un premier comité d'experts; 2) comparaison des traductions individuelles des experts et consolidation en une version consensuelle de l'épreuve; 3) validation de la version adaptée de l'épreuve par un deuxième comité d'experts, indépendant du premier; 4) révision et ajustement de l'épreuve par les deux comités afin de former une version prétest de l'épreuve; 5) prétest auprès d'un petit groupe de potentiels utilisateurs de l'épreuve, qui a pour but d'évaluer la clarté des consignes, des items, de la présentation de l'épreuve, etc.; 6) révision et ajustement de l'épreuve par le premier comité d'experts; 7) validation et essai de la version finale.

Dans le deuxième cas, c'est la retraduction vers la langue d'origine qui permet d'obtenir une première estimation sur la qualité de la traduction (Grégoire, 2006). Par exemple, elle permet d'entrevoir si le contenu, le sens (signification), les consignes ou les catégories de réponses sont comparables à leur version d'origine. Cette traduction vers l'arrière devrait être faite à l'aveugle (*blind-back translation*), c'est-à-dire que ce sont des traducteurs qui n'ont pas participé à la traduction de la langue d'origine vers la langue cible qui devraient l'exécuter (Bornman *et al.*, 2010). De plus, une seule itération ne serait généralement pas suffisante pour obtenir une traduction de qualité (van Widenfelt *et al.*, 2005). C'est pourquoi, le processus de traduction-retraduction entre la langue d'origine et la langue cible peut être répétée plusieurs fois. L'inconvénient de ce processus est que, traduire l'épreuve dans le but de retrouver le texte de départ lors de la retraduction, risque de conduire à une certaine standardisation des différentes versions offertes. En effet, la traduction risque alors de devenir plus littérale afin que les deux versions de l'épreuve soient linguistiquement équivalentes. Or, tel qu'il a été discuté précédemment (voir section 2.1.1), deux versions linguistiquement équivalentes ne le sont pas nécessairement psychologiquement (Gregorio,

2006). D'ailleurs, selon l'élément D.5 du guide de traduction et d'adaptation (ITC, 2010), il est recommandé de démontrer l'équivalence des deux versions d'une épreuve relativement à ces deux aspects. De plus, en appliquant le processus de traduction « vers l'arrière » de façon trop stricte, c'est la qualité de la traduction qui peut en souffrir puisque les nuances et les subtilités propres à une langue ou une culture auront été supprimées.

Selon Solano-Flores *et al* (2002), les différents modèles qui impliquent de traduire la version d'origine de l'épreuve comporteraient des limites théoriques, méthodologiques et pratiques importantes liées aux deux aspects suivants : d'une part, ces modèles ne tiendraient pas suffisamment compte du fait que la langue et la culture sont indissociables et, d'autre part, le processus de traduction de ces modèles aurait pour effet que le choix des mots utilisés dans les versions adaptées de l'épreuve ne découlerait pas du même processus de raffinement que celui utilisé pour la version d'origine de l'épreuve.

Afin de favoriser des évaluations plus équitables pour les différents groupes ethnoлингuistiques visés par une épreuve, un autre modèle, qui consiste à produire de façon parallèle les différentes versions de l'instrument, a été proposé. Ce développement simultané des différentes versions de l'épreuve (*simultaneous translation test development*) a l'avantage d'intégrer des représentants des différents groupes visés par l'épreuve dès ses premiers stades de développement (Rogers *et al*, 2003). Ainsi, toute l'information relative à certaines particularités linguistiques ou culturelles est-elle facilement accessible et les problèmes potentiels liés à ces particularités peuvent être détectés et traités au fur et à mesure que le test se développe. Le risque d'introduire des biais de construit s'en trouve diminué, de même que celui de développer une épreuve ethnocentrique puisque les différentes versions de l'épreuve étant toutes au même stade de développement, il est alors facile de les modifier. Par contre, un désavantage associé à ce modèle de traduction est qu'il peut être difficile de travailler en conciliant simultanément et continuellement au moins deux groupes cibles. De la même façon, il peut être difficile d'acquérir une certaine aisance avec le cursus scolaire associé à chacun des groupes cibles.

Enfin, plutôt que d'offrir une version unilingue adaptée à chacun des groupes linguistiques ou culturels, il est également possible de mettre en place une version bilingue de l'instrument permettant aux individus de chacun des groupes de choisir s'ils préfèrent répondre aux items de l'instrument dans sa langue d'origine ou dans sa version adaptée (Sireci et Khaliq, 2002; van Widenfelt *et al.*, 2005). En effet, dans une version bilingue, les individus ont la possibilité de consulter simultanément chaque question de l'épreuve dans les deux langues. Ils peuvent aussi fournir une réponse à chaque question dans la langue de leur choix en fonction de leur familiarité langagière avec le contenu de la question. Cela permet donc aux répondants de choisir la langue qui, à leur avis, leur permettra d'offrir la meilleure performance. Cette alternative pourrait ainsi neutraliser certaines incompréhensions d'ordre linguistique susceptibles d'être à l'origine de biais d'items et de concepts. Pour plusieurs, cette solution constitue une avenue prometteuse pour améliorer la validité des comparaisons interculturelles, particulièrement chez les groupes linguistiques minoritaires en réduisant l'impact d'un faible niveau d'habileté dans une langue sur le score du répondant (Duncan, Parent, Chen, Ferrera, Johnson, Oppler, et Shich, 2005; Sireci *et al.*, 1997; Sireci et Khaliq, 2002). Il serait ainsi possible de favoriser l'équité entre les différents groupes ethnolinguistiques visés par l'épreuve et de respecter la diversité pluriethnique et linguistique, reflet des populations scolaires d'aujourd'hui.

Conclusion

Ce texte nous rappelle à quel point les sources d'invalidité et d'erreur peuvent être nombreuses. Cela est d'autant plus vrai quand il s'agit de traduire ou d'adapter des épreuves qui visent à respecter l'hétérogénéité de la population à laquelle celles-ci sont destinées. Et si même l'opération de traduction ou d'adaptation est bien réalisée, il n'est pas non plus certain que les répondants de cultures diverses interpréteront de façon équivalente les questions (Kankaraš et Moors, 2010). Même si les méthodes se peaufinent, la vigilance reste de mise quant à l'équivalence des instruments.

Des pistes prometteuses existent et méritent d'être explorées pour améliorer les processus de traduction et d'adaptation. On peut penser, par

exemple, aux propositions de Sireci et Khaliq (2002) qui soulignent l'importance d'effectuer des recherches portant sur le développement et les pratiques d'administration d'épreuves bilingues pouvant améliorer la validité des comparaisons entre les groupes ethnolinguistiques. Ils soulèvent également l'importance de s'interroger sur l'utilisation que font les répondants des différentes versions linguistiques de chaque question de l'épreuve bilingue. Même s'il est possible de contribuer à l'amélioration de la validité des comparaisons entre les groupes linguistiques en administrant une version bilingue de l'épreuve, la modalité papier-crayon complexifie et rend très coûteux le processus d'administration étant donné le grand nombre de questions administrées et la possibilité de consulter chaque question dans les deux langues pour un nombre important de répondants. De plus, il faut leur accorder plus de temps pour répondre aux questions de l'épreuve bilingue, ce qui peut entraîner des problèmes de comparaison (Duncan, Parent, Chen, Ferrera, Johnson, et Oppler, 2005). Avec les développements importants des technologies de l'information et des communications (TIC), l'administration de versions bilingues d'une épreuve est maintenant une avenue qu'il est possible d'envisager sérieusement. D'autres études sur ce sujet seraient donc nécessaires pour approfondir la question. Par exemple, les répondants de l'épreuve bilingues pourraient-ils être défavorisés par rapport aux répondants unilingues? Même si, grâce à une version informatisée d'une épreuve, les répondants peuvent basculer d'une langue à l'autre presque instantanément, il n'en reste pas moins que les répondants qui prennent le temps de lire les deux versions d'une question risquent de prendre plus de temps que les autres à répondre (Sireci *et al*, 1997; Stansfield, 1997).

Bibliographie

- Achenbach, T. M., et Edelbrock, C. S. (1983). *Manual for the Child Behavior Checklist and Revised Child Behavior Profile*. Burlington : University of Vermont, Department of Psychiatry.
- American Educational Research Association (AERA), American Psychological Association (APA), et National Council on Measurement in Education (NCME) (1999). *Standards for educational*

- and psychological testing*. Washington, D.C. : American Educational Research Association (AERA).
- Auchter, J. E., et Stansfield, C. W. (1997). *Developing parallel tests across languages: Focus on the translation and adaptation process*. Communication présentée à la conférence annuelle Large Scale Assessment, Colorado Springs, CO. June 15-18.
- Bastin, G. L. (1990). Traduire, adapter, réexprimer. *Meta : journal des traducteurs*. 35:3.470-475.
- Blais, J.-G., et Grondin, J. (2011). *Les sources d'influences liées aux questionnaires d'enquête auto-administrés : le bilan d'une étude effectuée à l'aide de la famille des modèles de Rasch*. Communication présentée dans le cadre du colloque sur L'interdisciplinarité de la mesure et de l'évaluation, lors du 79^e congrès de l'ACFAS, Université de Sherbrooke, Canada. 10 mai.
- Bornman, J., Sevcik, R. A., Ronski, M., et Pae, H. K. (2010). Successfully Translating Language and Culture when Adapting Assessment Measures. *Journal of Policy and Practice in Intellectual Disabilities*. 7:2.111-118.
- Bravo, M., Canino, G., Rubio-Stipec, M., et Woodbury-Farina, M. (1991). A cross-cultural adaptation of a psychiatric epidemiologic instrument: The diagnostic interview schedule's adaptation in Puerto Rico. *Culture, Medicine and Psychiatry*. 15.1-18.
- Canales, S., Ganz, P. A., et Coscarelli, C. A. (1995). Translation and Validation of a Quality of Life Instrument for Hispanic American Cancer. *Quality of Life Research*. 4:1.3-11.
- Chalifoux, J.-J. (1993). Culture : une notion polémique? *Service social*, 42.1.11-23.
- Chavez, L. M., Matias-Carrelo, L., Barrio, C., et Canino, G. (2007). The cultural adaptation of the Youth Quality of Life instrument: Research version for Latino children and adolescents. *Journal of Child and Family Studies*.16.75-89.
- Comité consultatif mixte (1993). *Principes d'équité aux pratiques d'évaluation des apprentissages au Canada*. Edmonton, Alberta :

- Centre for Research in Applied Measurement and Evaluation. Repéré à http://www2.education.ualberta.ca/educ/psych/crame/files/fr_princ.pdf
- Cyr, D., et Trevor-Smith, H. (2004). Localization of Web Design: An empirical comparison of German, Japanese, and U.S. website characteristics. *Journal of the American Society for Information Science and Technology*. 55.13.1-10.
- Duncan, T. G., Parent, L. d. R., Chen, W.-H., Ferrera, S., Johnson, E., et Oppler, S. (2005). Study of a dual-language test booklet in eighth-grade mathematics. *Applied measurement in education*. 18:2.129-161.
- Dunn, L. M., et Dunn, L. M. (1997). *Peabody Picture Vocabulary Test - Third Edition (PPVT-III)*. Repéré à <http://psychcorp.pearsonassessments.com/HAIWEB/Cultures/en-us/Productdetail.htm?Pid=PAa12010>
- Elosua, P., Hambleton, R. K., et Zenisky, A. (2006). *Improving the methodology for detecting biased test items*. Communication présentée à la 5^e conférence de l'International Test Commission (ITC), Brussels, Belgium. July 6-8.
- Elosua, P., et López-Jauregui, A. (2008). Equating Between Linguistically Different Tests: Consequences for Assessment. *The Journal of Experimental Education*. 76:4.387-402.
- (2007). Potential Sources of Differential Item Functioning in the Adaptation of Tests. *International Journal of Testing*. 7:1.39-52.
- European Federation of Psychologists' Associations (EFPA)*, et *European Association of Work and Organizational Psychologists (EAWOP)*. (2005). *European Test User Standards for test use in Work and Organizational settings, version 1.92*. Repéré à <http://www.eawop.org/uploads/datas/10/original/European-test-user-standards-v1-92.pdf?1297020028>
- Grégoire, J. (2006). *Some obstacles ahead for meeting the guidelines for test translation, and potential solutions*. Communication présentée à la 5^e conférence de l'International Test Commission (ITC), Brussels, Belgium. July 6-8.

- Gregorio, S. C. (2006). *The influence of item content in cultural adaptations*. Communication présentée à la 5^e conférence de l'International Test Commission (ITC), Brussels, Belgium. July 6-8.
- Grisay, A., et Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in educational evaluation*. 33:1.69-86.
- Hambleton, R. K. (2006). *International Test Commission Guidelines for test adaptation, second edition (draft)*. Communication présentée à la 5^e conférence de l'International Test Commission (ITC), Brussels, Belgium.
- (1993). *Translating achievement tests for use in cross-national studies*. International Association for the Evaluation of Educational Achievement, New York, NY. National Center for Education Statistics (ED), Washington, DC.
- Hambleton, R. K., et Patsula, L. (2000). *Adapting tests for use in multiple languages and cultures. Laboratory of Psychometric and Evaluative Research Report*. Massachusetts University, Amherst School of Education.
- Hambleton, R. K., Yu, J., et Slater, S. C. (1999). Fieldtest of the ITC Guidelines for Adapting Educational and Psychological Tests. *European Journal of Psychological Assessment*. 15:3.270-276.
- He, W., et Wolfe, E. W. (2010). Item Equivalence in English and Chinese Translation of a Cognitive Development Test for Preschoolers. *International Journal of Testing*. 10.80-94.
- Hess, R. D., et Azuma, M. (1991). Cultural support for schooling: Contrasts between Japan and the United States. *Educational Researcher*. 20.2-8.
- Hui, C. H., et Triandis, H. C. (1989). Effects of Culture and Response Format on Extreme Response Style. *Journal of Cross-Cultural Psychology*. 20.296-309.
- International Test Commission (ITC) (2010). *International Test Commission Guidelines for Translating and Adapting Tests*. Repéré à <http://www.intestcom.org>

- Joint Advisory Committee (1993). *Principles for Fair Student Assessment Practices for Education in Canada*. Edmonton, Alberta : Committee Centre for Research in Applied Measurement and Evaluation. Repéré à http://www2.education.ualberta.ca/educ/psych/crame/files/eng_prin.pdf
- Kankaraš, M., et Moors, G. (2010). Researching Measurement Equivalence in Cross-Cultural Studies. *Psihologija*. 43:2.121-136.
- Kappelhof, J. W. S. (2014). The impact of method bias on the cross-cultural comparability in face-to-face surveys among ethnic minorities. *methods, data, analysis*. 8:1.79-118.
- Le grand dictionnaire terminologique (1988). *Fiche terminologique du mot retraduction relativement au domaine de la linguistique*. Repéré à http://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=17035764
- (1982). *Fiche terminologique du mot adaptation relativement au domaine de l'édition*. Repéré à http://www.gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=8433203
- (1979). *Fiche terminologique du mot norme relativement au domaine de la sociologie*. Repéré à http://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=8369954
- Leplège, A. (2001). Adaptation transculturelle d'instruments existants. In A. Leplège et J. Coste (dir.). *Mesure de la santé perceptive et de la qualité de vie : méthodes et applications*. Paris : Editions Estem. 191-205.
- Linton, R. (1968). *Le fondement culturel de la personnalité*. Paris, France : Dunod.
- Lessler, J. T. (1995). Choosing questions that people can understand and answer. *Medical Care*. 33:4.AS203-AS208.
- Martin, R., et Blais, J.-G. (2011). Une nouvelle perspective et de nouveaux défis pour les enquêtes internationales : vers un testing assisté par ordinateur. *Mesure et Évaluation en Éducation*. 34:2.87-112.
- Mounin, G. (2000). *Clefs pour la linguistique*. Paris : Editions 10-18.

- Norenzayan, A., et Schwarz, N. (1999). Telling what they want to know: Participants tailor causal attributions to researchers' interests. *European Journal of Social Psychology*. 29:8.1011-1020.
- Rogers, W. T., Gierl, M. J., Tardif, C., et Lin, J. (2003). *Differential validity and utility of successive and simultaneous approaches to the development of equivalent achievement tests in French and English*. Communication présentée le 24 avril à la réunion annuelle du *National Council on Measurement in Education* (NCME), Chicago, Illinois.
- Rugg, D. (1941). Experiments in wording questions : II. *Public opinion quarterly*. 5:1.91-92.
- Sanchez, M. E. (1992). Effects of questionnaire design on the quality of survey data. *The Public Opinion Quarterly*. 56.2.206-217.
- Serpell, D. (1979). How specific are perceptual skills?. *British Journal of Psychology*. 70.365-380.
- Simões Forte, L. A., et Dionne, E. (2013). L'évaluation formative de la compétence interculturelle : regard sur le développement d'un outil alternatif. *Actes du 25^e colloque de l'ADMÉE-Europe*. Fribourg.
- Sireci, S. G., et Bastari, B. (1998). *Evaluating construct equivalence across adapted tests*. Communication présentée à la réunion annuelle de l'American Psychological Association (APA), San Francisco, CA. August.
- Sireci, S. G., Fitzgerald, C., et Xing, D. (1998). *Adapting credentialing examinations for international uses*. Communication présentée à la réunion annuelle de l'American Educational Research Association (AERA), San Diego, CA. April.
- Sireci, S. G., Foster, D. F., Robin, F., et Olsen, J. (1997). *Comparing dual-language versions of an international computerized-adaptive certification exam*. Communication présentée à la réunion annuelle du National Council on Measurement in Education (NCME), Chicago, Illinois. March.
- Sireci, S. G., et Khaliq, S. N. (2002). *An analysis of the psychometric properties of dual language test forms*. Amherst, MA: School of Education, University of Massachusetts.

- Sireci, S. G., Xing, D., et Fitzgerald, C. (1999). *Evaluating adapted tests across multiple groups: Lessons learned from the IT industry*. Amherst, MA: School of Education, University of Massachusetts.
- Société Française de Psychologie (SFP) (2007). *Recommandations internationales sur les tests informatisés ou les tests distribués par Internet*. Repéré à http://www.sfpsy.org/IMG/pdf/ITC_Guidelines_on_Computer_-_french_version_2007_def.pdf
- (2000). *Recommandations internationales sur l'utilisation des tests. Pratiques psychologiques (numéro spécial hors série)*. Repéré à <http://www.intestcom.org>
- Solano-Flores, G. (2012). *Smarter Balanced Assessment Consortium: Translation Accommodations Framework for Testing English Language Learners in Mathematics*. Repéré à <http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/09/Translation-Accommodations-Framework-for-Testing-ELL-Math.pdf>
- Solano-Flores, G., Backhoff, E., et Contreras-Niño (2009). Theory of Test Translation Error. *International Journal of Testing*. 9.78-91.
- Solano-Flores, G., Trumbull, E., et 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.
- Stansfield, C. W. (1997). *Experiences and Issues Related to the Format of Bilingual Tests: Dual Language Test Booklets versus Two Different Test Booklets*. Massachusetts Department of Education.
- Stern, M. J., Smyth, J. D., et Mendez, J. (2012). The effects of item saliency and question design on measurement error in a self-administered survey. *Field Methods*. 24:1.3-27.
- Swanson, H. L., et Watson, B. L. (1982). *Educational and psychological measurement and evaluation*. Englewood Cliffs, NJ: Prentice-Hall.
- Tassé, M. J., et Craig, E. M. (1979). Critical issues in cross-cultural assessment of adaptive behavior. In R. L. Schalock et D. L. Braddock

- (dir.). *Adaptive Behavior and its Measurement: Implications for the Field of Mental Retardation*. Washington, DC : American Association of Mental Retardation.
- TERMIUM Plus. (2007). *Fiche terminologique du mot rétro-traduction relativement au domaine de la traduction (généralités) (fiche 1)*. Repéré à <http://www.btb.termiumplus.gc.ca/tpv2alpha/alpha-fra.html?lang=fra&i=1&index=frt&srchtxt=RETRO%20TRADUCTION>
- Tomás, I., Hernández, A., et González-Romá, V. (2006). *Evaluating test measurement equivalence across languages using the MACS model*. Communication présentée à la 5^e conférence de l'International Test Commission (ITC), Brussels, Belgium. July 6-8.
- Tourangeau, R., et Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, 103:3.299-314.
- Van de Vijver, F., et Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist*. 1:2.89-99.
- Van De Vijver, F. J. R., et Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda et C. D. Spielberger (dir.). *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*. Mahwah, NJ: Lawrence Erlbaum Associates. 39-63.
- Van de Vijver, F., et Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: an overview. *Revue européenne de psychologie appliquée*. 54.119-135.
- Van de Vijver, F. J. R., et Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*. 13:1.29-37.
- van Widenfelt, B. M., Treffers, P. D. A., de Beurs, E., Siedelink, B. M., et Koudijs, E. (2005). Translation and cross-cultural adaptation of assessment instruments used in psychological research with children and families. *Clinical Child and Family Psychology Review*. 8:2.135-147.

- Wainer, H. (2002). On the automatic generation of test items: Some whens, whys, and hows. In S. H. Irvine et P. C. Kyllonen (dir.). *Item generation for test development*. Mahwah, NJ: Lawrence Erlbaum Associates. 287-305.
- Wänke, M., et Schwarz, N. (1997). Reducing question order effects: The operation of buffer items. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz et D. Trewin (dir.). *Survey measurement and process quality*. New York : John Wiley et Sons, Inc. 115-140.
- Weisz, J. R., Suwanlert, S., Chaityasit, W., Weiss, B., Achenbach, T. M., et Walter, B. A. (1987). Epidemiology of behavioral and emotional problems among Thai and American children: parent reports for ages 6-11. *Journal of the American Academy of Child Psychiatry*. 26.890-898.
- Wooten, A. (2010). *Text Expansion & Contraction in Translation*. Repéré à <http://www.globalization-group.com/edge/2010/05/text-expansion-contraction-in-translation/>

-
- ¹ Bien qu'il soit complexe de définir la notion de *culture*, car c'est un terme polysémique et polymorphe (Chalifoux, 1993), nous pouvons toutefois dire que la culture fait référence à un héritage culturel (valeurs) transmis par la société. Ces valeurs sont socialement apprises et ne sont transmises que par l'individu (Linton, 1968). Nous renvoyons le lecteur à l'article de Chalifoux (1993).
- ² Le terme *adaptation* posséderait un sens plus large qui reflète mieux les différentes activités entourant la préparation d'une épreuve pour différents groupes ethnolinguistiques. En effet, selon le grand dictionnaire terminologique (1982), l'adaptation se définit comme la rédaction de textes autrement que par leur traduction ou par la transposition vers un médium différent. Selon Bastin (1990), l'adaptation sous-entendrait « une traduction soucieuse de la plus grande adéquation possible aux aspirations du lecteur et, partant, concernée par des écarts de forme particulièrement grands qu'impliquent deux réalités sociolinguistiques différentes » (p. 470). En ce sens, la traduction ne constituerait qu'une des étapes impliquées dans un processus d'adaptation.
- ³ Estimés par Wainer (2002) à 1000 \$ par item.
- ⁴ Par groupe ethnolinguistique, et en prenant appui sur la définition de Mounin (2000), nous faisons référence à un groupe d'individus ayant en commun une ou plusieurs langues qui fonctionnent dans des contextes socioculturels donnés.
- ⁵ Terme utilisé par l'auteur pour signifier « dans des contextes culturels et linguistiques différents » (page 191).
- ⁶ Selon le grand dictionnaire terminologique (1979), les normes correspondent à tout type concret ou formule abstraite de ce qui doit être relativement à tout ce qui admet un jugement de valeur. Culturellement, elles correspondraient à l'ensemble des règles collectives ou communes qui servent de guides ou de standards dans l'orientation de l'action.
- ⁷ Comment seront récupérées les copies à la fin de l'épreuve (faut-il les sceller, les poster, etc.)? Qui corrigera les copies? S'agira-t-il d'un seul correcteur pour toutes les copies ou de différents correcteurs? Est-ce que les copies seront corrigées mécaniquement (par lecteur optique pour des

questions à choix multiples par exemple)? Est-ce que les correcteurs se verront offrir une formation, un guide de correction, une grille? Une fois les copies corrigées, comment seront comptabilisés les scores?

- ⁸ Selon le grand dictionnaire terminologique (1988), le terme retraduction est admis pour définir ce processus de validation. Le lecteur notera cependant que « rétro-traduction » pourrait également être utilisé (TERMIUM Plus, 2007). Enfin, le terme « traduction renversée » a aussi été trouvé dans la littérature, mais ne semble pas reconnu par les bases de données terminologiques.