La gestion des spécificités linguistiques et culturelles dans les évaluations internationales de la lecture

Ariane BAYE *

Si la théorie nous dit qu'il n'est pas possible d'obtenir une traduction parfaite, l'expérience nous dit que dans ce monde on traduit. (Umberto Eco)

Cet article propose une analyse rétrospective de la gestion des diversités culturelles et linguistiques dans les évaluations internationales portant sur la compréhension de l'écrit. L'équivalence des supports d'évaluation est en effet gage d'équité pour les pays participants. Mais traiter équitablement les pays engagés dans ce type d'évaluation ne va pas sans tenir compte de la variété de leurs systèmes éducatifs ainsi que de leurs spécificités linguistiques et culturelles [1].

Ce problème se pose avec d'autant plus d'acuité lorsque l'on s'intéresse aux évaluations portant sur la compréhension de la lecture car, même si le consensus sur les objectifs de l'enseignement de la lecture est large, la préparation des tests eux-mêmes reste problématique. C'est que les tâches évaluées sont intimement liées à la langue et aux cultures des pays impliqués. "Portant la trace de l'organisation du monde que notre culture a élaborée, la langue nous situe dans l'univers et dans la société. Car, à l'instar de la race ou de la religion, elle sert volontiers de drapeau aux collectivités humaines et signifie puissamment les appartenances de leurs membres. [...] Faut-il donc s'étonner que les groupes sociaux investissent autant dans la langue et la chargent d'un poids symbolique aussi considérable ?" (Klinkenberg, 1997). Et faut-il s'étonner que l'utilisation d'un matériel de test standardisé pour évaluer les compétences en lecture d'individus d'origines linguistiques et culturelles variées ait souvent fait l'objet de suspicions (par exemple, Blum et Guérin-Pace, 2000, Romainville, 2002) ou de remises en question (Bonnet *et al.*, 2001, 2003) ? Les critiques sur la comparabilité des évaluations internationales se sont d'ailleurs aiguisées à

*Service de Pédagogie théorique et expérimentale, Université de Liège. Email : ariane.baye@ulg.ac.be

[1] L'impact du format des questions est discuté par Messieurs Monseur et Demeuse dans ce numéro.

mesure que la visibilité de ces évaluations, et leur influence sur les politiques éducatives, augmentaient (Baye, 2001).

Outre le constat de la difficulté de construire et de traduire des instruments valides et fiables pour évaluer la compréhension de l'écrit – les concepteurs des premières évaluations internationales ont d'emblée posé ces constats (Thorndike, 1973 ; Purves, 1973) – les promoteurs d'évaluations "culturellement non biaisées" (Bonnet *et al.*, 2001, 2003) adressent aux évaluations internationales "traditionnelles" certaines critiques plus précises en matière de gestion des diversités linguistiques et culturelles. Au niveau de la traduction, ils dénoncent l'usage prédominant de traductions à partir de l'anglais, le manque de références aux théories de la traduction et de la linguistique ainsi que le recours à des traductions réalisées par des agences généralistes pour les besoins d'une enquête. Les auteurs distinguent ce type de traduction "artificielle" et "décontextualisée" de la traduction "naturelle", effectuée par un professionnel spécialisé dans un genre particulier, et qui correspond à des livres que l'on peut trouver dans des librairies ou des bibliothèques. Ils reprochent également, tant en ce qui concerne la vérification des traductions que la production des tests, une gestion centralisée et verticale des processus. Enfin, ils pointent la difficulté de contrôler correctement l'influence que des épreuves traduites peuvent avoir sur les résultats des élèves : "it is also believed that a fair interpretation of comparability should put all pupils in the same situation ; i.e. all students should have the possibility of reading and interpreting a text originally written in their own mother tongue to reduce the influence of variables which cannot be properly controlled, such as influence of translation on performance" (Bonnet *et al.*, 2003, 12).
Ces critiques ont été mises à l'épreuve d'un examen des procédures appliquées dans différentes évaluations internationales pour assurer l'équivalence linguistique et culturelle du matériel de test. Nous envisagerons d'abord les procédures de traduction, puis la provenance des supports d'évaluation, pour conclure par une approche quantitative qui permettra d'aborder la question de l'impact de tests traduits sur les performances des élèves.

LES PROCÉDURES DE TRADUCTION

Sans verser dans un radicalisme linguistique qui voudrait qu'une langue soit intrinsèquement liée à une culture particulière, et que ses richesses et particularités ne puissent être traduites d'une communauté à une autre (Von Humboldt, 2000), il ne faut toutefois pas sous-estimer le problème de la traduction dans les évaluations internationales.
En effet, certaines difficultés de traduction/adaptation ont un impact direct sur l'équivalence des tests. Par exemple, les traducteurs/adaptateurs doivent

produire dans la langue cible[2] des textes comparables au niveau de la longueur, du degré d'abstraction et de la fréquence d'utilisation, moyens du lexique, car ces paramètres influencent leur degré de lisibilité[3]. À cela s'ajoute le problème de la polysémie : un mot charrie parfois un réseau de significations différentes. Un auteur peut utiliser à dessein un terme polysémique – particulièrement dans des textes à caractère littéraire. À l'inverse, il peut "cadenasser" le sens en veillant à ce qu'il n'y ait qu'une interprétation possible. Au traducteur de veiller à ce que son travail ne rétrécisse ou n'élargisse à l'excès les réseaux de significations et d'interprétations possibles. Dans le premier cas, le lecteur serait trop facilement encouragé dans une voie à sens unique. Dans le second, le lecteur ne disposerait pas d'une signalisation adéquate au "carrefour du sens". Le traducteur doit également être attentif à la connotation que peut prendre un mot dans un contexte précis, alors que ce mot n'aura pas la même acception dans un autre contexte linguistique. Enfin, certains termes n'ont pas d'équivalent dans une autre langue, tandis que d'autres en ont plusieurs[4].

Chaque langue a ses caractéristiques propres : il n'est pas gênant de répéter un mot dans une même phrase en anglais, alors qu'en français on préférera utiliser des synonymes. Cette caractéristique linguistique peut se révéler particulièrement embarrassante pour des questions basées sur la correspondance littérale. Les spécificités syntaxiques de certaines langues peuvent également poser des problèmes. Ainsi, certains formats de questions sont peu adaptés aux langues qui rejettent les verbes après une subordonnée ou qui emploient des verbes à particules séparables. Enfin, certaines langues tendent à plus de concision que d'autres. La relative économie des moyens d'expression peut compromettre la comparabilité formelle de deux versions. Il faudra parfois faire preuve de beaucoup d'ingéniosité pour conserver la forme de la version originale, et éviter que des textes ou graphiques ne "débordent" par rapport à l'espace prévu, obligeant par exemple l'élève à tourner la page, alors que, dans la version originale, textes et questions se trouvent vis-à-vis. Ces écueils évités, la traduction devra encore se surpasser en s'effaçant elle-même, elle ne devra pas "sentir" la traduction. On pourra alors parler de matériel de test "équilisible",

---

[2] Il peut s'agir d'une même langue parlée dans un autre pays. On peut par exemple distinguer l'anglais des Etats-Unis, du Royaume-Uni ou d'Australie. Pour prendre un exemple francophone, un "dépanneur" n'est pas associé à la même réalité pour un Belge francophone ou un Québécois. Pour ce même Québécois, il conviendra de traduire "fast food" par "restaurant à service rapide", traduction qui sera bien moins claire pour un Français que l'original anglais !

[3] Les mots longs tendent à être moins fréquents, plus techniques et/ou plus abstraits que les mots brefs. Le vocabulaire de base d'une langue est, le plus souvent, composé de mots très courts.

[4] Les exemples de potache ne manquent pas : le malais n'a qu'un mot pour désigner "frère" et "sœur" ; par contre, le hongrois en possède quatre, suivant qu'il s'agit des aînés ou des cadets... et le coréen utilise neuf formes différentes… Le "mouton" francophone deviendra "mutton" ou "sheep" en anglais selon qu'il garnisse votre assiette ou qu'il paisse.

dans le sens d'également compréhensible, interprétable et recevable par différents groupes linguistiques.
L'examen des procédures de traduction exploitées dans les évaluations internationales de la lecture indique que, contrairement à certaines idées reçues, les responsables des évaluations internationales menées sous l'égide de l'Association internationale pour l'évaluation des rendements scolaires (IEA) ou par l'Organisation de coopération et de développement économiques (OCDE) ont toujours été très sensibles à l'adaptation linguistique et culturelle des tests. En effet, dès le début, il a fallu convaincre les sceptiques de la comparabilité des résultats et pour cela, résoudre notamment "the formidable translation problem" (Husén, dans Thorndike, 1973 :11).
En effet, même si, en plus de trente ans d'évaluations internationales de la lecture, des progrès en matière d'outillage statistique ont permis un contrôle de plus en plus précis de l'équivalence des tests, on trouve dès 1971, à l'occasion de l'étude de l'IEA "Reading Comprehension", une description des contrôles empiriques effectués pour valider les instruments. Ces contrôles avaient inclus l'examen de la fidélité du test, l'analyse de la stabilité relative des indices de difficulté et de discrimination des items d'un pays à l'autre, ainsi que des analyses factorielles des indices de difficulté et de discrimination (Thorndike, 1973).

En amont, au niveau de la conception des instruments, les procédures de gestion et de contrôle de la traduction se sont également affinées. Après les deux premières études fondatrices de 1971 ("Reading Comprehension" et "Reading Literature"), l'IEA a recommandé que chaque version nationale du test soit traduite par deux traducteurs indépendants, et conciliées par un tiers (Elley, 1994)[5]. Cette recommandation s'est commuée en prescrit assorti d'un contrôle externe dans le cycle 2000 du Programme International pour le Suivi des Acquis des élèves (PISA). PISA a par ailleurs introduit une réelle innovation, à mettre en relation avec le commanditaire de l'étude : l'OCDE étant une organisation bilingue, le matériel de test est fourni en anglais et en français à tous les pays. On recommande aux traducteurs d'utiliser les deux versions, ce qui augmente la finesse de la compréhension du matériel et permet de résoudre certaines ambiguïtés[6]. Les analyses comparant le pourcentage d'items problématiques en fonction de la méthode de traduction utilisée dans les différents pays indiquent que le recours aux deux versions sources (soit via

[5] Il semble que cette recommandation n'est pas été largement suivie dans IEA "Reading Literacy". Dans PIRLS, une vérification externe a été prévue (Martin, Mullis, Kennedy, 2003).

[6] Ce système a aussi ses limites. La version "originale" est en anglais. Certains textes ou items apportés par les pays participants sont donc traduits une première fois vers l'anglais. Cette version "source" est traduite vers le français selon la procédure générale valable pour toutes les langues (deux traductions indépendantes, conciliation par un tiers, vérification externe de la qualité de la traduction). Les traducteurs qui partiraient uniquement de la version "source" en français auraient donc une version deux (voire trois) fois traduite. Ce cas de figure augmenterait potentiellement le nombre d'altérations liées au nombre de traductions.

une double traduction à partir des deux versions sources, soit via une double traduction à partir d'une des versions sources avec vérifications approfondies à l'aide de l'autre version source) s'est révélé plus efficace que les autres méthodes, et notamment, que la double traduction à partir d'une seule version source (Grisay, 2002).

Des dispositifs de plus en plus complexes et contraignants ont donc accompagné l'attention initiale à l'équivalence linguistique des épreuves. Ce processus s'est certes assorti d'une centralisation accrue. Mais nous allons voir que c'est plus au profit d'une vérification finale externe, de nature à garantir une égale attention aux productions nationales et une plus grande comparabilité des épreuves, qu'au préjudice de la responsabilité locale. En effet, dans toutes les études considérées, les responsables nationaux ont été sollicités pour fournir des textes. La traduction du matériel de test a, elle aussi, toujours été confiée aux centres nationaux. Et c'est précisément cette gestion décentralisée qui a été à l'origine de quelques incidents critiques qui ont amené à un contrôle externe plus strict du respect des procédures.

Ainsi, dans l'étude de l'IEA "Reading Comprehension" (1971), la gestion décentralisée de la traduction s'est heurtée à une limité de faisabilité : la quantité de matériel à traduire (50 textes, 445 items) était trop importante pour pouvoir être traitée dans les délais par chaque pays participant en vue du pré-test (Thorndike, 1973)[7].
Dans l'étude de l'IEA "Reading Literature", menée en parallèle et selon les mêmes procédures, les problèmes détectés grâce aux analyses du pré-test, particulièrement au niveau des items à choix multiples, conduisent le Comité international à augmenter les contraintes en vue du test principal. On trouve ainsi dans le rapport final de l'étude les premières traces écrites du recours à des méthodes de traduction professionnelles dans les évaluations internationales de la lecture. La "professionnalisation" se marque de deux façons. D'une part, on recommande aux Comités nationaux de faire appel aux services de traducteurs littéraires pour vérifier les textes avant le test définitif (les questionnaires d'attitude par rapport à la littérature seront cependant exclus de cette vérification). D'autre part, le matériel de test traduit dans les différentes langues nationales est retraduit vers l'anglais (*back translation)*, la comparaison de l'original et de la version retraduite permettant de détecter certaines erreurs de traduction (Purves, 1973). Ce mode de vérification par *back translation* sera également exploité dans l'étude de l'IEA "Reading Literacy" (1991), mais sera

[7] Par la suite, lors des essais de terrain, les responsables internationaux essayeront d'obtenir plus d'informations sur la comparabilité linguistique des items en les testant dans quatre langues différentes, et en essayant que chaque passage soit soumis à des échantillons de 200 à 300 élèves dans chaque pays participant (Thorndike, 1973).

abandonné par la suite, car il ne permet pas, par exemple, de détecter les traductions trop littérales[8].

Lorsque l'on évoque les problèmes de comparabilité des évaluations internationales, l'étude "International Adult Literacy Survey" (IALS, menée par Statistiques Canada et l'OCDE de 1994 à 1998) est souvent pointée du doigt (Blum et Guérin-Pace, 2000). Or, il semble qu'ici aussi, l'étude ait moins pêché par manque de vigilance en matière de procédures de traduction – vigilance d'autant plus cruciale que le format des réponses (ouvertes) appelait une grande qualité des guides de correction – que par une trop grande autonomie des pays par rapport aux procédures prévues. En effet, certains représentants nationaux *"chose not to incorporate a number of changes which were identified during the course of the review, believing that they "knew better"*(Murray, Kirsch, Jenkins, 1998, 77). Cette trop grande latitude est d'ailleurs dénoncée par les experts indépendants engagés pour vérifier le bien-fondé des critiques françaises concernant la qualité méthodologique de l'étude. Les experts notent par ailleurs la bonne qualité psychométrique des instruments de test. Ils trouvent cependant quelques problèmes de comparabilité concernant la mise en forme des documents et le non-respect des correspondances littérales entre des énoncés de question et les textes auxquels ils se rapportent (Kalton, Lyberg, Rempp, 1998).

Même si ces divergences ne sont pas, selon ces experts, de nature à affecter les résultats de l'étude, cet élément nous permet d'aborder un autre aspect contesté par d'aucuns : le recours à des textes traduits pour les besoins d'une enquête, là où des traductions réalisées pour un autre usage et disponibles dans le commerce sont considérées comme plus naturelles, et *de facto* mieux adaptées aux contextes culturels dans lesquels elles sont disponibles (Bonnet *et al.*, 2003). On peut opposer à cet argument qu'il convient surtout à des textes littéraires dont les plus "classiques" sont de fait disponibles dans de nombreux pays et dans de nombreuses traductions. Or, après les deux études fondatrices de l'IEA, où le champ de l'évaluation de l'écrit était clivé selon une dichotomie cognitif *versus* affectif, qui correspondait aussi à une différenciation au niveau des supports d'évaluation (informatifs *versus* littéraires), les évaluations internationales ont cherché à mesurer la littératie, "qui associe la compréhension et les usages sociaux de la lecture [...], et qui s'accompagne d'une entrée en force des documents comme supports de l'évaluation [...]" (Lafontaine, 2001, 73). Il semble difficile de trouver une grande variété de textes, et par exemple des articles de presse, adaptés à un public en âge scolaire qui soient diffusés et traduits internationalement... On pourrait chercher du côté des modes d'emplois d'appareils électroniques ou de téléphones portables,

[8] Pour prendre un exemple grossier, si un traducteur traduit par "Il pleut des chiens et des chats" l'expression anglaise "It's raining dogs and cats", le responsable de la *back translation* retraduira sans problème l'erreur française dans une version anglaise identique à l'original, la *back translation* n'aura donc pas permis de déceler la mauvaise qualité de la traduction française.

mais l'expérience quotidienne nous apprend que ces textes, que l'on trouve en versions traduites dans de nombreux pays et contextes linguistiques, requièrent du lecteur un certain degré d'inférence – voire un bon niveau de connaissance des langues étrangères – pour être compris, tant la qualité de leurs traductions commerciales "naturelles" laisse à désirer.

En outre, le recours à des traductions "naturelles" risque de poser plus de problèmes de comparabilité que des traductions spécifiquement réalisées pour une évaluation internationale, car les premières ne sont pas adaptées au contexte précis de leur réception. En effet, dans le cas d'évaluation de la compréhension de l'écrit, il est fondamental que les élèves, quel que soit leur environnement linguistique, soient soumis à des épreuves d'un niveau de difficulté équivalent. Cela implique, nous l'avons vu, beaucoup de précision, ne fût-ce que pour le choix du lexique utilisé (niveau de difficulté, correspondance littérale, longueur des mots, etc.). Même un traducteur professionnel expérimenté pourrait ne pas être assez attentif à tous ces aspects s'il n'est pas informé du contexte et du lectorat particulier auquel il s'adresse, et des qualités à l'aune desquelles sera apprécié son travail. C'est pourquoi toutes les enquêtes examinées ont été attentives à ce que la traduction soit adaptée à ce contexte particulier de réception, en fournissant des guides de traduction précisant notamment les buts de l'évaluation et les pièges à éviter dans ce type de traduction.

L'ORIGINE DES TEXTES

Une des critiques adressées aux évaluations internationales de la compréhension de l'écrit concerne la prédominance d'un matériel original en anglais.

Pour certains, la comparabilité des textes n'est pas seulement une affaire linguistique. Elle est intimement liée à l'origine culturelle des textes. Jocelyne Giasson évoque les travaux de Steffensen *et al.* pour expliquer que "les sujets lisent plus vite les textes portant sur leur propre culture et en retiennent mieux l'information. Ils font également plus d'erreurs d'interprétation dans les passages concernant une autre culture" (1990, 170). L'auteure évoque un texte soumis à des étudiants américains et indiens. Dans ce texte, il est question d'un mariage américain où la mariée porte la robe de sa grand-mère, robe dans laquelle la jeune mariée est charmante. Un lecteur indien conclut que la mariée était charmante, mais que sa robe est démodée : dans la culture indienne, un mariage est l'occasion d'investir dans des vêtements dernier cri. Peut-on dire que cet Indien a mal lu ? N'est-il pas concevable que certains textes soient impossibles à comprendre, en tous cas finement, car produits par et pour une culture donnée ? Warwick B. Elley (1994), responsable d'IEA "Reading Literacy", reprend cet exemple, mais conclut que les diversités culturelles peuvent être dépassées au nom des expériences communes des élèves : ils sont

généralement élevés dans des familles, par des adultes avec lesquelles ils communiquent. Ils ont souvent accès aux mêmes biens et services.

On le voit, il n'est pas aisé de trancher entre la possibilité d'une "lisibilité interculturelle" et l'"illisibilité pour cause culturelle". Peut-être ces débats reflètent-ils la persistance d'une vision holistique de la relation entre langue et culture et peuple, oubliant que la langue est un outil qui permet précisément l'appropriation d'univers variés, lointains ou inventés. Les évaluations internationales de la lecture examinées cautionnent en tous cas la première approche, puisqu'elles soumettent à des personnes d'origines linguistiques et culturelles variées des supports d'évaluation communs. Ont-elles ménagé les diverses sensibilités linguistiques et culturelles dans le choix des textes utilisés ?
Concernant l'adéquation culturelle, un premier filtrage est opéré par les comités de sélection des textes qui éliminent *a priori* les supports manifestement plus accessibles aux sujets de certaines cultures. Le matériel de test est également soumis à l'appréciation des responsables nationaux. À l'occasion de PISA 2000, un comité international d'experts non impliqués à d'autres niveaux de l'étude a été chargé d'examiner la convenance culturelle du matériel de test[9].

La question de la diversité linguistique sera abordée via l'analyse de la variété des pays et des langues d'origine des textes proposés aux répondants.
Malheureusement, l'origine des textes n'est pas toujours précisée dans les rapports internationaux. Pour l'étude de l'IEA "Reading Comprehension", "each National Center was invited to contribute reading passages to a pool from which an eventual selection might be made, and *a number did in fact do so* [10]" (Thorndike, 1973, 20). Pour l'étude de l'IEA "Reading Literacy", l'origine nationale des supports d'évaluation n'est pas mentionnée non plus. Il est intéressant de noter que l'origine des textes est indiquée dès 1971 pour l'étude de l'IEA "Reading Literature". Les concepteurs de l'étude étaient en effet particulièrement sensibles aux spécificités culturelles nationales, puisqu'une partie de l'étude visait à mettre en évidence les différences nationales dans la manière d'aborder l'analyse littéraire. Pour cette étude, quatre textes ont été sélectionnés. Deux auteurs sont américains, le troisième est espagnol, le dernier est belge. Vu le petit nombre de textes prévus pour le test final, cette sélection témoigne d'une relative diversité dans le choix des langues et des pays représentés.
Pour IALS et PISA, nous disposons également d'informations sur l'origine nationale et/ou linguistique des textes.

---

[9] Le travail du panel d'experts a cependant été contraint par l'agenda de l'étude : deux jours de réunion, pour lequel les documents de travail avaient été envoyés très peu de temps à l'avance (Hambleton, 1999). En outre, les avis du panel culturel n'ont pas pu être exhaustivement pris en compte, car la suppression de certains textes jugés problématiques aurait compromis l'économie générale du test en matière de type d'items présentés aux élèves.

[10] Nous soulignons.

Tableau 1. Origine linguistique des textes pour l'évaluation de la compréhension de l'écrit des tests définitifs de IALS et PISA 2000

IALS			PISA 2000		
Langue	***Nombre de textes***	**Proportion**	**Langue**	***Nombre de textes***	**Proportion**
Allemand	4	10 %	Allemand	1	3 %
Anglais	19	47 %	Anglais	17	46 %
Espagnol	2	5 %	Danois	1	3 %
Français	2	5 %	Espagnol	1	3 %
Néerlandais	9	23 %	Finnois	3	8 %
Non identifiée[11]	4	10 %	Français	4	11 %
Total	39	100 %	Grec	1	3 %
			Suédois	1	3 %
			Non identifiée[12]	8	22 %
			Total	37	100 %

Sources : IALS : National Center for Education Statistics, 1998 ; PISA 2000, Adams et Wu, 2002.

Malgré la relative diversité linguistique[13], la prédominance de l'anglais est manifeste dans les deux cas. Cette constatation mérite cependant d'être nuancée par la prise en compte des variations entre langues employées dans des contextes géographiques, sociaux et culturels différenciés, tant il est simpliste de considérer le français, l'anglais ou l'espagnol comme des systèmes monolithiques[14]. Soulignons aussi qu'une origine nationale et linguistique donnée n'implique pas un ancrage culturel marqué au niveau du contenu des textes. Ainsi, dans PISA 2000, un texte suédois évoque une réserve de rhinocéros au Kenya, une épreuve néo-zélandaise est fondée sur les instructions pour l'usage des préfixes téléphoniques dans un hôtel turc, un texte anglais parle de gravures rupestres en Afrique sub-saharienne. Ces quelques exemples – la liste n'est pas exhaustive – indiquent bien que l'impératif qui guide tant les responsables nationaux qui fournissent des textes que le Comité international de sélection, est l'exigence d'universalité des contenus.

Au niveau de la couverture des pays participants, on note que dans IALS, tous sont représentés dans les textes proposés aux répondants, à l'exception de la Suède qui a rejoint trop tard les autres pays (National Center for Education Statistics, 1998). Pour IEA "Reading Literacy", vingt des trente-deux représentants nationaux ont proposé des supports d'évaluation (Elley, 1992). Pour PISA, dix-huit des trente-deux pays participants ont soumis des textes (McQueen et Mendelovits, 2003). Dans l'étude de l'IEA "PIRLS", quatorze des vingt-cinq pays participants ont fourni des textes, et huit

[11] Textes canadiens en anglais ou en français.
[12] Textes repris de IALS dont l'origine nationale/linguistique n'a pu être identifiée.
[13] Pour PISA 2000, treize langues étaient représentées dans les supports d'évaluation de la littératie pour le pré-test. Les textes norvégien, russe, japonais, ainsi que les trois textes coréens n'ont pas été retenus pour l'épreuve définitive.
[14] Dans IALS et PISA 2000, les textes anglais proviennent de divers continents.

pays sont finalement représentés dans le choix des stimuli du pré-test (Martin, Mullis, Kennedy, 2003)[15].

En fin de compte, même si la diversité des origines culturelle et linguistique des stimuli n'a pas encore fait l'objet d'un calibrage minutieux dans la construction des épreuves, il semble que ce soit moins en raison d'un manque d'intérêt pour ces paramètres qu'en raison du nombre d'autres critères à prendre en compte pour assurer la couverture des domaines et des processus évalués. Pour sélectionner les supports d'évaluation, il faut concilier les avis des experts de la lecture et des responsables nationaux, tout en prenant en compte la diversité des types de textes évalués et la qualité psychométrique des items. S'il est essentiel d'inclure du matériel provenant de divers horizons linguistiques et culturels, il faut alors que les responsables internationaux aient à leur disposition des textes variés envoyés par tous les pays participants. Si, dans la hiérarchie des priorités internationales, la diversité linguistique et culturelle peut apparaître comme un parent pauvre, c'est peut-être que ces considérations, et l'investissement qu'elles impliquent au niveau national, n'ont pas encore trouvé assez de relais dans tous les pays participants.

L'IMPACT DES TESTS TRADUITS SUR LES PERFORMANCES DES ÉLÈVES[16]

La difficulté de contrôler correctement l'influence que des épreuves traduites peuvent avoir sur les résultats des élèves (Bonnet *et al.*, 2003) est réelle. L'analyse du fonctionnement différentiel des items permet bien de repérer et de supprimer des versions finales des épreuves les items "atypiques", dont les mauvaises qualités psychométriques peuvent être le résultat de problèmes de traduction ou d'adaptation culturelle, mais pas d'expliciter ou de quantifier les causes des phénomènes observés, ni de détecter une source de biais qui agirait uniformément sur l'ensemble des items. Une analyse en *clusters* menée par Lie et Roe (2003) à partir de la version définitive du test PISA 2000 fait apparaître des regroupements entre pays partageant, au moins en partie, une même langue – en l'occurrence, l'anglais ou l'allemand. Répliquée sur la base des différentes versions linguistiques nationales du test, et plus seulement de l'unité " pays"[17], l'analyse en

[15] Le nombre de pays participants peut être supérieur au nombre de pays présents au moment de fournir du matériel de test.

[16] Nous remercions Christian Monseur et Marc Demeuse pour leurs conseils précieux pour l'analyse des données de PISA 2000.

[17] Excepté pour le Luxembourg, pour lequel nous n'avons pas les données pour les différentes versions linguistiques utilisées.

clusters[18] présentée ci-dessous accentue les regroupements linguistiques, faisant émerger un sous-groupe anglophone, germanophone, mais aussi francophone, néerlandophone et italophone.

Plusieurs hypothèses pourraient expliquer ces regroupements : à côté de la parenté linguistique de certains pays, on peut mettre en avant le fait qu'un grand nombre d'entre eux a travaillé sur la base de versions communes (version francophone et anglophone internationales, mais également version germanophone, italophone et néerlandophone communes). Il serait également intéressant de creuser l'hypothèse de proximités pédagogiques entre pays qui partagent une histoire commune. D'autre part, on remarque que certains pays partageant des caractéristiques linguistiques (Mexique, Brésil) ou culturelles (Suède) avec d'autres pays ne se regroupent pas sur ces bases avec ces derniers. Les huit pays (de la Suède à la Corée) qui semblent "rejetés" par l'analyse en *clusters* sont par ailleurs particulièrement performants ou contre performants sur l'échelle combinée de compréhension de l'écrit. Comme l'avait déjà montré Thorndike (1973), les qualités psychométriques des items sont plus instables dans les pays où l'épreuve est sensiblement mieux ou moins bien réussie.

Mais que représentent les convergences et divergences mises en avant par l'analyse en *clusters* par rapport à l'ensemble des ressemblances et des différences entre les différentes versions du test ? Autrement dit, quelle est la part de la variance totale attribuable aux différences linguistiques entre groupes de pays suggérées par l'analyse en *clusters* ? Pour tenter de répondre à cette question, une analyse factorielle en composantes principales a été effectuée[19].

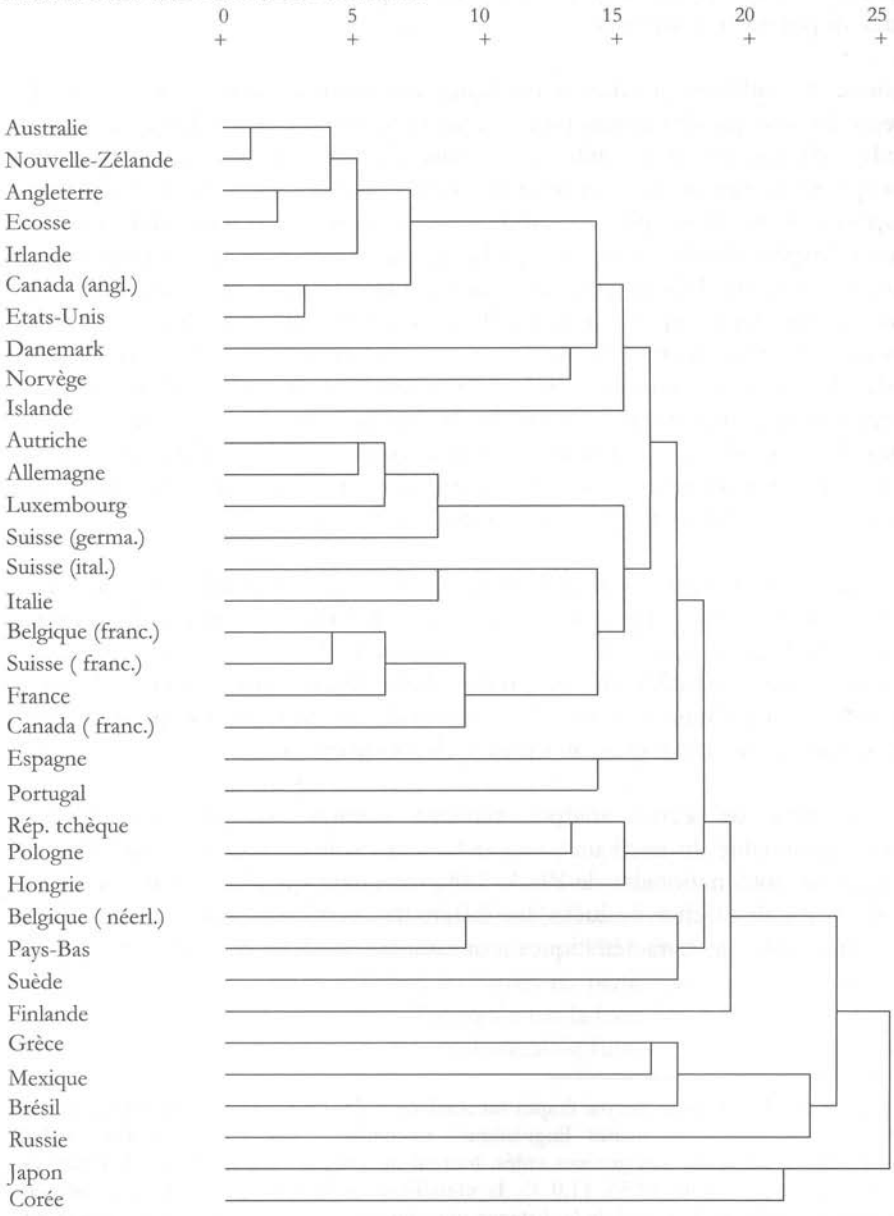
Les résultats de cette analyse rendent compte de l'impressionnante unidimensionnalité du test : un premier facteur explique 83,7 % de la variance entre les versions nationales de PISA 2000, indiquant que, lorsqu'on s'intéresse à la difficulté des tâches évaluées, les différentes versions du test partagent un vaste ensemble de caractéristiques communes, au-delà des langues et des cultures.

---

[18] L'analyse en *clusters* procède, par étapes successives, à des regroupements de variables (ici les pays ou les communautés linguistiques nationales). Elle vise à maximiser les similitudes à l'intérieur des groupes créés, tout en maximisant également les divergences entre les groupes. Sous SPSS 11.0 ©, la classification hiérarchique a été effectuée en utilisant la méthode du calcul de la distance moyenne entre classes. La mesure utilisée est le carré de la distance euclidienne.

[19] L'analyse factorielle en composantes principales est une technique statistique permettant de réduire un système complexe de corrélations en un plus petit nombre de dimensions. Ces dimensions ou facteurs, sont générées de manière à maximiser la part de variance totale expliquée. Sous SPSS 11.0 ©, l'analyse factorielle en composantes principales a été effectuée en retenant les facteurs dont la valeur propre est supérieure à 1, afin que chaque facteur dégagé explique plus de variance qu'une seule des variables de départ, soit une des versions nationales de PISA 2000.

Figure 1. Dendrogramme représentant l'analyse en *clusters* sur la base des versions nationales du test PISA 2000 de littératie



Données : PISA 2000. Deltas centrés pour les items de littératie [20]

[20] La base de données est constituée des indices de la difficulté des items de littératie estimée par le modèle de Rasch à un paramètre. Ces indices sont calculés pour chacun des 136 items de littératie, et pour chacune des 35 versions nationales de PISA 2000.

Peut-on néanmoins dégager, dans la variance inexpliquée à ce stade, d'autres facteurs communs de variation ? L'analyse factorielle n'extrait plus qu'un second facteur, expliquant seulement 2,9 % de la variance[21] Examinons maintenant la corrélation entre ces deux facteurs et les différentes versions nationales du test, afin de mieux comprendre ce que recouvrent ces facteurs, et notamment si le second peut être qualifié de "linguistique" ou "culturel".

Comme l'indique le tableau 2, toutes les versions nationales ont fortement contribué à la création du premier facteur, puisque ce dernier sature partout à plus de 0.80, et à 0.90 ou plus dans 29 des 35 versions du test considérées. Les quatre pays pour lesquels la corrélation est un peu moins forte (Corée, Japon, Mexique et Brésil) appartiennent, pris deux à deux, à des espaces géographiques et culturels proches – et peut-être spécifiques par rapport à un ensemble "occidental"– et se distinguent positivement ou négativement par leurs résultats sur l'échelle de littératie. Le second facteur est beaucoup plus difficilement interprétable. Les quatre pays qui ont le plus contribué à sa constitution (Mexique, Brésil, Grèce, Fédération de Russie) étaient "rejetés" par l'analyse en *clusters* et sont des pays où l'ensemble du test a été moins bien réussi que dans la plupart des autres pays. Mis à part cela, il est plus aisé de trouver une caractéristique commune aux pays qui n'ont pas contribué à la constitution de ce facteur – le groupe francophone – que de trouver des qualificatifs appropriés pour les pays où la corrélation est légèrement positive (pays "latins" et "de l'Est") ou négative (pays anglophones, scandinaves et germaniques).

Tableau 2. Saturation des versions nationales de l'épreuve de littératie PISA 2000 sur les deux facteurs extraits de l'analyse en composantes principales

	Facteur 1 Explique 83,7 % de la variance		Facteur 2 Explique 2,9 % de la variance
	Saturation sur le facteur 1		Saturation sur le facteur 2
Belgique (fran.)	0.96	Mexique	0.46
Suisse (fran.)	0.95	Brésil	0.41
Australie	0.95	Grèce	0.29
Canada (angl.)	0.95	Russie	0.27
Luxembourg	0.95	Portugal	0.22
Irlande	0.95	Italie	0.17
Nouvelle-Zélande	0.95	Pologne	0.16
Etats-Unis	0.94	Rép. tchèque	0.11

21 Si l'on examine les facteurs dont la valeur propre est inférieure à 1 – ce qui n'a qu'un sens heuristique, car ces facteurs expliquent moins de variance que chacune des variables introduites dans l'analyse – on trouve un "facteur 3", expliquant 2 % de la variance, auquel un groupe de pays germanophones est corrélé positivement et un groupe de pays anglophones est corrélé négativement, et un "facteur", expliquant 1,5 % de la variance, à la création duquel ont fortement contribué la Corée et le Japon.

Canada (fran.)	0.94	Suisse (ital.)	0.11
Italie	0.94	Corée	0.11
Islande	0.94	Hongrie	0.07
Ecosse	0.94	France	0.05
Suisse (ital.)	0.94	Espagne	0.04
Angleterre	0.94	Canada (fran.)	0.01
Portugal	0.93	Belgique (fran.)	0.00
France	0.93	Islande	-0.01
Espagne	0.93	Suisse (fran.)	-0.02
Allemagne	0.93	Luxembourg	-0.07
Pologne	0.92	Autriche	-0.08
Hongrie	0.92	Suisse (germa.)	-0.08
Norvège	0.92	Norvège	-0.09
Autriche	0.92	Allemagne	-0.09
Suisse (germa.)	0.92	Canada (angl.)	-0.10
Danemark	0.92	Etats-Unis	-0.10
Belgique (néerl.)	0.92	Danemark	-0.10
Suède	0.91	Belgique (néerl.)	-0.10
Pays-Bas	0.90	Finlande	-0.11
Rép. tchèque	0.90	Japon	-0.13
Finlande	0.90	Irlande	-0.14
Russie	0.86	Pays-Bas	-0.16
Grèce	0.85	Nouvelle-Zélande	-0.19
Corée	0.83	Suède	-0.19
Brésil	0.82	Ecosse	-0.20
Japon	0.82	Australie	-0.20
Mexique	0.81	Angleterre	-0.21

Données : PISA 2000. Deltas centrés pour les items de littératie.

DES ARBRES... ET DES FORÊTS

Si l'examen minutieux de la comparabilité linguistique et culturelle des évaluations internationales est légitime, non seulement d'un point de vue scientifique, mais aussi d'un point de vue politique, en raison de l'impact de leurs résultats, l'examen rétrospectif des procédures de gestion et de contrôle mises en place en trente ans d'évaluations internationales de la lecture indique que ces aspects n'ont jamais été négligés. Des progrès importants ont néanmoins été réalisés, tant au niveau du contrôle des processus que de l'examen des qualités psychométriques des épreuves, garantissant toujours plus la comparabilité des résultats.

A ce titre, les analyses menées sur les données de PISA 2000 ne permettent pas de confirmer l'hypothèse d'un impact significatif de facteurs géographiques, linguistiques ou culturels sur le facteur mesuré par le test, la compétence en lecture. Une fois la comparabilité linguistique et culturelle des épreuves analysée – et nous avons montré, comme l'avait déjà fait Thorndike en 1973, combien sur ce point ce qui est commun à l'ensemble des pays dépasse largement les différences entre eux – il nous semble urgent de se centrer sur les différences de performances entre les systèmes éducatifs, tant en termes d'efficacité que d'équité. De ce point de vue, il reste tant à défricher.

RÉFÉRENCES

Adams, R. & Wu, M. (Eds) (2002) *PISA 2000 Technical Report.* Paris : OECD.

Baye, A. (2001) Evaluations internationales : les enjeux du jour et de l'histoire. *Les Cahiers du Service de Pédagogie expérimentale*, 7-8, 11-23.

Blum, A. & Guérin-Pace, F. (2000) *Des lettres et des chiffres. Des tests d'intelligence à l'évaluation du "savoir lire", un siècle de polémiques.* Paris : Fayard.

Bonnet, G., Braxmeyer, N., Horner, S., Lappalainen, H-P., Levasseur, J., Nardi, E., Rémond, M., Vrignaud, P., White, J. (2001) *The use of national reading tests for international comparisons : ways of overcoming cultural bias.* Paris : Ministère de l'Education nationale. Direction de la Programmation et du Développement.

Bonnet, G., Daems, F., Gloppper, C., Horner, S., Lappalainen, H-P., Nardi, E., Rémond, M., Robin, I., Rossen, M., Solheim, R., Tonnessen, F-E., Vertecchi, B., Vrignaud, P., Wagner, A.K. (2003) *Culturally balanced assessment of reading (c-bar).* http://cisad.adc.education.fr/reva

Eco, U. (2003) L'expérience de la traduction. In Centre Roland Barthes. *Le plaisir des formes.* Paris : Le Seuil, 113.

Elley, W.B. (1992) *How in the World do Students Read ?* Hamburg : IEA.

Elley, W.B. (Ed.) (1994) *The IEA Study of Reading Literacy : Achivement and Instructions in Thirty-Two School Systems.* Oxford : Pergamon.

Giasson, J. (1990) *La comprehension de l'écrit.* Bruxelles : De Boeck.

Grisay, A. (2002) Translation and cultural appropriateness of the test and survey material. In Adams, R. & Wu, M. (Eds), *PISA 2000 Technical Report.* Paris : OECD, 57-70.

Hambleton, R. K. (1999) *Summary of the main points of the cultural review panel meeting.* Document non publié.

Kalton, G., Lyberg, L., Rempp, J.-M. (1998) Review of Methodology. In National Center for Education Statistics. *Adult Literacy in OECD Countries. Technical Report on the First International Adult Literacy Survey.* Washington D.C.: US Department of Education, Annexe A.

Klinkenberg, J-M. (1997) Pour une politique de la langue française. *La revue nouvelle.* Numéro spécial.

Lafontaine, D. (2001) Quoi de neuf en littératie ? Regard sur trente ans d'évaluation de la lecture. *Les Cahiers du Service de Pédagogie expérimentale*, 7-8, 71-90.

Lie, S. & Roe, A. (2003) Unity and diversity of reading literacy profiles. In Lie, S., Linnakylä, P., Roe, A. (Eds) *Northern Lights on PISA. Unity and Diversity in the Nordic Countries in PISA 2000.* Oslo : Department of teacher education and school development, Univerity of Oslo, 147-157.

Martin, M. O., Mullis, I. V. S., Kennedy, A. M. (Eds.) (2003) *PIRLS 2001 Technical Report*. Boston : International Study Center, Lynch School of Education.

McQueen, J. & Mendelovits, J. (2003) PISA Reading : cultural equivalence in a cross-cultural study. *Language testing*, vol. 20, issue 2, 208-224.

Murray, S.T., Kirsch, I.S., Jenkins, L.B. (1998) *Adult Literacy in OECD Countries : Technical Report on the First International Adult Literacy Survey*. Washington D.C. : U.S. Department of Education, National Center for Education Statistics.

National Center for Education Statistics. (1998) *Adult Literacy in OECD Countries. Technical Report on the First International Adult Literacy Survey*. Washington D.C. : U.S. Department of Education.

Purves, A. (1973) *Literature Education in Ten Countries*. Stockholm : Almquist & Wiksell.

Romainville, M. (2002) Du bon usage de Pisa. *La revue nouvelle*, vol. 115, issue 3-4, 86-99.

Thorndike, R. (1973) *Reading Comprehension Education in Fifteen Countries*. Stockholm : Almqvist & Wiksell.

Von Humboldt, W. (2000) *Sur le caractère national des langues et autres écrits sur le langage*. Paris : Seuil. Traduit de l'allemand par D. Thouard.