UNIVERSITÉ DE LIÈGE
FACULTÉ DE DROIT

Année académique 1999 – 2000

DE CONSENSUS EN BLOCUS

La gestion des spécificités nationales, linguistiques et culturelles Dans les évaluations internationales de la lecture

Mémoire présenté par
Ariane Baye
Dans le cadre du D.E.A. en
Relations internationales et intégration européenne

Merci aux passeurs, aux lecteurs et relecteurs

Particulièrement à Madame Dominique Lafontaine
À Monsieur Marc Jacquemain
À Marc Demeuse
Pour leur patience et leurs conseils avisés

Le respect des différences ne procède pas
D'une simple modalité méthodologique,
mais ne prend sa signification que
dans une politique sociale globale.
Gilbert de Lansdheere

Sommaire

Introduction

Des réseaux internationaux de communication à l'expansion des échanges économiques, en passant par l'importance croissante des organismes supranationaux, notre quotidien résonne de plus en plus à l'échelon mondial. Intégratrice, la globalisation entraîne de nouvelles formes d'organisation et de régulation et peut faire rêver à une meilleure compréhension de l'humain, mais elle charrie aussi de nouveaux questionnements identitaires, favorisant aussi le repli sur des entités plus « proches » (d'aucuns diraient à dimension plus humaine) comme la langue, la culture.

> L'unification et l'intégration s'accompagnent d'un processus de concentration et de monopolisation et, du même coup, de dépossession [...]. L'ambiguïté du processus d'unification se voit bien dans l'ordre de la culture et de la langue. L'unification culturelle et linguistique s'accompagne de la constitution d'une langue et d'une culture particulières en langue et en culture légitimes ainsi érigées en normes centrales, et par là « délocalisées », « départicularisées », « universalisées » et, du même coup, constituées en capital linguistique et culturel capable de procurer un profit de distinction [...] qui ont pour contrepartie le discrédit des arts de vivre différents, renvoyés à la barbarie ou à la vulgarité, comme les langues dominées sont réduites au statut déprécié de jargon ou de patois (Bourdieu, 2000).

La problématique abordée ici s'inscrit au cœur de ce double mouvement qui va du global au local en posant la question de la gestion de la diversité culturelle et des spécificités nationales dans les évaluations internationales des acquis des élèves. Pour être valides, les comparaisons internationales supposent en effet un haut degré de standardisation tout au long du processus de recueil et de traitement des informations. La similarité des conditions de passation des tests, des supports d'évaluation et d'analyse des données est aussi gage d'équité pour les pays participants. D'un autre côté, traiter équitablement les pays engagés dans ce type d'évaluation ne va pas sans tenir compte de leurs spécificités linguistiques et culturelles, et de l'extrême variété de leurs systèmes éducatifs. La comparabilité des données est donc intimement liée à la gestion de la diversité.

Dans ce travail, nous nous attacherons à examiner la manière dont les évaluations internationales en lecture ont pu répondre aux questions suivantes : comment s'assurer de la comparabilité des données et de l'équité des évaluations ? Et, partant, quelle est la validité des comparaisons internationales en lecture ?

Outre des raisons très pragmatiques, comme notre intérêt personnel pour le sujet, et l'obligation, pour respecter le cadre restreint de ce travail, d'opérer un choix, ne fût-ce que thématique, dans le corpus des enquêtes internationales, l'intérêt de se concentrer sur l'aspect « lecture » ou « littératie[1] » se justifie à plus d'un titre. D'une part, par l'importance déterminante accordée à la lecture, ou plus exactement, à la littératie, dans nos sociétés, et, d'autre part, par rapport à la problématique de la comparabilité des évaluations internationales.

Littératie, sauvegarde des peuples ?

Les performances en lecture sont plus que jamais l'objet de l'attention particulière des pays membres de l'OCDÉ. Pour le dire brièvement, dans des sociétés qui demandent de plus en plus de main-d'œuvre qualifiée (well-educated), la compétitivité semble inévitablement passer par la capacité à fournir aux employeurs des travailleurs « lettrés », c'est-à-dire capables de comprendre, d'utiliser ou de réfléchir sur des textes écrits pour atteindre des buts personnels, pour développer leurs connaissances et leur potentiel.[2]

> Because literacy has an effect on the ability of workers to learn efficiently and to be flexible in learning, it also has an effect on the rate at which a culture of lifelong learning can be realized. For some, it is the *sine qua non* of workplace learning (OCDÉ, 1995, p. 22).

La flexibilité des marchés implique une flexibilité des travailleurs, et les habiletés qu'acquièrent les personnes apprenant à maîtriser une grande variété de supports écrits (textes au sens large) leur permettront d'apprendre tout au long de la vie, et ainsi de « rester dans la course » sur les marchés de l'emploi à venir.

On attribue aussi à un meilleur niveau en littératie un meilleur niveau de qualité de vie – ce que d'aucuns attribuent à une meilleure instruction tout court (OCDÉ, 1994), et donc une diminution des dépenses à caractère social, et plus de chance de participer activement à la vie politique.

---

[1] Cette notion reviendra souvent de ce travail. Nous la définirons provisoirement comme l'habileté à comprendre, utiliser et réfléchir sur des textes écrits, à des fins personnelles, pour développer leurs connaissances et leur potentiel, et pour se débrouiller dans la société (adaptation de la définition donnée dans le cadre de l'évaluation Pisa 2000).

[2] OCDÉ 2000, nous traduisons.

Équité, qualité de l'enseignement, compétitivité... tout se confond, et sans doute tout est lié pour rappeler aux gouvernements l'urgence d'augmenter les performances de leurs systèmes éducatifs, et pour cela, de mieux les comprendre et les analyser en créant de solides bases de données qui sous-tendront les décisions politiques. Entre coopération et compétition, les pays en lice conjuguent leurs efforts pour évaluer le présent de leurs systèmes éducatifs, et préparer ainsi leur devenir.

De Babel à nos jours, un débat sans fin

Malgré les motivations communes qui poussent les gouvernements à évaluer les performances des élèves en littératie, la question des modalités évaluatives ne fait pas l'unanimité parmi les experts nationaux.

Assurer la comparabilité semble en effet présenter plus de difficultés lorsque l'on évalue la lecture que lorsque l'on s'intéresse aux mathématiques ou aux sciences par exemple. En effet, dans toute évaluation, la question de ce que l'on mesure effectivement – the adequate coverage of the same construct (Cook, 1999, p. 3) – est fondamentale. Dans les évaluations internationales en lecture, ce problème se pose avec d'autant plus d'acuité que « 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 » (Lafontaine, à paraître).

Cela n'est pas pour étonner. Dès qu'il est question de l'usage – ou de l'évaluation – de la langue, on voit souvent s'éveiller les sensibilités et les susceptibilités nationales. Pour les groupes en effet, la langue est un puissant vecteur de cohésion sociale et d'identité.

> 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, p. 55)

Si chaque langue implique une vision du monde originale et unique, si elle « forme un tout dont les différentes parties ne correspondent à aucune de celles du système des autres langues » (Von Humblodt, 2000, p. 7), comment passer de l'une à l'autre, les traduire ou les comparer ? La question de la relation ontologique entre langue, culture et peuple n'a cessé de préoccuper les linguistes, philosophes et sémiologues, de Babel à la naissance de la linguistique comparée, dans l'Allemagne romantique du 18e siècle. Elle resurgit encore aujourd'hui, précisément au moment où le concept de nation s'étiole, et où les groupes sont en quête de fondements identitaires.

Cette question sempiternelle concerne aussi ceux qui évaluent la maîtrise de la langue. Dans le cadre des enquêtes internationales, dont l'histoire est jalonnée de questionnements et de débats relatifs aux spécificités linguistiques et culturelles, cette problématique peut devenir éminemment politique, tant il est vrai que les débats linguistiques émeuvent l'opinion publique dans un domaine qui touche les individus dans ce qu'ils ont de plus précieux : leur outil pour penser.

> En trente ans, […] les études comparatives se sont modifiées sous bien des aspects, mais le défi majeur reste inchangé : assurer la comparabilité, s'approcher toujours d'un peu plus d'équivalence que d'aucuns diront mythique (Lafontaine, à paraître).

1. Les enjeux politiques et pédagogiques des évaluations internationales

L'évaluation des acquis des élèves n'est pas uniquement l'affaire des chercheurs en pédagogie. L'évaluation comparée des systèmes d'enseignement, en termes de rendement des élèves, peut avoir un impact profond sur les politiques éducatives des pays concernés. Le préambule à une brochure de présentation « grand public » de l'enquête Pisa 2000 de l'OCDÉ, tout en évoquant l'intérêt heuristique de telles évaluations, montre aussi les répercussions qu'elles peuvent avoir au niveau des politiques gouvernementales en matière d'éducation.

> Des analyses comparatives réalisées à l'échelon international peuvent compléter et enrichir ces travaux menés au plan national, en déterminant les niveaux de compétence atteints par les élèves de pays différents et en offrant un contexte plus large au sein duquel interpréter les résultats nationaux. Elles permettent de définir des orientations en ce qui concerne l'action des établissements en matière d'enseignement et l'acquisition de connaissances par les élèves ; elles donnent également des indications sur les points forts et les points faibles des programmes d'enseignement. *Associées à des mesures d'incitation appropriées*[3], elles peuvent pousser les élèves à mieux apprendre, les enseignants à mieux enseigner et les établissements scolaires à se montrer plus efficaces. Elles peuvent enfin offrir au pouvoir central des instruments lui permettant de suivre l'évolution des niveaux d'acquis, même dans les cas où la gestion des systèmes éducatifs est décentralisée et où la direction des établissements est assurée en coopération avec les collectivités locales (OCDÉ, 2000, p. 3).

Les évaluations internationales peuvent s'inscrire dans une perspective persuasive, et permettre à la communauté internationale et/ou aux gouvernements impliqués d'exercer une pression considérable dans les classes. D'une part, les résultats de telles évaluations peuvent servir d'incitant – ou de prétexte – à l'introduction d'une réforme pédagogique. D'autre part, si des évaluations externes[4] sont associées à des sanctions ou à des récompenses, les pressions sociales exercées sur les enseignants peuvent les conduire à focaliser leurs pratiques sur les objectifs évalués par les épreuves externes. Cela peut aider les professeurs et les élèves à mieux travailler à l'acquisition des objectifs évalués, mais cela peut aussi les amener à

---

[3] Nous soulignons.

[4] Une évaluation externe est une épreuve centralisée, tant dans sa conception que dans son administration et sa correction (Monseur - Demeuse, 1998, p. 11). Les évaluations internationales que nous envisageons dans ce travail correspondent à cette définition.

« bachoter », à s'entraîner, par exemple, à la pratique d'un type de questions particulier (Monseur, Demeuse, 1998).

Au-delà des influences sur les pratiques et les politiques éducatives, les performances des étudiants et des systèmes éducatifs dont ils ressortent ont d'autres implications éminemment politiques. La publication et la médiatisation des résultats de telles enquêtes provoquent de nombreuses réactions dans l'opinion publique, à tel point que certains gouvernements ont intérêt à voir retarder la publication des résultats internationaux. Parmi d'autres exemples, Margaret Brown (1998, pp. 33-34) mentionne celui de la Troisième étude internationale en mathématiques et en sciences (TIMSS). L'auteure explique que, bien que les résultats de l'étude étaient connus depuis mai 1996, ils ont été maintenus dans la confidentialité jusqu'à la mi-novembre, afin que les élections présidentielles américaines ne soient pas « affectées » par les faibles performances de ce pays.

Par ailleurs, la presse nationale peut orienter la perception des résultats selon ses propres sensibilités ou selon l'image qu'elle a de son système éducatif. Ainsi, lors de la publication des résultats de l'enquête internationale sur la compréhension en lecture de 1991 (IEA Reading Literacy), la presse belge francophone titre « Nos ados : derniers en lecture »[5], ou encore « Analphabètes, nos élèves ? »[6]. En fait, si la Belgique francophone se situait bien au dernier rang des pays industrialisés pour les élèves de 14 ans, les élèves de 9 ans étaient quant à eux dans la moyenne des pays européens. Par contre, en France, *Le Monde* annonce « Premiers de lecture »[7], se focalisant sur la très bonne performance des 14 ans, sans toutefois préciser que les résultats peuvent être revus à la baisse si l'on tient compte de la moyenne d'âge des jeunes français – l'échantillon français étant l'un des plus âgés.

On le voit, même si les évaluations internationales « should not be used to support arguments about the superiority or exceptionality of nations as if the international comparative study is the equivalent of a horse race with winners and losers » (Westbury 1999, dans Hambleton, 1999, p. 13), les performances des pays et leur position dans les classements internationaux ne manquent de susciter fierté ou inquiétudes au niveau publique et politique. Pour les « vainqueurs », il s'agit de tirer parti des résultats obtenus ; pour les « perdants », il

[5] *La Meuse* - La Lanterne, 15 octobre 1992.
[6] *Le Vif/L'express*, 9 octobre 1992.
[7] *Le Monde*, 24 septembre 1992.

s'agit de promettre ou de réaliser des réformes, il y va de la crédibilité sur les scènes nationale et internationale, mais il y va aussi de la compétitivité économique (pour des raisons déjà esquissées dans l'introduction).

Mais les comparaisons internationales en matière d'éducation n'ont pas toujours fait converger autant d'intérêts, ni été au centre d'enjeux si importants. Apparues voici une quarantaine d'années, leurs résultats sont longtemps restés confidentiels, suscitant surtout l'attention (ou les inquiétudes) des chercheurs.

Dans ce chapitre, nous nous pencherons sur la naissance et le développement des évaluations internationales dans le contexte politico-économique de l'après-guerre. Cette première approche historique nous aidera à comprendre comment la transformation des structures internationales a pu engendrer l'intérêt croissant pour l'évaluation des systèmes éducatifs. Cette analyse nous aidera par la suite à mieux mesurer l'importance de la problématique de la gestion des diversités culturelles dans le contexte des évaluations internationales en lecture.

1.1 Historique : les enquêtes de l'Association Internationale pour l'Évaluation des rendements scolaires et leur contexte politico-économique

L'IEA à l'heure des grandes alliances

Depuis 1959, l'Association Internationale pour l'Évaluation des rendements scolaires (IEA) réalise des études comparatives des acquis des élèves. Pionnière en la matière, cette coopérative de centres de recherches aura pendant près de quarante ans le quasi-monopole de l'évaluation et de la comparaison internationales du rendement scolaire. Dans cette première partie, nous allons examiner la naissance de l'IEA et le développement de ses activités dans une époque caractérisée par l'émergence d'organisations internationales.

L'après-guerre est marqué par de profondes transformations des sphères politique, économique, sociale et culturelle. L'internationalisation des structures peut être considérée comme un puissant vecteur de changement, touchant peu à peu les principaux pôles d'activités de nos sociétés.

Avant d'en venir à la problématique de la comparabilité dans le cadre des enquêtes internationales en lecture, nous avons choisi de nous intéresser à leur « internationalité ». L'internationalisation des systèmes a en effet pris toute son ampleur dans les dernières décennies, et la comparaison des systèmes éducatifs ne peut être envisagée sans un regard sur les conditions qui ont influencé sa naissance : de nouveaux enjeux internationaux et de nouveaux défis pour l'enseignement. C'est à cette croisée des chemins que se situe la naissance des évaluations internationales.

1.1.1. De nouveaux enjeux internationaux

Dans un premier temps, nous allons nous pencher sur les différents secteurs d'activités redessinés par les nouvelles donnes internationales. Ensuite, nous verrons comment l'intérêt et les attentes envers les systèmes éducatifs ont eux aussi évolué, pour enfin montrer comment ces systèmes ont pu devenir les centres de convergence d'intérêts dans la vie internationale.

Dès la fin de la seconde guerre mondiale, l'idée de coopération fait son chemin en Europe. La signature du traité de Bruxelles en mars 1948 affirme la volonté de coopérer pour reconstruire l'économie européenne et assurer la sécurité. S'intensifiant sans cesse en Europe par la suite, les alliances vont très vite dépasser ce cadre géographique. Ainsi, un an plus tard, les États-Unis et le Canada cosignent avec une dizaine de pays européens le traité d'Alliance atlantique donnant naissance à l'Organisation du Traité de l'Atlantique Nord. On le voit, l'immédiat après-guerre favorise les unions entre nations. Cependant, si l'on peut percevoir à travers ces exemples les prémices de changements radicaux dans les politiques extérieures des pays, on ne peut pas encore parler de collaboration tous azimuts. Il s'agit surtout d'unions défensives marquées au sceau du « plus jamais ça ».

La finalisation d'une union défensive, qui aurait pu constituer le premier pas décisif dans la construction d'une union européenne, va être tuée dans l'oeuf. Les trois piliers européens (France, Allemagne, Royaume-Uni) craignent tour à tour pour leur souveraineté, tandis que le principal facteur de cohésion, à savoir la hantise d'une nouvelle guerre, tend à perdre de son poids dans le climat de détente internationale qui s'instaure après la mort de Staline. L'échec des négociations sur la création d'une Communauté Européenne de Défense en 1953 en est le meilleur exemple, ce n'est pas sur le terrain politico-militaire que se créeront les nouvelles alliances.

La transformation des structures internationales va passer par le secteur économique. L'une des tâches assignées à la nouvelle Organisation des Nations Unies consiste à créer des instruments indispensables au maintien de courants d'échanges réguliers entre les peuples. L'économie s'impose alors comme une composante essentielle de la vie internationale. Ainsi, la nationalisation du canal de Suez fait prendre conscience aux Européens de leur dépendance vis-à-vis de l'extérieur en matière énergétique, secteur clé de l'économie.

Dans ces conditions, et grâce au travail de longue haleine d'un Jean Monet et des premiers « euro convaincus », deux nouvelles communautés voient le jour avec la signature des traités de Rome en 1957 : le Marché commun et EURATOM. On pourrait penser qu'il s'agit là d'une simple union économique, mais l'intention politique apparaît déjà dans le préambule du traité : « établir les fondements d'une union sans cesse plus étroite entre les peuples européens ». En réalité, « il s'agissait, par la fusion économique ainsi créée, de rendre à la fois possible et indispensable la réalisation d'une union politique ultérieure. L'objet du Marché

commun était économique, mais, indiscutablement, sa finalité, dans l'esprit de ceux qui l'ont négocié, était une finalité politique » (Gerbet, 1994). Dans les années 50, les nécessités économiques vont ainsi créer des alliances transnationales qui rendront ultérieurement possibles les alliances politiques.

Avec l'interdépendance des économies et leur multilatéralisation, on assiste à une transformation radicale des instruments de communication. L'économie-monde et la « villagéisation » de la planète appellent en effet l'organisation de réseaux mondiaux d'échanges d'informations. L'information et la communication deviennent ainsi des biens de consommation courante. Pour créer, gérer et diffuser l'information, de nouveaux organismes apparaissent, dépassant le cadre des États-nations.

> Thème essentiel de l'Unesco, l'information est devenue un droit, avec son corollaire d'indépendance tant à l'égard des intérêts étatiques que des intérêts économiques. Un « nouvel ordre international de la communication » surgit ainsi. […] Apparaît la notion de « transnational », moins liée à l'État-nation que la notion d'international, laissant suggérer un dépassement de l'État-nation (Roche, 1999, p. 43).

L'après-guerre est marqué par la transformation des structures internationales : l'internationalisation des économies et l'intégration d'acteurs non étatiques dans le jeu international modifient des systèmes d'organisations nationaux. L'état doit maintenant intégrer de nouveaux intervenants comme les multinationales, les organisations internationales, les organisations non gouvernementales, et doit tenir compte de nouveaux centres d'intérêts liés à l'apparition de ces intervenants extérieurs.

> Le concept de système international, fondé sur le primat du politique, était donc appelé à se diversifier et apparurent très vite ce que les Anglo-Saxons devaient appeler les « issues areas » (que l'on pourrait traduire par « systèmes d'objectifs »), correspondant à un *élargissement des mobiles à agir dans la vie internationale*[8] (Roche, 1999, p. 41).

Nous venons de voir comment les modifications structurelles des relations entre les nations ont pu ouvrir la voie à de nouvelles collaborations, notamment par le biais d'une demande accrue en information, mais ceci n'explique pas encore pourquoi les performances des systèmes éducatifs ont pu être à leur tour considérées comme un « système d'objectifs »

[8] Nous soulignons.

communs. Les théories des économistes néoclassiques développées à la même époque peuvent à cet égard se révéler très éclairantes.

1.1.2. De nouveaux enjeux pour les systèmes éducatifs

Élaborée dès la fin des années 50, la théorie du capital humain (Schultz, Becker) met en relation la durée de scolarisation d'une personne et ses gains sur le marché du travail. Ainsi, les économistes ont pu mettre en évidence l'augmentation du salaire moyen en fonction de l'augmentation du niveau scolaire. Les travailleurs instruits ont également plus de chances de trouver un emploi, et de le conserver en cas de récession économique (Oi, 1962, dans OCDÉ 1994), ils bénéficient également de meilleures conditions de travail et d'avantages sociaux plus importants (Mathios, 1989, dans OCDÉ 1994).

Outre ces avantages pécuniaires individuels, des recherches ont mis en avant les bénéfices sociaux des investissements dans l'éducation. Denison a par exemple étudié les facteurs ayant contribué à la croissance économique des États-Unis depuis 1929 :

> Une amélioration continue du bagage scolaire des travailleurs américains s'est traduite par un accroissement des compétences et de l'adaptabilité de la main-d'œuvre et a contribué à la progression du revenu national (1979, dans Rumberger, 1994).

Au rang des bénéfices sociaux, une éducation plus poussée peut également réduire le nombre de personnes et de familles pauvres, et donc les dépenses publiques à caractère social - par exemple en matière de santé. Enfin, les individus plus instruits ont de plus grandes possibilités de participer à la vie politique, et de meilleures chances de réussir par rapport à la génération précédente (Rumberger, 1994).

À côté du bien-être individuel et social annoncé par les économistes, l'éducation pour tous est aussi gage, dans des discours plus philanthropiques, de justice, de liberté et de paix. Paix qui, « fondée sur les seuls accords économiques et politiques des gouvernements, ne saurait entraîner l'adhésion unanime, durable et sincère des peuples, et qui, par conséquent, doit être établie sur le fondement de la solidarité intellectuelle et morale de l'humanité ». (Unesco, 1945). Dans le chef de l'Unesco, la diffusion de la culture et l'éducation de tous favoriseront

la connaissance et la compréhension mutuelles des nations et le développement harmonieux des sociétés.

Ces arguments économiques et philanthropiques misent sur une congruence d'intérêts et d'objectifs entre les nations, qui, en combinant leurs efforts, pourraient parvenir à augmenter la qualité de vie individuelle et sociale, et prospérer ensemble. Or, l'augmentation globale de la croissance n'implique pas l'égalité des chances individuelles. Tout dépend en effet de la manière dont est distribué le capital à l'intérieur et entre les nations. De plus, « la force de pronostic des théories de la convergence qui, dans les années 60, avaient misé sur la rationalité universelle de la dynamique du développement des sociétés industrielles, a été démentie par la force d'inertie de profils culturels ayant un caractère national spécifique, aussi et justement dans les économies nationales libérales et capitalistes avancées » (Schriewer, 1997, p. 122).

Sans vouloir lier directement la naissance des évaluations internationales à un mouvement ou à une théorie particulière, il n'en reste pas moins intéressant de constater, dès le début des années 60, une convergence d'intérêts, de courants d'idées, voire d'idéologies, qui conduit peu à peu différents acteurs sociaux à apporter un regard nouveau sur leur système éducatif. Ce regard sera orienté par le prisme de la nouvelle donne internationale aux lendemains de la seconde guerre mondiale.

1.1.3. Naissance de l'Association Internationale pour l'Évaluation des rendements scolaires (IEA)

Les premières évaluations internationales sont organisées par des équipes universitaires réunies dans une association internationale fondée en 1959.

> IEA is an independent, international cooperative of research centers. Its mission is to conduct comparative studies that focus on educational policies and practices so as to enhance learning within and across systems of education (www.iea.nl).

La création d'une telle association relève de ce que décrit l'école du mondialisme (Roche, 1994, pp. 62-68) : l'état n'est plus un acteur unique dans la société-monde. On voit émerger une pluralité d'acteurs aux statuts très divers allant des organisations internationales aux firmes multinationales, en passant par des organisations non gouvernementales ou des

mouvements de libération nationale. Les mobiles d'agir dans la vie internationale se diversifient, les acteurs aussi. On n'est plus dans le « tout politique » : les États doivent concilier leurs intérêts avec ceux des nouveaux acteurs de la vie internationale.

Et c'est précisément sous la houlette d'une organisation internationale –l'UNESCO – que se réunissent à Hambourg une douzaine de spécialistes de l'expérimentation « rêvant de contrôler objectivement les assertions, sinon les clichés relatifs aux vertus et aux faiblesses de divers systèmes scolaires » (De Landsheere, 1986, p. 232). C'est ainsi que naissent les évaluations comparatives par la technique de *surveys* normatifs, et c'est ainsi qu'est lancé un projet pilote en 1959 dont les résultats démontreront la possibilité d'évaluations internationales comparatives. L'IEA est créée officiellement en 1961.

Le but premier de l'IEA est d'organiser des études évaluatives à réaliser parallèlement, selon un même plan général, par un ensemble de pays. Le tableau récapitulatif des activités de l'IEA[9] dans ce domaine montre qu'en quatre décennies, l'association a organisé pas moins de quarante-six campagnes de tests, et couvert une belle variété de disciplines, créant ainsi une banque de données sur les systèmes éducatifs d'une richesse extraordinaire.

L'IEA a également joué un rôle décisif dans la diffusion des méthodes et des techniques de recherche quantitative, spécialement en matière de *surveys* normatifs de rendement.

[9] Tableau présenté en annexe.

1.2 Développements et perspectives : l'intérêt croissant des gouvernements par le biais de structures internationales

Jusqu'ici, nous nous sommes surtout intéressée aux éléments du contexte international favorisant la naissance des évaluations à grande échelle. Dans les années 60, les gouvernements nationaux doivent également gérer un autre problème : l'accroissement continuel de la population scolaire au sein de leur structure éducative soulève l'épineuse question des moyens à mettre en place.

> À la lecture des rapports internationaux, on ne peut manquer d'être frappé par la similitude des évolutions constatées et par la convergence des interrogations soulevées dans les différents pays, en dépit de la diversité de leurs systèmes éducatifs. L'expansion quantitative des années 60 et 70 a revêtu partout des caractères communs : l'augmentation des dépenses publiques consacrées à l'éducation ; l'allongement de la durée de l'enseignement obligatoire ; la quasi-généralisation du collège unique, symbole de démocratisation (Lesourne, 1987, dans OCDÉ 1994, p. 187).

Confrontés aux mêmes problèmes, les gouvernements vont s'associer pour chercher ensemble des moyens de les résoudre. Nous avons vu que de telles associations, inédites avant la fin de la guerre, étaient peu à peu devenues incontournables. Des structures intergouvernementales comme l'Organisation de Coopération et de Développement Économiques[10] (ou plus tard, l'Union Européenne) vont offrir aux gouvernements nationaux des lieux de prise d'information et de concertation propices aux prises de décisions communes.

Entre une coopérative de centres de recherches (IEA) et l'OCDÉ, on retrouve la notion de collaboration, mais on passe de l'action de chercheurs à des réalisations (inter)gouvernementales. Il s'opère ainsi un saut politique, et les résultats des évaluations vont avoir de plus en plus de poids au niveau des pays impliqués.

---

[10] L'Organisation de Coopération et de Développement Économiques, en vertu de la Convention signée le 14 décembre 1960 à Paris, a pour objectif de promouvoir des politiques visant :
- à réaliser la plus forte expansion de l'économie et de l'emploi et une progression du niveau de vie dans les pays Membres, tout en maintenant la stabilité financière, et à contribuer ainsi au développement de l'économie mondiale;
- à contribuer à une saine expansion économique dans les pays Membres, ainsi que les pays non membres, en voie de développement économique;
- à contribuer à l'expansion du commerce mondial sur une base multilatérale et non discriminatoire conformément aux obligations internationales.

La coopération interétatique par le biais d'organisations internationales va être envisagée, en théorie politique, comme un procédé rationnel destiné à optimiser l'emploi des moyens mis en commun par les états membres. La « rationalité » peut être par ailleurs envisagée comme le nécessaire ajustement entre les demandes des États et les nouvelles contraintes internationales. Les états sont ainsi conduits à s'associer pour répondre plus efficacement à des besoins communs. Les décisions des organisations internationales se présentent comme le plus petit commun dénominateur sur lequel se sont accordés les états.

> Parce qu'elles apportent des solutions que les états ne peuvent offrir, les organisations internationales sont à leur tour en mesure de modifier les règles du jeu. Elles disposent d'un pouvoir d'influence, mais celui-ci est insuffisant pour résister aux demandes des états. Leur pouvoir de dire est supérieur à leur pouvoir de faire (Roche, 1999, p. 96).

Il n'empêche, même si leurs recommandations ne sont pas assorties d'effets immédiats, les organisations internationales ne vont cesser de prendre de plus en plus d'importance dans les affaires de leurs états membres. Et, à mesure que les états nations délégueront de leur souveraineté à des entités qu'ils composent mais qui les dépassent, les décisions internationales porteront de plus en plus à conséquence. C'est peut-être pour ces raisons que l'on peut interpréter le positionnement stratégique d'organismes comme l'OCDÉ dans le secteur de l'éducation. Ce secteur, même s'il reste très attaché au contexte national dans lequel il s'inscrit, est aussi l'un des domaines où les enjeux communs amèneront de plus en plus à des orientations et à des politiques convergentes.

La première étape du positionnement de l'OCDÉ dans le secteur éducatif passe par la création du Centre pour la Recherche et l'Innovation dans l'Enseignement, en 1968. Les principaux objectifs du Centre sont les suivants (OCDÉ, 1994) :

- encourager et soutenir le développement des activités de recherche se rapportant à l'éducation et entreprendre, le cas échéant, des activités de cette nature ;
- encourager et soutenir des expériences pilotes en vue d'introduire des innovations dans l'enseignement et d'en faire l'essai ;
- encourager le développement de la coopération entre les pays Membres dans le domaine de la recherche et de l'innovation dans l'enseignement.

Le CERI exerce ses activités au sein de l'OCDÉ, et est placé sous le contrôle direct d'un Comité directeur composé d'experts nationaux dans le domaine de compétence du Centre, chaque pays participant étant représenté par un expert.

La partie la plus visible du travail du CERI consiste en la réalisation d'indicateurs internationaux qui, dans le chef de l'OCDÉ « n'est pas simplement un exercice technique planifié et contrôlé par les statisticiens, mais est, d'abord et avant tout, un exercice politique […] Les indicateurs fournissent également une base permettant de susciter de nouvelles visions et de nouvelles attentes » (OCDÉ, 1994, p. 30).

Par ailleurs, l'OCDÉ publie *Un système d'indicateurs de l'enseignement visant à orienter les décisions des pouvoirs publics*. (1973). Ce titre reflète un nouveau cadre épistémologique dans l'appréhension de l'enseignement. Celui-ci peut se prêter à une analyse systémique et fonctionnelle : le système éducatif ne fonctionne pas en vase clos, il fait partie intégrante de la société dans laquelle il s'inscrit, à l'intérieur comme à l'extérieur de lui jouent les forces sociales, politiques et économiques. Produit social, il doit à son tour produire des résultats, résultats que l'on peut mesurer et comparer.

À partir des années cinquante, la croissance économique et l'expansion des effectifs scolaires donnent un nouveau cadre épistémologique à l'analyse des systèmes éducatifs. Les indicateurs sur l'enseignement sont un outil parfait pour répondre à la demande des gouvernements qui voient dans l'état de leurs systèmes éducatifs une projection de leur développement. Plus qu'un outil d'analyse, les indicateurs sont des instruments de construction guidant, orientant les décisions politiques.

Depuis le début des années 80, dit John Lowe (OCDÉ, 1995), on a vu un intérêt sans précédent de la part des décideurs politiques, des praticiens et des chercheurs pour mettre au point et évaluer les standards de l'éducation. C'est également dans les années 80 que renaît l'attention du public et des décideurs sur les résultats de l'enseignement. En 1983, la publication de *A nation at risk* aux USA relance le débat sur les performances des systèmes éducatifs. La publication des résultats de l'IEA renforce encore le sentiment de la nécessité d'améliorer les résultats de l'enseignement. Pour l'opinion publique, les systèmes d'enseignement ne rencontrent pas les attentes qualitatives et quantitatives qu'on place en eux.

Les conditions sont prêtes pour qu'il y ait une véritable volonté politique d'évaluation des systèmes. En 1987, un nouvel élan est donné à l'action internationale en matière d'indicateurs de l'enseignement, car les États-Unis se sont déclarés en faveur d'un projet sur cette question. En 1990 à Paris, les Ministres de l'Éducation des pays Membres de l'OCDÉ conviennent que l'évaluation des étudiants, des institutions et des systèmes dans leur globalité est une indissociable composante des politiques éducationnelles. Les autorités veulent répondre à des questions comme « Comment la recherche sur l'apprentissage peut-elle aider à mettre au point des buts réalistes pour les écoles ? » « Quels sont les moyens les plus efficaces pour permettre à chaque étudiant d'atteindre son plein potentiel ? » « Comment peut-on établir des standards ? ». Les ministres de l'éducation de différents pays en arrivent à la conclusion « qu'il faut des données pour la transparence et la qualité d'un débat politique ».

La publication des *Regards sur l'éducation*, est un premier pas dans cette voie. Il s'agit d'un nouveau produit fournissant une pléthore d'informations sur le cadre de fonctionnement, le financement et les résultats des systèmes scolaires des pays de l'OCDÉ. Les informations recueillies proviennent principalement des ministères de l'éducation des pays concernés. Ce type d'informations va cependant vite se révéler insatisfaisant pour analyser les performances des systèmes au niveau des acquis des élèves. Les statistiques des pays montrent en effet quel pourcentage de la population fréquente les différents degrés d'études, mais elles ne disent pas quelles compétences et connaissances ont acquis les élèves ainsi formés. Pour répondre à ces questions, l'OCDÉ va donc, après l'IEA, se lancer dans des campagnes internationales de test pour évaluer les acquis des élèves.

> Dans la période antérieure (les années de croissance), la demande des décideurs portait surtout sur une standardisation des définitions statistiques, et sur des mesures quantitatives globales : taux bruts de scolarisation, indicateurs de flux scolaires, dépenses publiques d'éducation, coût par élève, taux d'analphabétisme, classification internationale type de l'éducation. La période actuelle de mondialisation de l'économie caractérisée par une interdépendance et une compétition économique accrues se traduit par une demande d'indicateurs reflétant des caractéristiques plus qualitatives : non seulement la qualité des services éducatifs, mais aussi les résultats obtenus, notamment les acquisitions cognitives et les compétences acquises, l'efficacité des systèmes éducatifs, les inégalités régionales et entre les sexes, les formations extrascolaires, la direction des évolutions dans le temps (Debeauvais, 1997, p. 105).

La décision de l'OCDÉ de mener elle-même des *surveys* à grande échelle implique un redécoupage des terrains d'investigation, à des spécialisations entre l'IEA et l'OCDÉ, mais l'évolution paraît nette : là où l'IEA analyse et informe, l'OCDÉ dont l'éducation n'est ni le premier ni le seul domaine d'activité, influence et prescrit. Le travail de fond mené par l'IEA va maintenant pouvoir être récupéré dans le cadre d'enquêtes ayant un objectif clair : orienter le débat et les décisions politiques.

Les difficultés méthodologiques qu'entraîne ce genre d'évaluation demeurent :

> When large number of countries are involved in such surveys, development of the test instruments involves inevitable compromise about content and coverage. There is always then the risk that international differences in achievement will be compounded by international differences in the validity of tests (Cook, 1999, p. 7).

Question d'aujourd'hui et de toujours, la difficulté d'assurer la comparabilité des données recueillies dans différents pays aux contextes linguistiques, culturels et éducatifs si variés a suscité de nombreux débats chez les experts. Dans la seconde partie de ce travail, nous allons examiner comment les diversités et spécificités culturelles, linguistiques et nationales ont été envisagées dans les évaluations internationales de la lecture.

2. La comparabilité dans un cadre multiculturel ...

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

En guise de préambule

Quand méthodologie rime avec idéologie

En recherche pédagogique comme ailleurs, les techniques utilisées ne sont jamais neutres, elles traduisent des choix idéologiques. Comme l'a montré Stephen Jay Gould pour la mesure du quotient intellectuel, la science, même lorsqu'elle s'appuie sur des méthodologies de pointe, sur des chercheurs de renom et sur d'impressionnantes procédures statistiques, peut, lorsqu'elle est consciemment ou non au service d'un a priori ou d'une idéologie particulière, causer des catastrophes sociales ou individuelles (Gould, 1997).

Les choix posés par les experts lorsqu'ils évaluent le rendement des élèves en littératie, même lorsqu'ils portent sur des aspects apparemment mineurs, peuvent eux aussi avoir un impact plus ou moins important sur les résultats finaux. Tout choix a une raison et une finalité idéologique et politique. Nous n'entendons pas porter le discrédit sur l'ensemble des résultats fournis par les évaluations internationales de la lecture, mais simplement de mettre un instant en lumière des aspects apparaissant trop souvent dans les seuls rapports techniques réservés aux spécialistes.

> Le dogme partagé se cache souvent sous le masque de l'objectivité (Gould, 1997, p. 320).

Les choix méthodologiques ne sont ni bons ni mauvais, mais opérés selon des principes qui méritent au moins des justifications, car les enjeux sont publics. Ils concernent en effet la société dans son ensemble par le biais de ses choix éducatifs. Les questions méthodologiques ne peuvent être reléguées, en raison de leur difficulté apparente, à la seule appréciation des experts forcément partiaux, non par malhonnêteté, mais parce que, responsables de l'élaboration des tests et de l'analyse des résultats, ils sont à la fois juges et parties.

Les évaluations abordées dans ce travail sont construites dans une optique interculturelle. La question de l'adaptation d'un test d'une langue dans une autre ne se pose pas dans les mêmes termes que lorsqu'il s'agit de tests fabriqués dans un certain contexte culturel, et implantés tels quels dans une autre culture, comme cela a été le cas pour les test de QI, par exemple. Dans ce cas, les problèmes des praticiens se concentrent dans la validité et l'applicabilité du test dans la culture cible. Dans le cas de tests élaborés pour un usage multiculturel, il faut penser à l'applicabilité de l'épreuve non pas dans une, mais dans plusieurs cultures, et être attentif aux multiples biais que cela peut induire (Cook, 1999, p. 4). Les concepteurs des évaluations sont d'autant moins aidés dans cette lourde tâche que les « guides d'adaptation » ne sont pas légion[11].

En guise d'introduction

Les pays engagés dans les évaluations internationales diffèrent par leurs contextes politiques, sociaux, culturels, économiques... et éducationnels. Le contrôle des systèmes, leur administration, leur financement et leur évaluation peuvent se répartir sur une échelle allant d'une forte centralisation à une relative autonomie. Les standards éducationnels varient, les priorités assignées à l'école aussi. On peut en outre observer des différences fondamentales dans l'organisation des classes, le choix des méthodes pédagogiques, les relations écoles-parents, les méthodes d'enseignement et leur évaluation.

Les *surveys* internationaux doivent quant à eux réussir, pour produire des comparaisons valides[12] et fiables[13], à évaluer les élèves équitablement. Les outils d'évaluation, les conditions de passation des tests, la correction et l'analyse des données sont autant d'éléments dont il faudra assurer l'équivalence dans les pays testés, sous peine de voir invalidés les résultats

[11] En 1992, l'International Testing Committee a mis au point un guide d'adaptation. Ce document très court (quatre pages) donne des indications pour le développement et l'adaptation des tests, ainsi que pour l'analye des résultats dans un cadre multiculturel. Ces indications ont un caractère très général. Elles ont l'avantage de répertorier les problèmes liés à de telles adaptations, mais ne donnent aucune recommandation « concrète » pour faciliter le travail des « adaptateurs ». Les *Technical strandards for IEA studies* (1999) sont beaucoup plus pratiques, mais ne fournissent que très peu d'informations pour des probèmes à caractère « culturel ». On trouve seulement deux pages consacrées à la traduction.

[12] Pour être valide, un instrument doit mesurer ce qu'il prétend mesurer (Demeuse, 2000).

[13] Si une mesure est applicée à un même objet aujourd'hui et dans une semaine, le résultat doit être semblable (sauf bien sûr si l'objet s'est transformé) (Demeuse, 2000).

d'enquêtes très coûteuses[14]. Les responsables des évaluations internationales doivent donc sans cesse concilier l'hétérogénéité des différents systèmes impliqués et l'obligation de la cohérence internationale. Autrement dit, fabriquer un instrument qui satisfasse le tout et les parties.

> Pour que les interprétations apportent du sens, il faut aussi bien considérer les différences sociales, politiques et économiques entre les nations, que les opportunités offertes par le système éducatif. Les développeurs de tests doivent donc être bien informés des problèmes culturels qui peuvent influer sur les scores (Hambleton, 1999, p. 14).

La culture, telle que la décrit ici Hambleton, subsume toute une série de données sociales, économiques et politiques. Ces données ont fait l'objet de nombreuses études, et ont été largement diffusées ces dernières décennies. Les développeurs de tests ont à leur disposition diverses sources d'information bien étayées se rapportant au contexte socio-économique des pays et aux systèmes d'enseignement impliqués dans leurs enquêtes[15]. Par ailleurs, les tests incluent des questionnaires « contextuels »[16] permettant d'affiner ces informations et de voir leur incidence au niveau local (des élèves, des classes ou des établissements).

Nous ne nous inscrivons pas en faux par rapport à la définition englobante d'Hambleton, car elle a l'avantage de ne pas reléguer les « problèmes culturels » au banc des faits marginaux, voire folkloriques. Cependant, il manque ici d'autres spécificités, appartenant aussi au champ de la culture, comme l'ensemble des aspects intellectuels, comportementaux ou linguistiques propres aux différentes sociétés. Ce genre d'information ne fait pas l'objet de publications systématiques. On ne trouve pas de rapport classant les pays par ordre croissant de dominance culturelle ou linguistique, même si l'on sait que, sur ces marchés-là aussi, certains sont plus « compétitifs » que d'autres[17] (Klinkenberg, 1997).

Nous allons donc examiner comment ont été gérées les spécificités nationales, linguistiques et culturelles dans les évaluations internationales de la lecture. Ce sujet recouvre

---

[14] Les montants en jeu se comptent en millions de francs belges (ex. 615 000 dollars américains rien que pour les coûts internationaux de IEA Reading Literacy, 1991).

[15] On peut par exemple consulter les *Regards sur l'éducation* de l'OCDÉ, les bilans annuels de l'UNESCO ou les rapports d'Eurydice (UE).

[16] Il s'agit de questionnaires soumis aux établissements scolaires, aux professeurs et aux élèves. Ils sont entre autres destinés à recueillir des données sur l'environnement socio-économique des populations évaluées.

de nombreux aspects, allant de l'échantillonnage des populations testées, à la présentation des résultats internationaux, en passant par les conditions d'administration des tests. Nous avons délimité notre champ d'investigation à partir d'une question très pragmatique : les élèves des différents pays sont-ils égaux face à l'instrument de test qu'ils ont sous les yeux ? Ce fil conducteur a entraîné les sous-questions suivantes :

– L'élève a-t-il déjà eu affaire à une évaluation standardisée ?
– Les tâches qu'il doit effectuer lui sont-elles familières ?
– A-t-il déjà rencontré, dans un contexte scolaire, les types de textes ou de documents qu'on lui présente ?
– Est-il habitué aux formats utilisés pour poser les questions et pour y répondre ?
– Les textes proposés conviennent-ils à son environnement linguistique et culturel ?

Ces questions se sont probablement posées pour chaque évaluation internationale. Elles n'ont cependant pas toutes préoccupé les experts de la même manière : certaines ont fait l'objet d'un consensus tacite (le découpage de la matière évaluée, le fait de soumettre des tests standardisés); d'autres en revanche ont été au centre de débats houleux (le format des questions-réponses, la traduction des instruments). Consensus ou pas, elles méritent toutes d'être revisitées sous l'angle de la place accordée à la diversité et aux spécificités des populations évaluées.

Notre propos n'est pas de remettre en question les options théoriques et méthodologiques internationales au nom des divergences entre les systèmes et de particularismes nationaux. Cela reviendrait à rejeter d'emblée l'intérêt des évaluations internationales, intérêt fondé sur la présomption que les choix méthodologiques et organisationnels des différents systèmes éducatifs ont des répercussions sur les performances de leurs élèves. Si nous posons la question de la prise en compte des spécificités nationales et culturelles, c'est parce que nous pensons qu'elles peuvent aussi avoir un impact sur les performances aux tests internationaux. Il s'agit donc pour nous d'examiner à quel point ces spécificités sont prises en compte, aussi bien pour les choix consensuels, que pour ceux qui ont déjà fait l'objet de débats passionnés.

[17] « Le succès actuel de l'anglais est né de l'addition de l'expansion coloniale et commerciale de l'Empire britannique et de l'hégémonie du modèle technologique des États-Unis » (Eco, 1994, p. 374).

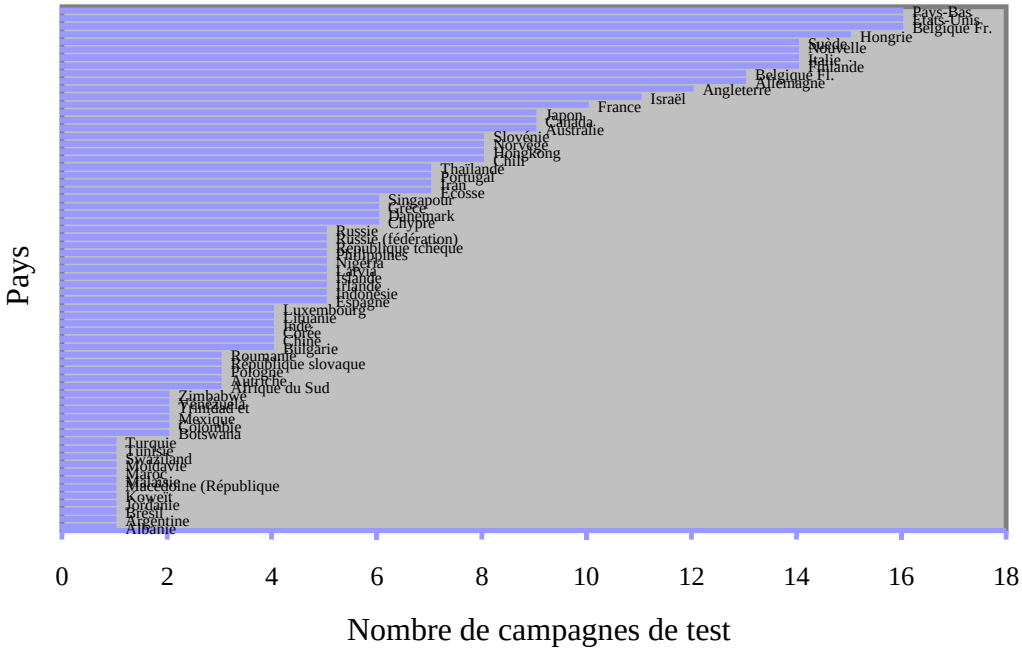
2.1 Les grands consensus

Tous les systèmes éducatifs réservent une place à l'évaluation des apprenants. Les modalités d'évaluation varient quant à elles de système en système. Les évaluations internationales en lecture offrent aux pays la possibilité d'évaluer l'efficacité de leur système éducatif en soumettant des échantillons d'élèves aux mêmes instruments. Si les instruments sont identiques, tous les élèves ne sont pas nécessairement égaux devant les instruments. Le matériel de test peut être mal traduit ou culturellement mal adapté, le mode de questionnement peut paraître inhabituel… Ou peut-être les élèves n'ont-ils jamais rencontré ce type d'évaluation standardisée. Nous commencerons par ce dernier aspect : quel type d'évaluation pratiquent les pays participant par ailleurs aux évaluations internationales ? Ces « cultures d'évaluation » peuvent-elles avantager certains pays lors des épreuves internationales standardisées ?

2.1.1. Des cultures d'évaluation

Dans un premier temps, nous avons recensé la participation des différents pays à dix-huit études internationales[18] - [19]. Nous les avons classés par ordre de fréquence de participation aux dix-huit enquêtes envisagées. En annexe, un tableau plus complet reprend tous les pays impliqués, les années d'évaluation et les matières abordées.

Participation aux évaluations internationales



[18] Cette sélection inclut seulement des études s'adressant (au moins en partie) à des élèves du secondaire. En effet, c'est surtout pour ce niveau d'enseignement que les pays utilisent, au niveau national, des épreuves standardisées.

Nous nous sommes alors demandé dans quelle mesure une « tradition » d'évaluation internationale pouvait être associée avec une « culture » nationale d'évaluation standardisée. Ne disposant pas de données pour tous les pays concernés, nous avons examiné le cas de dix-huit systèmes éducatifs européens[20].

La majorité de ces systèmes (douze sur dix-huit) prévoient une évaluation externe[21] certificative à la fin de l'enseignement secondaire obligatoire et/ou à la fin de l'enseignement secondaire. Peut-on mettre en relation ces pratiques et une tradition de participation aux évaluations internationales ? Les douze systèmes éducatifs européens qui recourent à une évaluation externe certificative au secondaire sont effectivement de grands « consommateurs » d'évaluations internationales standardisées : en moyenne, ils ont participé à la moitié des évaluations proposées, ce qui les place parmi les plus grands utilisateurs d'évaluations internationales au niveau mondial. En outre, huit d'entre eux se sont impliqués dans les évaluations de l'IEA dès les années 70. Pour ces pays, on peut dès lors parler d'une tradition de participation aux évaluations internationales. Nous ne pouvons cependant pas voir dans ces rapprochements – entre participation internationale et évaluation externe nationale – des liens de causalité. Parmi les systèmes éducatifs européens qui ne pratiquent pas d'évaluation externe à fins certificatives, on trouve aussi des pays à grande tradition d'évaluation internationale, comme la Belgique ou l'Allemagne. Par conséquent, il est difficile de soutenir qu'une « culture d'évaluation internationale » est synonyme de « culture d'évaluation standardisée au niveau national ». En réalité, la « culture d'évaluation internationale » apparaît davantage liée à la situation géopolitique et/ou économique des pays – selon une répartition pays « développés » *versus* pays « en développement » (nous y reviendrons).

Voyons maintenant si les pays qui pratiquent des évaluations externes certificatives au niveau national adoptent des modalités communes d'évaluation. Dans l'affirmative, nous comparerons ces modalités aux pratiques internationales. En cas de forte correspondance, nous pourrions conclure à un avantage possible pour les systèmes pratiquant déjà, au niveau national, des évaluations standardisées.

---

[19] Pour cette analyse, nous nous sommes servie des travaux de la Commission européenne (1999), basés sur des données recueillies entre 1996 et 1998.

[20] Allemagne, Autriche, Belgique (francophone et néerlandophone), Dannemark, Espagne, Finlande, France, Grèce, Irlande, Islande, Italie, Norvège, Pays-Bas, Portugal, Royaume-Uni (deux systèmes), Suède.

[21] Par évaluation externe, nous entendons toute évaluation centralisée (cf. note 2, chapitre 1).

Dans la plupart de ces pays, ces évaluations externes sont organisées par le Ministère de l'éducation, mais leur mise sur pied peut être confiée tantôt à une administration locale, tantôt à une commission académique, ou encore à des professeurs. Elles peuvent aussi avoir des enjeux plus ou moins importants pour l'accès à des études supérieures ou au marché de l'emploi. Exception faite du Royaume-Uni, les évaluations externes ont en commun une épreuve sur la langue d'enseignement, qui comprend au moins la rédaction d'un « essai ». Ici encore, les modalités du test divergent selon les systèmes : la longueur des tests, le choix des sujets, le système de cotation, les modalités de réponses sont autant d'éléments qui offrent peu de points de comparaison. On ne peut donc pas dégager de « standards d'évaluation externe » en Europe. De là, nous pouvons conclure que les pays européens pratiquant des évaluations externes au niveau national ne sont pas globalement avantagés par les modalités des évaluations internationales standardisées.

Rappelons que cette analyse n'inclut que des systèmes éducatifs européens. Il aurait évidemment été du plus grand intérêt de soumettre à un même examen tous les pays impliqués dans les évaluations internationales, et notamment les pays « en développement », puisque le tableau reprenant le nombre de participations aux évaluations internationales indique clairement une opposition suivant la situation géopolitique et économique des pays...

2.1.2. Curriculums, habiletés, types de textes

Lorsqu'on parle de comparer les performances en lecture entre les élèves de différents pays, la question de la comparabilité des curriculums (programmes) nationaux s'impose d'emblée. On peut penser que les systèmes éducatifs, même s'ils font tous place à un tronc de disciplines communes comme les mathématiques, les sciences ou la langue d'enseignement, abordent ces matières sous différents angles théoriques et méthodologiques, et que, par conséquent, les connaissances et compétences des élèves varient en nature ou en degré de pays en pays[22]. On peut aussi penser que l'on retrouve des fluctuations de cet ordre au sein de chaque système, en fonction de son degré de centralisation.

[22] Ce genre de considérations est relayé par la question de l'équivalence des diplômes avec laquelle se débat encore l'Union Européenne.

> Les évaluations internationales tentent d'offrir une meilleure compréhension des résultats et des mécanismes qui les déterminent. Dans ce type d'études, chaque système éducatif est situé, non plus par rapport aux objectifs qu'il poursuit, mais par rapport à une norme internationale construite sur la base d'une analyse détaillée de l'ensemble des programmes nationaux d'enseignement. Dans ce cadre, l'efficacité d'un système éducatif revient à se situer parmi les premiers dans le classement international. Toutefois, ces tests recouvrent assez bien les contenus et processus mentaux de certains programmes nationaux alors que pour d'autres, ce recouvrement peut s'avérer relativement faible (Beaton & *al*., 1996 et Martin & Kelly, 1997). *Dans ce cas, le pays risque, au niveau international, d'être considéré comme peu efficace, indépendamment des résultats des élèves à leur propre programme*[23] [...] (Demeuse, Crahay, Monseur, à paraître).

En mathématiques et en sciences, les experts internationaux ont très tôt pris ce paramètre en considération en présentant entre autres les résultats pondérés en fonction de l'adéquation entre les curriculums nationaux et la norme internationale. Par exemple, à l'occasion de la deuxième étude internationale en mathématiques et en sciences organisée par l'IEA (SIMSS), les résultats des États-Unis et du Japon en mathématiques diffèrent fortement. Si l'on pondère les résultats en fonction des curriculums nationaux, la différence tend à s'estomper (Hambleton et Kanjee, 1993)[24].

Pour la lecture, on a jamais présenté de tels ajustements. Dans ce domaine, tout se passe comme si les démarches et processus mentaux en œuvre dans la compréhension en lecture étaient assez universels pour que leur évaluation passe outre les spécificités des curriculums nationaux. Un large consensus international sur les compétences à évaluer serait-il l'écho d'une homogénéité entre les différents programmes nationaux pour la langue d'enseignement ?

Pour répondre à cette dernière question, il aurait fallu procéder à une analyse comparative des différents programmes nationaux en langue d'enseignement, ce qui n'était pas envisageable dans le cadre de ce travail. Nous allons envisager la question par un autre biais : si les curriculums nationaux n'apparaissent généralement pas en aval – lors de la présentation des résultats, les spécificités nationales sont-elles prises en considération en amont – dans la conception des épreuves ?

---

[23] Nous soulignons.

[24] Bien sûr, il faut aussi prendre ces ajustements avec précaution. Dans l'exemple cité, nous ignorons qui a a fait l'analyse de la convergence de l'épreuve internationale et des curriculums nationaux (experts internationaux/experts nationaux/...), et comment les résultats ont été recalculés.

IEA Reading comprehension - IEA Literature (1971) : des approches différentes, en fonction du domaine d'étude

L'étude fondatrice, IEA Six Subjects organise le champ « lecture » en deux études, l'une portant sur la « compréhension en lecture » et l'autre sur « l'interprétation littéraire ». Ces deux volets sont actuellement englobés dans ce que l'on appelle « littératie »[25]. La distinction opérée à l'époque est intéressante, puisqu'elle montre que l'on envisage séparément des compétences liées au contenu cognitif des textes et d'autres liées aux aspects esthétiques ou affectifs (Thorndike, 1973, p. 16). Cette division du domaine va aussi servir de catalyseur pour la prise en compte des programmes nationaux.

IEA Reading comprehension

Le test comprend trois parties, couvrant chacune une habileté (*skill*) particulière : la compréhension en lecture, la vitesse de lecture, et un test de connaissance du vocabulaire. Dans le rapport que Thorndike consacre à cette étude, on ne trouve pas de justification théorique conduisant à délimiter ainsi les habiletés évaluées. Ainsi, pour la vitesse de lecture, l'auteur indique que c'est « a relatively independent aspect of reading skill that has some importance as an academic accomplishment. » (Thorndike, 1973, p. 28). Quant aux items[26] retenus pour la compréhension, ils ont d'abord été choisis de manière à représenter un degré acceptable de difficulté et de discrimination[27], « and, within these limits, an attempt was made to include items covering as wide a range of reading skills as possible, i.e., items dealing with specific factual details, with the main idea, with inference beyond the literal content of the passage, with the author's point of view and purpose, and with the author's writing techniques » (Thorndike, 1973, pp. 22-23).

En fait, l'intuition empirique semble avoir présidé au choix des compétences évaluées. Comme le dit Dominique Lafontaine, « cette évaluation reflète bien l'état des recherches

[25] La littératie est l'habileté à comprendre, utiliser et réfléchir sur des textes écrits, à des fins personnelles, pour développer ses connaissances et son potentiel, et pour se débrouiller dans la société (adaptation de la définition donnée dans le cadre de l'évaluation Pisa 2000).

[26] Item : question, unité élémentaire entrant dans la composition d'un test psychologique (Schumann, 1977, dans Demeuse, 2000).

[27] « On dit qu'un test ou qu'un item de test discrimine bien les sujets quand les scores varient même pour une différence relativement faible de la caractéristique mesurée » (De Landsheere, 1979, p. 73).

théoriques dans le domaine de la lecture de l'époque, marquées par une conception behavioriste : une centration sur des prérequis et/ou des « skills » associés à la compréhension (le vocabulaire, la vitesse), la compréhension représentant quant à elle une sorte de boîte noire ou de continent encore inexploré » (à paraître)[28].

Quant aux curriculums nationaux, ils n'ont pas fait l'objet de l'attention des concepteurs de l'évaluation. Les compétences évaluées dans le test de compréhension en lecture (rapidité de lecture, compréhension, connaissance du vocabulaire) semblent avoir été considérées comme des « universaux » de l'enseignement de la lecture. La question des curriculums nationaux a quant à elle eu plus d'impact dans l'évaluation de la « littérature », deuxième volet du champ « langue d'enseignement », réalisée la même année.

IEA Literature

Ici, la question des curriculums nationaux est prépondérante. L'une des hypothèses de départ de cette étude est qu'il existe des modèles de réponse littéraire[29] (« pattern of expressed response to literature »), qui varient entre les pays suivant la place accordée à la littérature au sein de la nation et dans son système éducatif.

> One would expect there to be differences between nations in the ways by which Literature is approached and thaught, which ways represent the aesthetic, cultural, and pedagogical emphases of that nation. *Unlike other subjects, Literature study has no universal standards of achievement...* [30] It is this sort of difference which will determine the structure, sequence, and material in the Literature curriculum, and which will be reflected in the typical responses of the students to the literacy works that they read (Purves, 1973, pp. 35-36).

[28] On trouve toutefois dans l'ouvrage de Thorndike une référence à des *skills* dont la formulation ressemble furieusement aux compétences évaluées dans les enquêtes ultérieures : « Persons writhin about readint abilities [...] have intended to recognize and be concerned with a number of more or less distincts skills. Such skills as identifying the main idea of a paragraph, finding the answer to a question specifically answered in the passage, recognizing something implied by the passage but not specifically stated, and identifying the writer's purpose or point of view are frequently been mentionned » (1973, p. 28). Outre le manque de justification théorique a priori, cette enquête se différencie aussi des suivantes en ce qu'elle ne prévoit pas à l'avance un nombre d'items par *skill* évalué.

[29] « Response is best defined as the ongoing interaction between the individual and the work » (Purves, 1973, p. 36).

[30] Nous soulignons.

Les spécificités nationales sont largement prises en considération. Dès lors, les responsables nationaux sont plus sollicités : à eux de définir les accents principaux des programmes nationaux. Les résultats de l'évaluation viendront nuancer, confirmer ou infirmer les tendances décrites par les experts nationaux. Le Comité International pour la Littérature analysera les réponses des élèves en fonction de « modèles de réponse » prédéfinis. Les pays seront placés dans des tableaux en fonction de leur approche de la littérature (thématique, affective, analytico-formelle ou encore historique).

En fait, la séparation des champs d'investigation des deux études de 1971 correspond aussi à la distinction culturel-national / non culturel-universel : les performances littéraires sont censées être étroitement liées à une langue et une culture particulières, alors que les habiletés « de base » en lecture relèveraient de processus universels de compréhension partagés par les élèves de l'ensemble des pays participants.

IEA Reading Literacy (1991) – PISA (2000) – PIRLS (2001) : la Littératie par-delà les programmes

Vingt ans ont passé depuis la première enquête en lecture. Les modèles théoriques sur les processus cognitifs de la lecture se sont développés (on a ouvert la « boîte noire » des années 70). Au concept de « lecture », on préfère maintenant celui de « littératie »[31] (« reading literacy »), entendue comme « the ability to understand and use those written language forms required by society and/or valued by the individual. » (Elley, 1992, p. 3).

La nouvelle définition de la compréhension de l'écrit inclut des types d'écrits (destinés à un usage « utilitaire »/ destinés à un usage privé) et des objectifs de lecture (pour se débrouiller dans la société / pour le plaisir). Cette définition implique un dépassement de l'usage scolaire de la lecture. On sent ici les prémices de ce que PISA appellera l'évaluation des « compétences pour la vie ». Dès lors, le lien aux curriculums nationaux sera assez lâche dans les évaluations internationales en lecture, précisément parce que la définition du domaine

[31] Le terme n'est pas encore repris dans le *Petit Robert*. En anglais, l'acception courante (celle du *Collins - Robert*) est « le fait de savoir lire et écrire, le degré d'alphabétisation ». L'évolution sémantique du terme dans le monde de l'éducation reflète une évolution des exigences sociétales par rapport à la compréhension de l'écrit. Il ne suffit plus de savoir déchiffrer pour être un lecteur efficace dans la société.

investigué renvoie à des types de textes et des pratiques, dont on sent, ici plus qu'ailleurs, la finalité extrascolaire[32].

Le choix des types de textes et des habiletés évaluées va suivre l'évolution de la définition du domaine (de la lecture à la littératie) et de la recherche théorique sur les mécanismes de compréhension en lecture (d'une conception behavioriste à une approche cognitiviste ou constructiviste). Il n'est pas certain que les programmes nationaux et les pratiques enseignantes évoluent dans le même sens dans tous les pays. Nous avons vu que, à l'exception d'IEA Littérature, les programmes nationaux n'entrent pas en ligne de compte dans les choix des supports d'évaluation. Voyons maintenant comment ces divergences sont prises en compte dans l'analyse des résultats.

L'IEA n'a pas éludé la problématique des divergences nationales par rapport aux tâches évaluées et aux textes sur lesquels porte l'évaluation. Ces aspects sont étudiés par le biais de questionnaires soumis aux enseignants. Par exemple, dans IEA Reading Literacy (1991), on demande aux professeurs à quelle fréquence ils utilisent différents types de textes (narrations, textes didactiques ou documents), quelles activités de lecture ils privilégient, ou encore leurs objectifs dans l'enseignement de la lecture. Les indications fournies par ces questionnaires sont analysées comme autant de variables contextuelles pouvant être associées à la qualité des performances des systèmes éducatifs.

*

* *

Dégager les variables influençant les performances des différents systèmes éducatifs impliqués est par ailleurs l'un des buts fondamentaux des évaluations internationales. Ces questionnaires offrent donc aux responsables nationaux de précieux renseignements sur les divergences entre les pratiques des enseignants et « l'idéal international », fondé sur la

[32] On pourrait sans doute en dire autant de tous les domaines évalués, car sinon, cela reviendrait à prétendre que certaines matières sont réservées à l'acquisition de compétences « scolaires ». Or, le détachement par rapport au curriculum au nom de la nature du domaine évalué semble être l'apagage de la compréhension de l'écrit.

recherche fondamentale dans le domaine de la lecture, et impliquant un travail sur différents types de textes, sur l'acquisition de « stratégies de lecture » pour des usages variés.

Les standards internationaux sont par ailleurs déterminés par des groupes internationaux d'experts – nous examinerons « l'internationalité » de ces groupes dans la dernière partie de ce travail – et sont soumis à l'aval de tous les pays participants.

Nous pouvons cependant regretter que toutes les évaluations internationales n'interrogent pas les enseignants : l'enquête PISA de l'OCDÉ ne prévoit pas d'interroger les professeurs[33]. D'autre part, on n'interroge jamais les étudiants évalués sur le degré de familiarité avec les types de textes qu'on leur a soumis ou sur les habiletés qu'ils ont dû démontrer.

Par ailleurs, pour vérifier l'adéquation de l'évaluation par rapport aux pratiques nationales, on pourrait envisager une option nationale qui inclurait des questions du curriculum national dans l'instrument d'évaluation international. Ce genre de décision relève des gouvernements, et dépendent de leur volonté de voir confronté les pratiques nationales aux normes internationales.

[33] Le choix d'une population constituée d'élèves de 15 ans, pouvant par ailleurs fréquenter divers degrés d'étude, explique sans doute l'absence de questionnaires enseignant : il aurait été très difficile de réunir les différents professeurs des élèves interrogés. Le choix de ce type d'échantillon présente sans doute d'autres avantages, mais va priver les chercheurs et les décideurs de préciseuses informations sur les pratiques enseignantes.

2.2. Les grands blocus – où se cristallisent les divergences nationales

2.2.1. Le format des questions/réponses[34]

La problématique du format des questions est loin de rencontrer l'unanimité au sein des groupes d'experts. Entre questions à choix multiple et questions ouvertes, les concepteurs de tests doivent opérer des choix dans lesquels de nombreux paramètres interviennent. Parmi ceux-ci, le coût des corrections, leur fiabilité (la correction fait-elle intervenir une plus ou moins grande subjectivité des correcteurs ?), l'authenticité (certaines formes de questions semblent coller de plus près à la réalité[35]) et enfin les compétences évaluées (le fait de pouvoir exercer un jugement ou une réflexion critique serait plus aisé à évaluer par le biais de questions ouvertes que par celui des questions à choix multiple [QCM]).

Pour chacun de ces aspects, on peut aussi se poser la question de la comparabilité des résultats, que ce soit dans une perspective égalitaire (et l'on réfère ici aux habitudes des pays en jeu), dans une perspective technique (selon la plus ou moins grande fiabilité des procédures de traitement des différents types de réponses), ou au point de vue des compétences réellement évaluées (dans le cas de réponse construite, dans quelle mesure les compétences en expression écrite influencent-elles l'évaluation des compétences en lecture ?).

Au fil des enquêtes, les questions/réponses se sont déclinées en plusieurs formats : tantôt des QCM uniquement[36], tantôt seulement des réponses ouvertes[37]. La tendance actuelle est à la combinaison des deux. Nous parlons bien de tendance, car, dans ce domaine aussi, les choix varient et évoluent en fonction des avancées - ou des intuitions - empiriques et théoriques, mais aussi des revendications nationales. Nous allons examiner plus avant six évaluations internationales de la lecture, et tenter, ici encore, de voir si les choix opérés tiennent compte des divergences entre les pays impliqués. Nous verrons comment, dans les débats sur le format des questions, la notion d'« authenticité » liée à des recherches théoriques, va contribuer à modifier les pratiques.

---

[34] Dans cette section, nous nous sommes concentrée sur les tests cognitifs. Notre analyse n'inclut pas les questionnaires contextuels, ni les options nationales.

[35] On parle ici de la réalité sociétale ... qui n'est pas nécessairement parallèle à la réalité scolaire.

[36] C'est le cas dans IEA Reading Comprehension ou IEA Reading Literature.

IEA Reading comprehension (1971) : sans se poser de questions ?

Sans trop se poser de questions - c'est du moins l'impression que donne le rapport Thorndike, la première grande enquête internationale sur la compréhension en lecture opte pour un test « d'un type conventionnel », « in which a passage is presented to the pupil together with *multiple choice questions*[38] based on that passage » (Thorndike, 1973, p. 20).

Malgré l'affirmation de l'auteur, il n'est pas certain que l'usage des QCM ait été uniformément familier pour les quinze pays impliqués[39]. Les États-Unis, par exemple, recourraient souvent à cette pratique, tandis que d'autres pays, comme la Grande-Bretagne lui préférait les questions ouvertes à réponse longue – les essais – ou courte (Hambleton, 1999, p. 5). Il n'est pas sûr non plus que ce choix ait été largement consensuel, puisqu'à la même époque, l'enquête sur la littérature menée parallèlement par l'IEA est l'occasion de débats entre les experts internationaux sur le sujet (nous y reviendrons). Comme le souligne Dominique Lafontaine, une partie des enjeux de cette première grande évaluation internationale de la lecture consiste à montrer la faisabilité même de telles enquêtes,

> les choix proposés en matière d'items sont dès lors très prudents [...] et de toute façon cohérent[s] avec le modèle de la lecture que se sont donné ces évaluations. Pour toutes les questions posées, il existe une et une seule réponse correcte (Lafontaine, à paraître, p. 13).

IEA Literature (1971) : le choix multiple en question

Cette dernière conviction n'emporte pas l'adhésion de tous les membres du Comité International pour la Littérature. Pour certains, « the use of a right-answer or best-answer mode of testing seemed inimical to the ambiguity of literary response » (Purves, 1973, p. 60).

Sous des débats apparemment techniques se profilent ici des approches différentes de la signification. Pour les uns, le sens vient du texte, dans lequel le lecteur « trouve la

---

[37] C'est le cas dans l'International Adult Literacy Survey.
[38] Nous soulignons.
[39] Il s'agit en fait de 14 pays au sens strict. Les communautés flamande et francophone de Belgique ont été traitées séparément, et sont considérées comme deux unités indépendantes dans le compte des « pays ».

signification »; pour les autres, le sens résulte d'une interaction entre le lecteur et le texte[40]. Dans une telle optique, il n'y a pas une seule réponse possible au texte, et encore moins une seule *bonne* réponse possible. On le voit, l'utilisation de questions à choix multiples, assortie d'une classification dichotomique entre bonnes et mauvaises réponses, relève d'une problématique plus épistémologique que formelle.

Toutefois, ces principes théoriques ne font pas d'emblée pencher la balance dans le camp des partisans des questions ouvertes, car elles amènent d'autres questionnements : comment s'assurer de la fiabilité des correction, et, plus pragmatiquement, comment faire face aux coûts qu'elles impliquent ?

Pour éclairer le débat qui oppose les adeptes des QCM et ceux des questions ouvertes, le Comité International pour la Littérature décide de conduire une étude à partir de deux poèmes présentés à un échantillon composé d'étudiants anglais et américains[41]. Le Comité développe une série de questions à réponse ouverte brève (open-ended short-answer question), et le même nombre de questions à choix multiples à quatre propositions. « The open-ended items were designed to parallel the open-ended items exactly » (Purves, 1973, p. 60). Le taux de fidélité entre les correcteurs des questions ouvertes s'avère très encourageant. En revanche, le taux de corrélation entre QCM et question ouverte varie fortement selon les textes. L'incomparabilité des résultats est attribuée à la nature des textes (deux poèmes[42]), et non à la nature des items (QCM *versus* question ouverte). Le Comité procède dès lors à une seconde campagne de test[43], recourant cette fois à des textes en prose. Ici, le taux de corrélation entre les différentes formes du test est jugé satisfaisant, et le Comité de conclure :

> The results also show that multiple-choice items per se do not measure anything different from what is measured by open-ended questions on the same topic. (Purves, 1973, p. 67)

IEA Literature utilisera donc des QCM. Certains pays incluront tout de même des items à courtes réponses ouvertes dans le pré-test, non plus dans une optique exploratoire, ni parce

[40] « Response is best defined as the ongoing interaction between the individual and the work, an interaction that may continue long after the individual has finished reading. » (Purves, p. 36)

[41] L'échantillon est composé de 200 étudiants anglais et de 200 étudiants américains.

[42] Ces résultats ont aussi amené le Comité à modifier le plan d'évaluation de départ, qui prévoyait de tester des poèmes en langue originale dans chaque pays. Cette idée n'est pas tout à fait tombée dans l'oubli, puisqu'une enquête internationale vient de tester les élèves de différents pays selon les mêmes principes, nous y reviendrons.

qu'ils sont convaincus que ce format est mieux adapté chez eux[44], mais pour améliorer les questions à choix multiple. En effet, les réponses ouvertes recueillies dans les pays constituent une banque de réponses « spontanées » (*freely given responses*) dans laquelle pourra puiser le Comité pour la révision des items à choix multiples, notamment pour sélectionner des distracteurs pertinents.

IEA Reading Literacy (1991) : un vent de contestation

Le scénario de 1971 se répète peu ou prou dans l'enquête IEA Reading Literacy : une majorité considère que les QCM sont plus objectives, plus rapidement corrigées, et qu'elles évitent les frais et les difficultés organisationnelles liées à la correction de questions ouvertes. Des voix s'élèvent cependant parmi les responsables nationaux – particulièrement les États-Unis[45], arguant que les questions ouvertes sont plus « authentiques », et qu'elles permettent d'évaluer des habiletés de niveau taxonomique plus élevé. Les responsables d'IEA Reading Literacy décident de tester les deux formats à partir de textes du projet pilote. Ils en arrivent aux mêmes conclusions que leurs prédécesseurs d'IEA Literature :

> The result of this study confirmed the findings of other such studies in reading, namely, that in reading surveys both types of items measure essentially the same abilities, and that multiple-choice items do so in less time, with less cost and are more popular with students. While there are clear reasons for using more open-ended questions in diagnostic testing and other classroom assessment, the value of including many such items in an international survey of reading literacy had not been empirically demonstrated *before this study was conducted*[46]. (Elley, p. 5)

La dernière assertion est à mettre en exergue, car c'est précisément sur la base des données d'IEA Reading Literacy que va être à son tour démontrée l'utilité des questions ouvertes dans les études internationales en lecture. En effet, la version finale du test de 1991 comportait, à

---

[43] L'échantillon, plus important cette fois, est composé de 665 Anglais et de 627 Américains

[44] En tout cas, le rapport international ne mentionne pas ce type de motivations. Malheureusement, le rapport d'Elley ne mentionne pas les différents pays ayant inclu des questions à réponse ouverte dans le pré-test. Il aurait été intéressant voir si, parmi eux, on ne trouvait pas déjà les pays qui allaient prôner l'introduction de ce type de format de réponse dans les enquêtes ultérieures.

[45] Il faut dire qu'à la même époque aux États-Unis, les QCM étaient vivement critiqués. « The available evidence suggest that conventional tests have had at least some negative social consequences for teaching and learning » (Bennet, 1993, p. 24).

[46] Nous soulignons.

titre expérimental, quelques questions à réponse ouverte[47]. Ces données n'ont pas été incluses dans les rapports internationaux. En 1994, les États-Unis, dont le représentant national pour l'étude IEA Reading Literacy avait été l'un des fervents promoteur de l'introduction d'items à réponse construite, réexplorent la possibilité d'inclure ce type d'items dans les analyses internationales (Kapinus et Atash, 1994).

Les travaux de Kapinus et Atash permettent de conclure que, face à une question à choix multiple, les élèves procèdent plutôt par élimination des distracteurs, sans se référer au texte sur lequel porte la question. Ils interagissent dès lors plus avec la question et les propositions de réponses qu'avec le texte lui-même. Au contraire, dans le cas d'items à réponse ouverte, ils tendent à plus se reporter au texte et à construire leur interprétation sur cette base. Ces nouvelles analyses remettent en cause le fait que les QCM mesurent les mêmes processus que les questions ouvertes[48].

> These findings suggest that the two types of items measure different but related aspects of the domain of reading. [...] This finding supports the notion that the use of a combination of constructed-response and multiple-choice items might give a more complete sampling of the domain than multiple-choice items alone. [...] We believe that the use of constructed-response items on large-scales reading assessments would appear to tap the reading process more completely and to reflect good instruction and real-life reading responses more faithfully than relying completely on multiple-choice items. (Kapinus et Atash, pp. 124-107)

Les items à réponses ouvertes peuvent permettre de décrire plus finement les processus évalués, notamment parce qu'ils permettent de mieux appréhender les processus complexes utilisés par les lecteurs. Toutefois, les auteures montrent que ce type de format a aussi ses faiblesses. D'une part, la finesse de l'analyse dépend de la réalisation de guides de corrections adaptés à la variété des processus impliqués, et donc, à la diversité des réponses possibles. Les auteures recommandent dès lors de prévoir des crédits partiels[49]. Elles attirent aussi l'attention sur le fait que, dans ce type d'évaluation, les performances en lecture sont difficilement

[47] Pour la population A (9 ans), quatre réponses nécessitant un ou deux mots (completion-type item) et deux réponses longues (un paragraphe). Pour la population B (14 ans), vingt réponses courtes, et deux réponses longues.

[48] Britton, dans le cadre de IEA Reading comprehension et Elley pour IEA Reading Literacy avaient montré le contraire.

[49] Il s'agit de scores intermédiaires attribués à des réponses partiellement correctes. Ils permettent de prendre en compte des catégories de réponses variées qui témoignent de la diversité des processus de lecture. Dans le cas d'une correction binaire « vrai - faux », ce genre de réponses est amalgamé aux réponses qui ne reflètent pas une compréhension minimale de la tâche demandée.

isolables des performances rédactionnelles[50]. Enfin, la correction des questions ouvertes peut s'avérer moins fiable, car plus sujette à la subjectivité des correcteurs, à l'intérieur d'un pays et entre eux.

International Adult Literacy Survey (1994) : au gré du public

Fait inédit depuis le début des évaluations internationales en lecture, IALS utilise exclusivement des questions ouvertes. On pourrait croire que ce choix est l'aboutissement logique des travaux démontrant la pertinence de ce format de réponse, mais curieusement, ce genre d'argument n'apparaît pas dans les rapports consacrés à IALS. Le choix a été motivé par la nature du public testé : des adultes.

> All of the literacy tasks were open-ended rather than multiple-choice because it was thought that adults would be more interested in performing such tasks. (Kirsch et Murray, p. 19)

L'explication laconique figurant dans le rapport technique de l'étude ne fait pas état d'une support théorique particulier dans le choix du format des questions/réponses[51]. Le plus étonnant ici est que le choix de ce format neuf dans les enquêtes internationales n'ait pas été exploité pour son principal avantage. En effet, les guides de correction de IALS ordonnent les réponses selon une catégorisation « correct - incorrect - omission », sans affiner ce découpage par l'utilisation d'un crédit partiel permettant de classer les types de « bonnes » ou « mauvaises » réponses[52]. Sans crédit partiel, les questions ouvertes privent le chercheur d'informations plus précises sur les stratégies de lecture. Ce mode de correction dépasse donc la dualité correct/incorrect pour permettre d'explorer les voies intermédiaires, les chemins de la compréhension qu'empruntent les lecteurs lorsqu'ils répondent « partiellement ».

---

[50] Les auteures montrent par exemple qu'à qualité égale, une réponse longue obtient souvent de meilleurs résultats qu'une réponse brève. Deux lecteurs « égaux » n'auront donc pas nécessairement des scores identiques, si l'un est meilleur rédacteur.

[51] Le silence sur les fondements théoriques du choix d'un modèle marque peut-être simplement une attitude prudente des responsables de IALS, au moment où la polémique sur les formats de réponses divise toujours les chercheurs (voir à ce sujet Bennet et Ward, 1993).

[52] C'est d'autant plus curieux que le National Center for Education Statistics (États-Unis) a été à la fois responsable de la réanalyse des résultats d'IEA Reading Literacy - analyse qui avait entre autres examiné la problématique des questions ouvertes et recommandé l'utilisation de crédits partiels dans les guides de correction - et un acteur majeur dans la conception de IALS. De même, Marilyn Binkley, éditrice du rapport Kapinus - Atash préconisant les crédits partiels, était aussi la représentante nationale pour les États-Unis dans IALS.

PISA (2000) - PIRLS (2001) : des solutions mixtes, authentiques et pas chères ?

Ces deux enquêtes actuellement en cours se caractérisent par l'utilisation conjointe de questions à choix multiple et de questions ouvertes, elles-mêmes réparties en questions à réponse brève (qui ne nécessitent qu'un seul codeur) et à réponse construite (qui demandent un codage multiple). Ces dernières ont pour but d'amener les élèves à mettre en œuvre des stratégies de lecture plus variées et de niveau taxonomique plus haut (Framework, p. 6). À ces raisons théoriques[53] s'ajoutent des raisons psychologiques : il s'agit de rendre les tâches plus « authentiques » (Lafontaine, à paraître).

Le critère d'authenticité (associé à l'utilisation de réponses construites) revient comme un leitmotiv depuis IEA Reading Literacy. Il est aussi au centre de ce travail, puisque nous sommes partie du principe que ce qui est « authentique » / « vrai » / « naturel » pour les uns (certains modes d'évaluation dans certaines cultures), ne l'est pas nécessairement pour d'autres. Lorsque certains crient à l'authenticité, il est dès lors toujours intéressant de regarder qui parle. Dans le cas d'IEA Reading Literacy, nous avons vu qu'il s'agissait des États-Unis. Nous disposons justement pour cette étude de données concernant les modes d'évaluation utilisés par les professeurs de langue d'enseignement dans les différents pays testés. Dans le rapport international, il apparaît clairement que, si les évaluations du type « essais », « questions ouvertes à réponse brève », « discussions orales » sont utilisées dans tous les pays sans grande variation entre eux, le recours aux questions à choix multiple diffère fortement selon les pays.

> The general tendency seems to be that developing countries and Mediterranean countries use multiple-choice questions with higher frequency than the Nordic countries (Lundberg - Linnakylä, 1992, p. 78).

Si donc les questions ouvertes sont « authentiques » pour tous (encore que ceux qui y recourent le moins soient précisément des pays « en développement »[54]), les QCM semblent moins « naturelles » – entendez moins utilisées – dans les pays « du Nord »... ce qui pourrait expliquer qu'ils réclament, au nom de l' « authenticité », l'utilisation de questions ouvertes.

[53] On ne trouve cependant pas, dans les « frameworks » des deux études, de justification théorique explicite (auteurs ou études) étayant ces affirmations.
[54] Grèce, Zimbabwé, Trinidad-Tobago, Vénézuela.

Les débats théoriques entre les partisans des deux modèles d'évaluation sont loin d'être définitivement clos... « all these issues are complicated and badly in need of research » (Snow, 1993, p. 58). Les concepteurs de tests doivent pourtant prendre des décisions.

PISA et PIRLS ont opté pour des solutions mixtes, incluant différents formats de réponses. Tout en contentant les partisans des évaluations « authentiques », ils limitent les coûts de ce type d'évaluation. Ainsi, lorsqu'on évalue une tâche « simple », par exemple, localiser une information fournie explicitement dans le texte, on peut utiliser une QCM, qui aura l'avantage d'une correction rapide, fiable et peu coûteuse. Lorsqu'il s'agit pour le lecteur de développer une interprétation personnelle sur un texte, on préférera proposer une réponse ouverte construite, qui pourra être codée selon son degré d'abstraction ou son caractère littéral – on recourt au crédit partiel dont nous parlions plus haut.

*

* *

Si l'utilisation de différents formats de questions au sein de mêmes tests permet de concilier les impératifs pratiques et théoriques, elle ne constitue pas moins une source de contraintes supplémentaires pour les développeurs de tests. Pour des raisons psychométriques, il est impératif de veiller à ce que les questions ouvertes à réponse construite soient utilisées pour différents types de tâches. « Sans quoi les variables – format de la question et processus – se trouveraient confondues, avec les difficultés d'interprétation qui s'ensuivent. » (Lafontaine, à paraître). Ce type d'arguments amène les concepteurs de tests à planifier et à calibrer les tests a priori, en fonction de la nature des tâches à évaluer, des formats de questions/réponses, ou encore des types de textes.

Le recours aux questions ouvertes à réponse construite et l'emploi de crédits partiels impliquent la conception de guides de corrections. Il faut en effet prévoir les réponses possibles, et donner aux correcteurs un maximum d'indications et d'exemples pour éviter des biais liés à la subjectivité de la correction, en évitant de donner trop d'informations, pour éviter une surcharge cognitive ! Ce travail délicat engendre une autre difficulté : il faut en

effet que la traduction des guides de correction reflètent les subtilités des réponses prévues par les développeurs de tests. Souvent pointée du doigt lorsque l'équivalence des tests est mise en cause, les méthodes de traduction ont elles aussi évolué au gré des évaluations internationales. Voyons ce qu'il en est.

2.2.2. La traduction

Sans verser dans un radicalisme linguistique qui voudrait qu'une langue spécifique soit le produit d'une culture particulière, voire son essence, et que ses richesses et particularités ne puissent être traduites d'une communauté à l'autre[55], il ne faut pas sous-estimer le problème de la traduction dans les évaluations internationales. L'équivalence de tests passés dans des langues différentes ne coule pas de source. Bien au contraire, de la conception des épreuves à l'interprétation des résultats, la « traduisibilité » s'est toujours avérée problématique.

La question de la traduction du matériel de test est d'autant plus intéressante pour notre propos que les enquêtes que nous envisageons évaluent les performances des élèves en langue d'enseignement. Si on évalue « la capacité de comprendre, d'utiliser et d'analyser des textes écrits »[56], il faut que les supports écrits soient également compréhensibles, utilisables et analysables, par-delà les frontières linguistiques et culturelles.

Certaines difficultés de traduction ont un impact direct sur l'équivalence des tests. Au niveau lexical, les traducteurs doivent trouver dans la langue-cible[57] des termes comparables en longueur, en degré d'abstraction et en fréquence d'utilisation. Tous ces paramètres interviennent pour influencer le degré de lisibilité d'un texte[58]. 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

[55] Les arguments des partisans de cette position « extrême » sont présentés dans l'introduction. Pour notre part, même si nous pensons qu'il existe des divergences dues à l'origine linguistique et culturelle, nous croyons à la « communicabilité » entre langues et cultures.

[56] Il s'agit ici de la définition de la littératie que donne l'OCDÉ (OCDÉ, 2000, p. 11).

[57] Il peut s'agir aussi d'une même langue parlée dans un autre pays. On peut par exemple distinguer l'anglais des États-Unis, du Royaume-Uni ou d'Australie. Un exemple, francophone celui-là : 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 Belge que l'original anglais !

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 le même sens connotatif dans un autre contexte linguistique. Enfin, certains termes n'ont pas d'équivalent dans une autre langue, tandis que d'autres en ont plusieurs[59].

Au niveau de la phrase, la simplicité ou la complexité des structures syntaxiques peuvent être difficiles à rendre. À la charnière entre la phrase et le texte, il faut aussi trouver les ressources stylistiques qui donnent à un texte un ton, une coloration ou une connotation précise.

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. Cela peut se révéler particulièrement gênant pour des questions basées sur la correspondance synonymique. D'autre part, 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. Le metteur en page – actuellement, on confie souvent cette tâche au traducteur – devra 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 » (Grisay, 1998, p. 8).

---

[58] « 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 (les 1500 à 3000 mots les plus fréquents et les plus faciles) est, le plus souvent, composé de mots très courts. » (Grisay, 1998, p. 8).

[59] 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'ils s'agit des aînés ou des cadets... et le coréen utilise neuf formes différentes ! Le fameux « mouton » français sera « mutton » ou « sheep » selon qu'il garnisse votre assiette ou qu'il paisse.

Ces écueils évités, la traduction devra encore se surpasser en s'effaçant elle-même, elle ne devra pas « sentir » la traduction, le « ce n'est pas comme ça qu'on l'aurait dit ». On pourra alors parler de matériel de test « équi-lisible », dans le sens d'également compréhensible, interprétable et recevable par différents groupes linguistiques.

Dans chaque évaluation internationale en lecture, des précautions ont été prises pour assurer la qualité des traductions et, partant, la comparabilité des supports de test et l'équivalence des résultats. En amont, on met en place des moyens de contrôler la qualité du travail des traducteurs. En aval, la qualité du matériel est vérifiée de manière empirique et statistique lors d'un prétest. Ces contrôles sont aussi effectués lors du test définitif.

Le processus de traduction

Nous l'avons dit, les développeurs de test doivent s'assurer que les formes linguistiques utilisées dans les épreuves conviennent également pour toutes les populations concernées. Cela implique des procédures de contrôle lors du processus de traduction.

> Instrument developers/publishers should insure that the adaptation process takes full account of linguistic and cultural differences among the populations for whom adapted versions of the instrument are intended (International Test Commission, 1992)[60].

Pour rencontrer ce critère, on peut agir à trois niveaux : le recrutement des traducteurs, leur formation, la vérification de la correspondance entre leur version et l'original.

Les traducteurs choisis peuvent être ou non des professionnels. On peut en engager un seul, ou préférer un groupe de personnes qui travaillent ensemble, ou encore plusieurs traducteurs travaillant indépendamment.

Les traducteurs peuvent être ou non formés par les responsables de l'étude. La formation a pour but de familiariser les traducteurs avec ce type d'étude (buts de l'étude, matière abordée, formats des questions, types d'informations visées par les questionnaires...) ou avec le travail de traduction lui-même (les « pièges » à éviter).

[60] L'International Test Commission a réuni un groupe de treize personnes pour développer des standards techniques d'adaptation de test en 1992. Elle produit une série de recommandations, dont sept concernent directement les précautions à prendre en matière de traduction. Celle-ci est la première.

Pour vérifier la concordance entre l'original et l'adaptation, on recourt généralement à l'une des deux méthodes suivantes :

La *backward translation* : un groupe de traducteurs adapte le test, un second groupe prend le test adapté (version langue-cible) et le réadapte dans l'autre sens, vers la langue-source. On compare l'équivalence des deux tests en langue-source. Les divergences indiquent des erreurs de traduction. Ici, les développeurs de tests peuvent eux-mêmes vérifier les adaptations. C'est pratique, mais ça ne suffit pas pour démontrer la validité des adaptations : des automatismes de traducteur peuvent être faciles à retraduire par un autre traducteur. Imaginons qu'il y ait, dans la version traduite, une tournure syntaxique propre à la langue d'origine. Cette tournure sera difficile à interpréter pour un élève. Si on demande à un traducteur de retraduire cette tournure syntaxique dans sa langue – la langue d'origine, il retrouvera sans doute la tournure de départ, d'autant plus facilement que cette forme syntaxique lui est familière. La version originale et la *backward translation* seront donc similaires. La version traduite sera quant à elle considérée comme correcte, alors qu'elle contient une tournure syntaxique erronée qui « sent la traduction ».

La *forward translation* : un traducteur ou un groupe de traducteurs adapte le test d'une langue source vers la langue cible. L'équivalence entre les deux versions est évaluée par un autre groupe de traducteurs. Cette méthode permet de poser directement un jugement sur l'équivalence des deux versions. Le désavantage est le haut niveau d'inférence que les traducteurs ou les vérificateurs doivent faire à propos de l'équivalence. Ce procédé fait intervenir la subjectivité des juges, qui ne possèdent pas de support matériel ou statistique pour confirmer leurs impressions. De plus, la *forward translation* ne permet pas aux développeurs de tests de juger de l'équivalence entre les versions. Or, leur avis est précieux, puisqu'ils savent ce qu'ils veulent tirer du matériel de test.

Le pré-test

Le pré-test est une expérimentation sur le terrain, dans des conditions relativement similaires à celles du test définitif[61]. Il a pour but de valider le matériel de test (textes, items,

[61] L'échantillon est plus petit, il n'est pas toujours sélectionné selon les critères prévus pour le test définitif. Tous les pays participants n'effectuent pas toujours un pré-test, soit par manque de moyens, soit parce qu'ils décident de participer au dernier moment, voire quelques mois après que les premiers ont commencé.

questionnaires) en établissant ses caractéristiques psychométriques. Il sert aussi à repérer tous les problèmes organisationnels ou communicationnels.

Les psychométriciens profitent des données du pré-test pour détecter les items qui fonctionnent « bizarrement » dans un ou des groupes donnés.

> Instrument developers/publishers should apply appropriate statistical techniques to 1) establish the equivalence of the different versions of the instrument, and 2) identify problematic components or aspects of the instrument which may be inadequate to one or more of the intended populations (International Test Commission, 1992).

Prenons un exemple : on soumet un item parmi d'autres à dix groupes linguistiques. Pour neuf d'entre eux, le taux moyen de réussite est de 50 %. Le dixième obtient quant à lui un taux moyen de réussite de 20 %. Ce résultat médiocre peut s'expliquer de deux façons. Si le dixième groupe linguistique est plus faible que les neuf autres, le score obtenu est « normal » puisqu'il confirme les faibles performances du groupe. Par contre, si le dixième groupe linguistique obtient généralement (pour les autre items) des résultats comparables à ceux des autres groupes, on dira que l'item fonctionne « anormalement » dans ce groupe linguistique, puisqu'on attendait ici un taux de réussite moyen d'environ 50 %.

L'analyse du fonctionnement différentiel des items (DIF – Differential Item Functionning) permet de repérer les items atypiques, dans le sens d'*anormalement* faciles ou difficiles pour un groupe par rapport à un autre groupe.

Ces comportements atypiques peuvent être attribués entre autres à de mauvaises traductions (rendant certaines questions plus faciles ou plus difficiles pour certains groupes), à des problèmes d'adaptation culturelle ou encore à des formats de questions inadaptés. Bien sûr, il s'agit là de conjectures, car une analyse DIF permet d'observer des comportements différentiels, pas de les expliquer. Dans les évaluations internationales en lecture, on élimine généralement les items qui « fonctionnent mal », sans analyser en profondeur les causes de ces anomalies. On risque alors de supprimer de la version définitive du test des items intéressants malgré - ou en raison de - leur comportement anormal.

En effet, les « spécialistes matières » ne maîtrisent pas nécessairement les arcanes des analyses psychométriques. Ils sont parfois amenés à faire une confiance « aveugle » aux

recommandations des psychométriciens... qui utilisent des méthodes comme l'analyse DIF pour valider a posteriori la traduction, et supprimer les items qui sortent des normes. Certains avis plus « subjectifs », ceux des experts qui « sentent » pourquoi ça doit fonctionner différemment dans leur langue ou dans leur système éducatif, ne trouvent pas toujours un support statistique, et n'ont donc pas toujours le poids qu'ils mériteraient.

Chaque évaluation internationale en lecture a géré à sa façon l'exécution et le contrôle de la traduction. Nous allons les passer en revue. Cet examen nous permettra de suivre l'évolution de la prégnance de la question de la traduction, et de mettre en rapport les choix méthodologiques avec la manière d'envisager le problème de l' « équi-lisiblité » du matériel de test.

IEA Reading compréhension (1971) : entre les paroles et les actes

La traduction apparaît en première ligne dans les préoccupations des concepteurs d'IEA Reading compréhension. Même si la faisabilité d'une étude internationale en lecture a été encouragée par les résultats de l'étude pilote réalisée dix ans auparavant, l'enjeu principal de cette étude fondatrice est bien de confirmer les premiers résultats par une évaluation « grandeur nature ». Il faut dès lors convaincre les sceptiques de la comparabilité des résultats, et pour cela, résoudre notamment « the formidable translation problems » (Husén, dans Thorndike, 1973, p. 11)

En regard de ces précautions oratoires, les mesures prises pour assurer la qualité des traductions semblent très minces : le bon sens et un pré-test réduit.

> [...] it was felt that *with some care*[62], the test could be maintained as nearly enough the same task from one language to another to make the cross-national comparisons interesting and fruitful (Thorndike, 1973, p. 15).

Au bon soin des représentants nationaux d'envoyer textes et questions traduits vers l'anglais au Comité international pour la lecture, de commenter le matériel international

[62] Nous soulignons.

sélectionné[63] par la même instance, et finalement de traduire vers leur langue nationale les textes choisis.

Après cette phase d'élaboration, un pré-test est organisé afin de détecter les items qui auraient un niveau de difficulté atypique dans l'une ou l'autre langue. La quantité de matériel à traduire semble cependant trop importante pour pouvoir être traitée par chaque pays participant. Les pays pré-testeront donc seulement un tiers du matériel, sur des échantillons de deux à trois cents élèves (« but samples were in some cases smaller than that ») (Thorndike, 1973, p. 21). Le Comité International pour la lecture s'assure tout de même d'une diversité linguistique minimale en pré-testant chaque partie du test dans quatre pays, dont un seul anglophone (« but a few passages were tried out in only two or three countries. ») (Thorndike, 1973, p. 21)

Les résultats du taux moyen de discrimination des items, ainsi que la fiabilité du test, donnent aux pays anglophones un léger avantage, ... considéré comme normal, et même « gratifying », puisque les items ont été développés et édités en anglais (Thorndike, p. 27). Ce bel optimisme n'empêchera pas le même auteur de remettre en cause le test de vocabulaire en raison d'importants problèmes de traduction[64] (« ... it must be admitted that the final result is one in which the equivalence from language to language is suspect ») (Thorndike, 1973, p. 33).

IEA Reading Literature (1971) : la culture avant la langue ?

Il est plus facile de se convaincre de la « traduisibilité » d'un texte à vocation non littéraire[65], que de soutenir qu'un texte d'auteur ne perd rien à passer d'une langue à l'autre. Un texte littéraire est un produit culturel par excellence, exploitant les possibilités stylistiques

[63] Le Comité sélectionne les textes en fonction de leurs qualités stylistiques, de leur contenu et de leur « quasi-universalité » (Thorndike, 1973, p. 20).

[64] Les paires de mots (antonymes ou synonymes) les plus compliquées en anglais avait été simplifiées dans de nombreuses traductions nationales. Par exemple, le couple antonymique « pessimistic - sanguine » avait été rendu par « optimiste - pessimiste » en français. (Thorndike, 1973, p. 132) On imagine l'embarras des traducteurs improvisés des centres nationaux devant trouver dans leur langue, et sans recourir à des périphrases, des couples de mots de difficulté comparable, d'autant que leur dictionnaire ne leur offrait pas d'indice de fréquence des mots.

[65] Encore que la lecture de modes d'emplois peu compréhensibles peut faire facilement douter de cette affirmation simpliste.

d'une langue, prenant toute sa signification dans le contexte culturel où il naît, et pour lequel il est créé.

Pour les concepteurs de la première enquête internationale en littérature, l'apprentissage de la littérature fait partie intégrante de la construction de l'identité culturelle (Purves, 1973, p. 15). L'une des hypothèses générales présidant au développement des tests est qu'il existe des modèles de réponses littéraires (« patterns of expressed response to literature »), et que ces modèles varieront de nation en nation, selon la place que chacune réserve à la littérature (Purves, 1973, p. 39).

Très sensibles aux spécificités culturelles nationales, les posant même a priori, les concepteurs de l'évaluation en littérature vont paradoxalement se montrer relativement moins vigilants quant à la qualité des procédures de traduction. En fait de paradoxe, il s'agit peut-être plus d'une conséquence : ayant posé des divergences nationales « naturelles » (ou naturellement culturelles), les différences de performances entre les groupes viendront appuyer cette hypothèse, que ferait vaciller au contraire une remise en question des résultats en raison de problèmes de traduction.

Ce qui précède ne signifie pas qu'il n'y ait eu aucun contrôle sur la qualité des traductions. À ce niveau, l'évaluation en littérature est même plus scrupuleuse que sa jumelle en compréhension en lecture. Examinons maintenant les précautions prises en matière de traduction.

Un choix prudent, étayé par les faibles résultats sur la comparabilité de la poésie obtenus à l'occasion des discussions sur le format des questions (cf. supra), va conduire le Comité à ne retenir que des textes en prose, pour lesquels « there was no similar problem [...] since the pilot study [1959] had already determined the comparability of prose passages. » (Purves, 1973, p. 60). Confiant dans la comparabilité des textes en prose, le Comité ne se focalisera pas sur la précision des procédures de traduction. Le matériel de test est ainsi envoyé en anglais aux comités nationaux pour traduction avant le pré-test. Ce dernier fait apparaître des problèmes de traduction, particulièrement au niveau des items des choix multiples, qui ne diffèrent parfois entre eux que par un ou deux mots. (Purves, 1973, p. 70) Le Comité décide alors de resserrer le contrôle des traductions, et nous trouvons dans le rapport de Purves les

premières mentions du recours à des méthodes de traduction professionnelles dans les enquêtes internationales. 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 items et questionnaires d'attitudes face à la littérature seront cependant exclus de cette vérification). D'autre part, le matériel de test va être retraduit vers l'anglais (*backward translation*). Les deux versions anglaises (la version source et la traduction anglaise de la traduction en langue nationale) vont ensuite être transmises par Allan Purves à des collègues du département de langue de l'Université de l'Illinois, afin qu'ils comparent les deux versions et y détectent des sources d'erreur. Ce travail donnera lieu à des révisions mineures. (Purves, 1973, p. 75)

IEA Reading literacy (1991) : l'équivalence sous le contrôle des procédures statistiques

> Any translation process entails the possibility that the meaning is lost in translation, and an international test is no exception. (Elley, 1992, p. 8)

Cette circonspection va se concrétiser dans une série de mesures pour éviter les biais dus à des erreurs de traduction, tant au niveau du processus de traduction qu'au niveau de la sélection finale du matériel de test.

On demande aux pays de réaliser deux traductions indépendantes. Les coordinateurs nationaux (NRC) supervisent et rectifient les traductions, à l'aide d'un guide de consignes fourni aux traducteurs et aux Coordinateurs nationaux par le Steering Commitee. Enfin, on demande aux NRC de fournir une *backward translation* pour une partie des textes et des items. On prendra soin de vérifier également la forme – mise en page, illustrations. Quant aux adaptations nationales permises, elles consistent surtout à des modifications des noms des personnages, des lieux, des monnaies ou des unités de mesure (Elley, 1992, p. 2).

Après avoir analysé les réponses du pré-test, l'équipe IEARL retire plus de la moitié des items pour le test final. En outre, quelques items du test définitif n'ont pas été pris en compte dans la communication des résultats finaux, en raison de leur mauvaise qualité psychométrique (Elley, 1992, p. 96).

IALS (1994) : un « heureux » incident

Les concepteurs de IALS ont-ils été moins scrupuleux que leurs prédécesseurs d'IEA Reading Literacy ? On mentionne bien l'emploi de directives pour traduire et adapter les instruments, ainsi qu'une révision minutieuse des instruments de test, mais les rapports ne parlent pas d'autres procédures de contrôle de traduction[66]. IALS n'avait-elle pas mesuré les enjeux et les risques d'écueils liés au choix du format des réponses ? Un 100 % questions ouvertes – et son caractère inédit – appelait des corrections « manuelles » dont il fallait assurer l'homogénéité. Dès lors, la qualité de la traduction des guides de correction devenait un élément crucial pour assurer la comparabilité des résultats. Le rapport technique ne dit rien des précautions prises à cet égard. Cependant, le recodage (recorrection) des réponses (cf. supra) montre une grande fiabilité entre les corrections de différents pays (94 % à 97 % d'accord sur les corrections) (Darcovich, Murray, 1998, p. 89).

Quoi qu'il en soit, la France met en doute la qualité méthodologique de l'étude, « *despite* having participated in every phase of development » (Kirsch et Murray, 1998, p. 16). Ceci expliquerait-il ce retrait ? Un rapport d'experts « neutres »[67], engagés pour vérifier le bien-fondé des accusations françaises, identifie certaines faiblesses méthodologiques, mais recommande la publication des résultats[68]. Rien à faire, la France se retire. Loin d'être anecdotiques, ces remous dans le monde de l'évaluation internationale provoqueront « un formidable processus de réflexion sur toutes les facettes destinées à assurer la comparabilité des données, dont les enquêtes ultérieures tireront un bénéfice majeur. » (Lafontaine, à paraître, p. 18)

---

[66] On peut même percevoir un certain laisser-aller des pays par rapport aux procédures en place, puisque les pays « chose not to incorporate a number of changes which were identified during the course of the review, believing that they " knew better " » (Darcovich, Murray, p. 77).

[67] Il s'agit de Graham Kalton, Lars Lyberg et Jean-Michel Rempp, engagés par Statistics Canada pour revoir tous les aspects de l'étude.

[68] Les auteurs disent toutefois n'avoir eu ni le temps ni les moyens financiers suffisants pour mener en profondeur les analyses qu'on leur demandait. Ils recommandent également aux auteurs des rapports internationaux d'avertir leurs lecteurs des effets possibles des biais méthodologiques qu'ils ont mis à jour. Cette dernière recommandation ne sera pas suivie d'effets dans les rapports internationaux les plus publics (seul le rapport technique fait état des difficultés méthodologiques).

PISA (2000) : sous le contrôle exactement

Après « l'incident français », rien ne sera plus laissé au hasard en matière de traduction. C'est du moins le sentiment que donne la lecture des « Consignes pour la traduction du matériel PISA » (OCDÉ, 1998). Le dispositif est impressionnant. Deux traducteurs professionnels[69] réalisent des traductions indépendantes, qui sont conciliées par une troisième personne (souvent un chercheur du centre national ayant l'expérience de ce type d'enquête et une bonne connaissance des matières évaluées). La version proposée doit ensuite être soumise à l'approbation du Comité National, avant d'être envoyée au Centre de Coordination Internationale, où une équipe de traducteurs « internationaux » vérifiera à son tour la qualité du travail.

On trouvait déjà des éléments de ce dispositif de traduction dans IEA Reading Literacy. Les contrôles sont cependant plus serrés ici. Par exemple, le feu vert pour l'impression des carnets de test n'est octroyé qu'après que le centre international est assuré de l'introduction effective de ses corrections dans le matériel de test. En outre, les responsables nationaux de projet (NPM) ont l'obligation de fournir des listes exhaustives des « adaptations nationales »[70] introduites.

En fait, la différence majeure par rapport aux autres études réside dans la multiplication des procédures de contrôle mises en place. Les pays n'ont plus la possibilité d'ignorer les recommandations des vérificateurs sous prétexte qu'il « savent mieux » (cf. note 53). Dans PISA, le contrôle est contraignant et fait autorité.

Par ailleurs, PISA a intégré de réelles innovations. La première est sans doute à mettre en relation avec le commanditaire. L'OCDÉ est une organisation bilingue anglais/français. Le matériel de test est fourni dans les deux langues à tous les pays. On recommande aux traducteurs d'utiliser les deux versions, ce qui augmente la finesse de la compréhension du matériel et leur permet de résoudre certaines ambiguïtés.

[69] Le recours à des professionnels n'est pas obligatoire, mais « fortement conseillé » (Grisay, 1998, p. 6).
[70] L'emploi de tournures « locales » pour le lexique ou la syntaxe, l'adaptation des noms, des lieux, des unités de mesure, des monnaies, ainsi que la correction des coquilles ou erreurs trouvées dans les versions source.

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 première version est traduite vers le français selon la procédure générale valable pour toutes les langues (deux traductions indépendantes, conciliations, vérification). 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 le nombre d'altérations liées au nombre de traductions.

Par ailleurs, la solution « bilingue » de l'OCDÉ double le nombre de versions successives du test. Les développeurs des tests envoient en effet aux traducteurs les différentes moutures des épreuves au fil de leur élaboration[71]. Les traducteurs doivent redoubler de vigilance pour intégrer dans leur version tous les changements, parfois ténus ou peu visibles, apportés à la version source. Il faut aussi être certain d'avoir travaillé sur la dernière version envoyée. C'est d'autant moins évident lorsque l'on reçoit *deux* dernières versions provisoires (l'une en anglais, l'autre en français), que ces versions n'arrivent pas nécessairement en même temps, et qu'elles ne concordent pas toujours parfaitement (l'une des deux versions pouvant avoir elle-même omis certaines modifications).

Deux autres innovations témoignent de l'importance névralgique qu'a pris le processus de traduction dans PISA.

D'une part, on crée des instruments à l'attention des traducteurs. Les « Consignes pour la traduction » (OCDÉ, 1998) expliquent aux traducteurs les enjeux de l'étude, l'importance de leur contribution pour garantir l'équivalence des instruments, les procédures de traduction, les pièges à éviter... Une formation aux procédures de traduction est également organisée lors d'un meeting pour les responsables nationaux des projets.

D'autre part, les développeurs de tests conçoivent des instruments *destinés à être traduits*. Ils rédigent des notes pour certains textes ou items qui indiquent au traducteur les nuances stylistiques que doivent inclure leur version. Ils attirent l'attention sur tel ou tel distracteur dans une QCM (qui doit par exemple reprendre une expression directement tirée du texte). De

[71] Si l'on envoyait tout le matériel fini, les traducteurs devraient travailler dans des délais « intenables ». De plus, certaines modifications sont le fait des remarques des traducteurs qui détectent des problèmes, il est donc important qu'ils commencent à travailler sur des versions provisoires.

plus, toutes les questions sont précédées d'une rubrique « objectif de la question », qui décrit la tâche évaluée. La nature de la tâche peut guider la traduction : s'il s'agit pour l'élève de « trouver une information – correspondance synonymique », le traducteur veillera à ce que les termes précis du texte se retrouvent dans l'énoncé de la question.

*

* *

Ces derniers exemples le montrent, la traduction est au cœur de l'évaluation. Trop à cœur ? Le déploiement d'une batterie de mesures destinées à assurer la qualité, notamment au niveau de la traduction, participe à l'augmentation des coûts des évaluations internationales... Certains pays ne pourront pas participer aux prochains cycles d'évaluation, faute de moyens.

Vaut-il mieux donner au plus grand nombre l'opportunité d'évaluer la qualité de son système éducatif, ou sacrifier à la quantité des informations recueillies la qualité des comparaisons internationales ? Les organisations d'évaluations internationales devront sans doute apporter des réponses qui satisferont le tout, et les parties.

2.2.3. L'origine des textes

Nous avons vu qu'au fil du temps, et des incidents, la qualité linguistique du matériel d'évaluation a été l'objet de vérifications de plus en plus scrupuleuses. Nous avons montré que, pour assurer la comparabilité des évaluations, il fallait présenter aux populations impliquées un matériel de test « équi-lisible ». Pour certains, la lisibilité 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, p. 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 tout cas finement, car produits par et pour une culture donnée ?

Warwick B. Elley (responsable d'IEA Reading Literacy 1991) 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 (1994, p. 95).

Nous ne pouvons pas trancher entre la possibilité d'une « lisibilité interculturelle », envisagée comme corollaire d'expériences universelles communes et une « illisibilité pour cause culturelle ». Les évaluations internationales de la lecture que nous avons examinées cautionnent 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 ?

Dès les premières évaluations internationales en lecture, les pays participants ont été invités à soumettre des textes et des items aux différents comités internationaux d'experts en lecture. Ces derniers se chargent de la sélection finale, elle-même soumise à l'approbation des comités nationaux.

Cette structure de base est invariable. Certains éléments conduisent pourtant à penser que la contribution effective des experts nationaux a varié.

Par exemple, les rapports internationaux n'insistent pas tous de la même façon sur la volonté d'inclure les responsables nationaux à différents niveaux d'élaboration et de décision des tests.

> The development of all tests and questionnaires, and decisions about research design, definitions, age levels, methodology, timetabling, and reporting were collective decisions, made by the Steering Committee in consultations with the N[ational] R[esearch] C[oordinators]. [...] Thus, every effort was made to ensure that national circumstances were taken into account in the preparation of instruments and design of the survey. [...]

> The vigilance and competence of the NRCs was crucial to the efficient organization of the whole project [...] (Elley, 1994, p. 4).

On ne trouve pas ce type de discours dans les deux premières évaluations de l'IEA. À l'occasion d'IEA Reading Literacy (1991), Warwick B. Elley ne témoigne pas nécessairement d'un changement dans les pratiques – les responsables nationaux collaboraient déjà lors des premières évaluations – mais son discours reflète la volonté d'exposer, d'expliciter, voire de souligner, l'importance des contributions nationales.

Les commanditaires de IALS et PISA sont des organisations intergouvernementales[72]. La participation active de tous les gouvernements impliqués et, par conséquent, des représentants nationaux, devient dès lors un enjeu politique[73].

> With the exception of Sweden and Ireland, who joined the study late, all countries participating in the first round of IALS data collection also participated in the development of the background questionnaire and the assessement instruments (Kirsch and Murray, 1998, p.15).

Le consortium responsable de PISA prévoit « à de nombreux stades cruciaux du projet » des échanges intensifs avec les Directeurs Nationaux de Projet :

> Cela concerne le développement des plans d'évaluation, la constitution de la batterie d'items ainsi que l'établissement d'indicateurs et d'échelles de compétence (OCDÉ, mai 1998, p. 7).

Pour assurer une réelle collaboration multiculturelle et multinationale, les responsables de PISA recrutent également les experts « matières » dans un grand nombre de pays participants. Dans les évaluations internationales en lecture, ces groupes d'experts ont un rôle prépondérant, puisqu'ils sont chargés de la sélection des textes, de la sélection ou la création des items et du développement des questionnaires (au moins du questionnaire « lecture »). La

---

[72] L'OCDÉ, l'Union Européenne et l'UNESCO pour IALS; l'OCDÉ pour PISA.

[73] « The international Adult Literacy Survey (IALS) was a collaborative effort by seven governments and three intergovernmental oranizations... » (OCDÉ, 1995, p.13)

« Un certain nombre de principes nous ont guidés dans la mise en place du consortium et la préparation de cette offre. Ces principes nous conduisent à attacher une grande importance à des procédures assurant une large collaboration internationale et la participation active des pays dans tous les aspects du développement et de l'exécution du projet » (OCDÉ, mai 1998, p. 7).

composition des groupes d'experts internationaux peut donc être un indice de la variété des points de vue linguistiques et culturels qui président aux choix des supports d'évaluation.

Le tableau qui suit permet de mettre en relation, pour chaque étude, l'origine nationale et linguistique des experts internationaux en lecture, le nombre de pays impliqués dans l'étude et le nombre de langues évaluées.

Nom de l'étude	IEA Reading comprehension	IEA Literature	IEA Reading Literacy	IALS	PISA
Nombre de pays participants	10[74]	15[75]	31	12[76]	32
Comités internationaux	International Reading Comprehension Committee (5 membres)	International Committee for Literature (5 membres)	International Steering Committee (7 membres)	Statistic Canada Educational Testing Service[77]	Groupe fonctionnel d'experts en lecture (10 membres)
Pays représentés au sein du groupe d'experts internationaux	États-Unis (présidence) Iran France[78] Angleterre Pays-Bas	États-Unis (présidence et 1 membre) Royaume-Uni Suède Belgique fr.	Nouvelle-Zélande (présidence) États-Unis (3 membres) Australie (2 membres) Suède	Non applicable.	États-Unis (présidence et 1 membre) Royaume-Uni Canada Pays-Bas Belgique fr. Finlande France Allemagne Japon
Langues[79] représentées au sein des groupes d'experts internationaux /nombre de langues évaluées	4/10	3/7	2/ ? (16 au moins)	Non applicable.	6/ 26

[74] La Belgique francophone et néerlandophone sont considérées comme deux « pays ».
[75] La Belgique francophone et néerlandophone sont considérées comme deux « pays ».
[76] Nous avons inclu la France, qui ne s'est retirée du projet qu'après la campagne de test, et qui a donc pu participer aux travaux préliminaires. Quatre pays ont participé à une deuxième campagne de test en octobre 1995 : l'Australie, la Belgique néerlandophone, la Nouvelle Zélande et le Royaume-Uni.
[77] IALS a la particularité de ne pas avoir créé un Comité international d'experts. « Statistics Canada, the statistical arm of the Canadian government, and Educational Testing Service, the leading private testing organization in the United States, coordinated the development and management of the IALS. These organisations were guided by national research teams form the participating countries, which assisted in developing the survey design » (National Center for Education Statistics, 1998, p. 16)
[78] La France ne participe pas à cette évaluation.
[79] Nous avons compté le nombre de langues au sens large. Nous n'avons pas considéré comme deux langues distinctes l'anglais des États-Unis et celui du Royaume-Uni. D'aucuns diront qu'il s'agit bel et bien de deux langues différentes, et que les chiffres que nous donnons à la dernière ligne du tableau pourraient être revus à la hausse. Ce n'est pas faux, mais cela ne changerait pas nos conclusions quant à la faible diversité linguistique au sein des groupes d'experts.

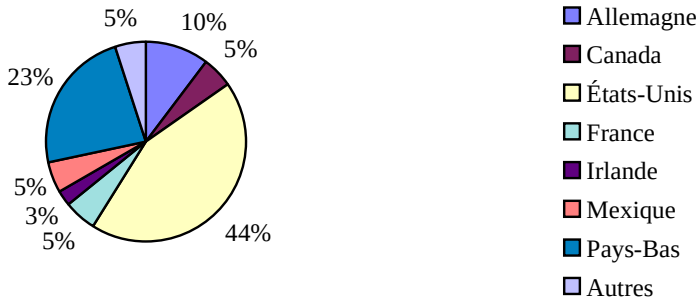
On peut difficilement conclure de ce tableau que les diverses langues et cultures des pays participants aient été largement représentées au sein des groupes d'experts en lecture. Bien sûr, la langue de travail des groupes d'experts a toujours été l'anglais, et on peut imaginer qu'il a été plus facile de communiquer entre pays anglophones. Voyons maintenant dans quelle mesure le choix des textes reflète la diversité linguistique des pays impliqués.

Malheureusement, l'origine linguistique des textes n'est pas toujours précisée dans les rapports internationaux. Pour 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*[80] » (Thorndike, p. 20). IEA Reading Literacy a aussi demandé aux pays de proposer des textes, mais nous n'avons pas pu retrouver leur origine.

Pour IEA Reading Literature, le Comité international a sélectionné quatre textes d'auteurs. Deux sont Américains, le troisième est Espagnol, et le dernier est Belge. Vu le petit nombre de textes prévus pour le test final, le choix du Comité témoigne d'une relative diversité linguistique et culturelle.

Par contre, pour IALS et PISA, nous disposons d'informations précises sur l'origine nationale et/ou linguistique des textes.

IALS : origine nationale des textes

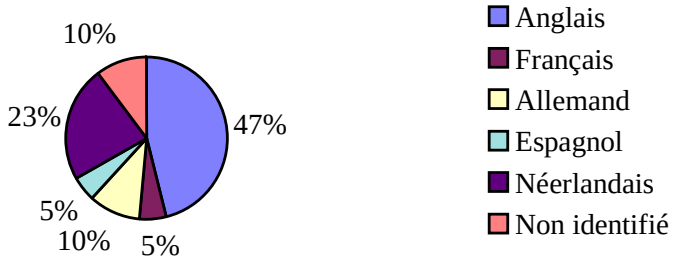


[81]

[80] Nous soulignons.
[81] Source : National Center for Educational Statistics, 1998, p. 19.

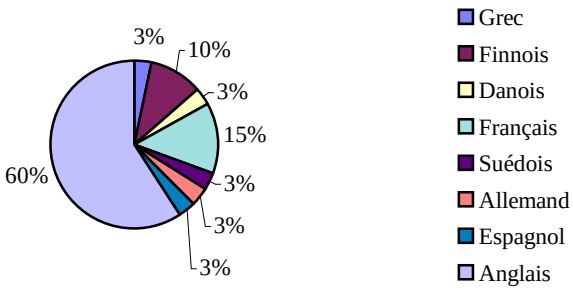
Nous constatons que, pour IALS, on a utilisé des textes fournis par tous les pays participants, à l'exception de la Suède qui, ayant rejoint trop tard les autres pays, n'est pas représentée. Nous nous sommes servie de ces données pour créer le graphique suivant[82] :

IALS : origine linguistique des textes



Les auteurs du rapport technique précisent que « approximately half of these tasks were based on materials from outside North America » (National Center for Educational Statistics, 1998, p. 19). On pourrait aussi souligner l'inverse... et souligner que, malgré la diversité des pays et des langues, la prédominance de l'anglais est manifeste. C'est aussi le cas dans PISA :

Pisa : origine linguistique des textes



Comme pour IALS, on retrouve dans l'évaluation PISA une prépondérance de textes en anglais. Comparativement à IALS, cette suprématie est plus marquée, car le nombre de pays

[82] La catégorie « non identifié » reprend les deux textes canadiens, dont nous n'avons pas présumé l'origine

impliqués dans PISA est bien plus important[83]. Dès lors, la diversité linguistique des populations concernées est moins représentée par l'origine linguistique des textes sélectionnés.

Culturally correct...

Pourtant, le consortium responsable de PISA s'est entouré de nombreux garde-fous pour assurer que les étudiants ne soient pas défavorisés par leur origine culturelle ou linguistique. Ainsi, fait inédit dans l'histoire des évaluations internationales, 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, pour apaiser les craintes exprimées par plusieurs pays d'un risque de biais liés à la provenance du matériel de tests.

> The Cultural Review Panel play an important part in the suit of quality assurance procedures that are employed by PISA to ensure that the assessment instruments :
>
> - are appropriate in the various cultural and linguistic contexts of OECD countries;
> - do not violate accepted cultural values and positions (OCDÉ, Novembre 1999, p. 5)

... ou presque

En regard de l'importance accordée ici au travail du Comité de révision culturelle, le temps qui lui a été octroyé semble dérisoire : deux jours de meeting, pour lequel les documents de travail avaient été envoyés très peu de temps à l'avance.

> The panel understood that every aspect of the project is on a fast track but at the same time, I was instructed to mention this concern about the lack of comprehensiveness of our review (Hambleton, OCDÉ, Novembre 1999, p. 7).

En outre, il semble que les avis du Comité culturel n'aient pas été largement pris en compte. Les items les plus problématiques pour le Comité culturel n'ont pas été supprimés pour le test définitif, et les nombreuses suggestions d'amélioration – tant sur la forme que sur

---

linguistique, et les deux textes répertoriés dans notre source sous la rubrique « autres ».

[83] 32 pays dans PISA, 12 dans IALS.

les adaptations favorisant une meilleure traduction – ne semblent pas avoir donné lieu à des modifications concrètes.

*

* *

Bien sûr, les développeurs de tests ont dû prendre en considération une quantité d'autres paramètres pour sélectionner le matériel du test définitif. Il fallait concilier les avis des experts en lecture et ceux des responsables nationaux, tout en tenant compte de la qualité psychométrique des items et des impératifs quantitatifs à respecter par catégorie de texte.

En fin de compte, il semble que la diversité culturelle et linguistique soient les seuls paramètres qui n'aient pas encore fait l'objet d'un calibrage minutieux... En fait, il s'agit bien d'une question de priorités. Nous venons de le voir, une quantité d'autres contraintes influencent les choix internationaux pour la sélection finale des tests. Si donc on considère comme essentiel le fait d'inclure du matériel provenant de divers horizons linguistiques et culturels, il faut aussi 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 apparaît comme un parent pauvre, c'est peut-être que ces considérations, et la participation active qu'elles impliquent, n'ont pas encore trouvé assez de relais au niveau des pays...

3. Pour des faux diplomatiques ?

On peut rêver d'équité, et dans ce rêve, inclure une plus grande prise en compte des particularités de chacun, tant dans les systèmes sociaux, dans les systèmes éducatifs, que dans les systèmes internationaux d'évaluation des acquis des élèves. Ce désir d'équité se heurte parfois à ses propres limites.

Pour certains, construire des instruments valables pour toutes les communautés linguistiques et culturelles reviendrait à créer des instruments trop " polis " – dans les deux sens du terme, qui finalement ne conviendraient plus à personne. L'équité aurait alors une saveur aseptisée, et ne serait que prétexte au gommage de l'hétérogénéité.

D'autres, à l'inverse, pensent que la prise en compte des particularismes de chacun est le seul moyen de résistance face à la globalisation. Pour eux, les systèmes de concepts, les classifications et les statistiques, les normes de qualité et les critères normatifs d'évaluation en vigueur dans les organisations internationales ou les centres de recherche anglo-américains, placent ceux qui les reçoivent, dans le monde entier, sous la pression d'un mouvement d'uniformisation – pression d'autant plus efficace qu'on ne la perçoit pas comme telle (Schriewer, 1997, p. 119).

Dans la première partie de ce travail, nous avons montré le caractère éminemment politique des évaluations internationales de la lecture. Dans ce contexte, et à l'heure où l'on revendique transparence et démocratie, la prise en compte des diversités n'est-elle pas un corollaire indispensable d'une certaine " démocratie internationale " ?

La deuxième partie a permis de constater combien l'existence de divers points de vue nationaux, linguistiques et culturels n'a cessé de provoquer, au-delà des consensus apparents, des recherches et des avancées méthodologiques, dans l'espoir de tendre à toujours plus d'équivalence.

Toutefois, l'a priori sur lequel se fondent les évaluations internationales, à savoir la possibilité d'évaluer les élèves sur la base d'instruments de test communs, n'emporte pas

l'adhésion de tous. Un projet européen actuellement en cours (CEDE, 2000) tente d'exploiter une autre voie. Partant du principe qu'il est quasi impossible de trouver des instruments d'évaluation qui ne soient culturellement " conditionnés ", et inquiets devant la prépondérance de textes anglais pour évaluer les compétences en lecture des élèves, les concepteurs de ce nouveau projet international vont proposer aux étudiants des textes originaux, non traduits. Ils devront bien sûr résoudre eux aussi de nombreux problèmes méthodologiques, dont le principal enjeu sera de déterminer l'équivalence de la difficulté des tâches évaluées à partir de supports et de questions différentes.

Avec ce projet européen, on sent poindre à nouveau le vieux problème de l'authenticité des évaluations, lié au caractère ontologique et unique entre langue, culture et peuple. Qu'est-ce qui est le plus " juste ", le plus " naturel " ou le plus " authentique " pour les groupes impliqués dans les évaluations ? Des supports " identiques ", susceptibles de pécher par leur manque d'équivalence culturelle et linguistique, ou des supports " authentiques ", risquant d'être incomparables quant aux traits qu'ils évaluent ? Les deux positions permettent de nourrir des débats passionnants, et de faire avancer la recherche... toujours en quête d'une toute relative vérité, dont la valeur, en fin de compte, dépendra des critères qu'on se donne pour établir " l'équivalence ".

> Malgré cela, même si aucun des critères pris individuellement n'est satisfaisant à cent pour cent, nous nous fions généralement à eux pour émettre des hypothèses raisonnables à partir d'une évaluation équilibrée des diverses méthodes de vérification. C'est comme dans un procès, où un témoignage peut paraître non crédible mais où trois témoignages concordants sont pris au sérieux ; un seul indice peut paraître mince, mais trois indices font système. Dans tous les cas, on s'en remet à des critères d'*économie de l'interprétation.* Les jugements d'authenticité sont le fruit de raisonnements persuasifs, fondés sur des preuves vraisemblables même si pas totalement irréfutables, et nous acceptons ces preuves parce qu'il est raisonnablement plus économique de les accepter ... (Eco, 1992, p. 209).

Mais nous devons aussi passer notre temps à les mettre en doute.

4. Bibliographie

Altbach, P. G., Kelly, G. P. (Eds). (1986). *New approaches to comparative education.* Chicago : The university of Chicago Press.

Atash, N., Kapinus B. (1994). Exploring the Possibilities of Constructed - Response Items. In Binkley M., Rust K., Winglee M (Eds), *Methodological Issues in Comparative Educational Studies : The case of the IEA Reading Literacy Study.* Washington, D.C. : U.S. Department of Education, National Center for Education Statistics.

Bennett, R. E. (1993). On the Meanings of Constructed Response. In Bennett, R. E., Ward W. C. (Eds), *Construction Versus Choice in Cognitive Measurement. Issues in Constructed Response, Performace Testing, and Portfolio Assessment.* Hillsdale : Lawrence Erlbaum Associates, IEA.

Bennett, R. E., Ward W. C. (Eds) (1993). *Construction Versus Choice in Cognitive Measurement. Issues in Constructed Response, Performace Testing, and Portfolio Assessment.* Hillsdale : Lawrence Erlbaum Associates, IEA.

Binkley, M., Rust, K., Winglee, M. (Eds) (1994). *Methodological Issues in Comparative Educational Studies : The case of the IEA Reading Literacy Study.* Washington, D.C. : U.S. Department of Education, National Center for Education Statistics.

Bourdieu, P., Passeron, J. - C. (1970). *La reproduction. Éléments pour une théorie du système d'enseignement.* Paris : Les Éditions de Minuit, coll. Le sens commun.

Bourdieu, P., Passeron, J.- C. (1985). *Les héritiers. Les étudiants et la culture.* Paris : Les Éditions de Minuit, coll. Le sens commun.

Bourdieu, P. (2000). *Les structures sociales de l'économie.* Paris : Éditions du Seuil, coll. Liber.

Brown, M. (1998). The tyranny of the International Horse Race. In Slee R., Weimer G., Tomlinson S. (Eds), *School Effectiveness for whom ? Challenges to the School Effectiveness and School Improvement Movement.* London : Auteurs.

CEDE (2000). *National Institute for the Evaluation of the Education System -CEDE. Yearbook 2000. Research projects and activities.* Roma : Auteur.

Commission Européenne (2000). *Les chiffres clés de l'éducation en Europe.* Bruxelles, Luxembourg : Auteur.

Cook, L., Schmitt A., Brown, C. (1999). *Adapting Achievement and Aptitude Tests : A Review of Methodological Issues.* Document présenté à la Conférence Internationale sur l'adaptation de tests. Washington, DC : Auteurs.

Cronbach, L. J., Drenth P.J.(Eds) (1972). *Mental Tests and Cultural Adaptation.* Mouton, The Hague, Paris : Psychological Studies.

Darcovich, N., Scott Murray, T. (1998). Data Collection and Processing. In Scott Murray T., Kirsch I. S., Jenkins L. B, *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.

De Landsheere, G. (1979). *Dictionnaire de l'évaluation et de la recherche en éducation.* Paris : Presses universitaires de France.

De Landsheere, G. (1986). *La recherche en éducation dans le monde.* Paris : Presses Universitaires de France, coll Pédagogie d'aujourd'hui.

Debeauvais, M. (1997). L'influence des organisations internationales sur les politiques nationales d'éducation. In Meuris, G. , De Cock, G. (Eds), *Éducation comparée. Essai de bilan et projets d'avenir.* Paris- Bruxelles : De Boeck & Larcier, Perspectives en éducation.

Demeuse, M. (mars 2000). *Les échelles unidimensionnelles : Les échelles de Thurstone, Likert, Guttman et le modèle de Rasch.* Liège : Publications du SPE, série « Notes techniques ».

Demeuse, M. , Crahay, M. , Monseur, C. (à paraître). Efficacité et équité dans les systèmes éducatifs : les deux faces d'une même pièce ? In *Clés de lecture de regards sur l'éducation n°6. Les indicateurs de l'ODE.* Bruxelles : Ministère de la Communauté Française, Secrétariat général, Direction des Relations internationales.

Downing, J. (1973). *Comparative Reading. Cross-National Studies of Behavior and Process in Reading and Writing.* New York : The Macmillan Company

Eco, U. (1992). *Les limites de l'interprétation.* (trad. de l'italien par Bouzaher M.) Paris, Éditions Grasset & Fasquelle, coll. Le Livre de Poche.

Eco, U. (1994). *La recherche de la langue parfaite dans la culture européenne.* (trad. de l'italien par Manganaro J.-P.) Paris : Éditions du Seuil, coll. Point - Essais.

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.

Gerbet, P. (1994). *La construction de l'Europe.* Paris : Imprimerie nationale Éditions.

Giasson, J. (1990). *La compréhension en lecture.* Bruxelles : De Boeck, Pédagogies en développement, pratiques méthodologiques.

Gould, S. J. (1997). *La mal-mesure de l'homme.* (trad. de l'anglais par Chabert J. et Blanc M.) Nouvelle édition .Paris : Éditions Odile Jacob.

Hall, E. T. (1979). *Au-delà de la culture.* (trad. de l'américain par Hatchuel M.-H.) Paris : Éditions du Seuil, coll. Points - Essais.

Hallak, J. (1998). Education and Globalisation. In *IIEP Contributions* n° 26. Paris : Unesco.

Hambleton, R. K. (1999). I*ssues, Designs, and Technical Guidelines for Adapting Tests in Multiple Languages and Cultures.* Document présenté à la Conférence Internationale sur l'adaptation de tests, Washington, DC.

IEA PIRLS (1999). *Draft Framework and Specifications for the PIRLS Assessment of Reading Literacy. Progress in International Reading Literacy Study.* Chestnut Hill, Ma : PIRLS International Study Center, Boston College.

Klinkenberg, J.-M. (1997). Pour une politique de la langue française. In *La revue nouvelle* (tiré à .part).

Lafontaine, D. (1996). *Performance en lecture et contexte éducatif. Enquête internationale menée auprès d'élèves de 9 et 14 ans.* Bruxelles : De Boek, Pédagogies en développement.

Lafontaine, D. (à paraître). From comprehension to literacy : thirty years of reading assessment. In *Compendium for the General Assembly of INES Tokyo.*

Lundberg, I., Linnakylä, P. (1992). *Teaching reading around the world.* The Hague : IEA.

Management Committee for the National School English Literacy Survey. (1997) *Mapping Literacy Achievement. Results of the 1996 National School English Literacy Survey.* Canberra : Australian Council for Educational Research.

Martin, M. O., Rust, K., Adams, R. J. (Eds). (1999). *Technical Standards for IEA Studies.* Amsterdam : IEA.

Meuris, G., De Cock, G. (Eds) (1997). *Éducation comparée. Essai de bilan et projets d'avenir.* Paris- Bruxelles : De Boeck & Larcier, Perspectives en éducation.

Monseur, C., Demeuse, M. (1998). Les évaluations externes sont-elles efficaces ? In Monseur, C., Demeuse, M. *Pour accroître l'efficacité des systèmes d'enseignement : Recherche des facteurs d'efficacité et étude comparative des dispositifs de pilotage.* Bruxelles : Commission européenne, Direction générale XXII (Jeunesse - Éducation - Culture).

Nettles, M. T. Nettles A.L. (Eds) (1995). *Equity and excellence in educational testing and assessment.* Boston, Dordrecht, London : Kluwer Academic Publishers.

OCDÉ (1992). *Adult Illiteracy and Economic Performance.* Paris : Auteur.

OCDÉ (1995). *Performance standards in education. In search of Quality.* Paris : Auteur.

OCDÉ (1996). *Knowledge Bases for Education Policies.* Paris : Auteur.

OCDÉ (1999). *Programme International pour le Suivi des Acquis des Élèves. Une nouvelle enquête périodique chez les jeunes de 15 ans - évaluer leur préparation à la vie d'adulte.* Paris : Auteur.

OCDÉ (2000). *Mesurer les connaissances et les compétences des élèves. Lecture, mathématique et science : l'évaluation de Pisa 2000.* Paris : Auteur.

OCDÉ - Minister of Industry (1995). *Literacy, Economy and Society. Results of the first International Adult Literacy Study.* Paris - Ottawa : Auteur.

OCDÉ - Minister of Industry (Canada) (1997). *Literacy Skills. For the Knowledge Society.* Paris - Ottawa : Auteurs.

OCDÉ - PISA (Mai 1998). *Procédures de traduction et de vérification des traductions.* Document provisoire préparé par Aletta Grisay. Paris : Auteur.

OCDÉ - PISA (Mai 1998). *Les indicateurs sur les acquis de l'élève de l'ODE. Extrait de l'offre du consortium.* Document présenté au Meeting des Directeurs Nationaux de projet. Paris : Auteur.

OCDÉ - PISA (Octobre 1998). *Consignes pour la traduction du matériel PISA.* Document provisoire préparé par Aletta Grisay. Paris : Auteur.

OCDÉ - PISA (Novembre 1999). *Summary of the main points of the cultural review panel meeting.* Document préparé par Ron Hambleton. Paris : Auteur.

OCDÉ - PISA (2000). *Manuel du directeur national de projet.* Paris : Auteur.

Postlehwaite, T. N., Ross, K.N. (1992). *Effective schools in Reading. Implications for Educational Planners.* La Haye : The Ìnternational Association for the Evaluation of Educational Achievement.

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

Roches, J.-J. (1999). *Théories des relations internationales.* 3e éd. Paris : Montchrestien, coll. Clefs - Politique.

Rumberger R. W. (1994). Les résultats économiques en tant qu'indicateur des résultats de l'enseignement. In Truijnman, A., Bottani, N. (dir.), *Évaluer l'enseignement. De l'utilité des indicateurs internationaux.* Paris : OCDÉ.

Schriewer, J. (1997). Système mondial et réseaux d'interrelation. L'internationalisation de la pédagogie, un problème des sciences comparées de l'éducation. In Meuris, G. , De Cock, G. (Eds), *Éducation comparée. Essai de bilan et projets d'avenir*. Paris-Bruxelles : De Boeck & Larcier, Perspectives en éducation.

Scott Murray 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.

Snow, R. E. (1993). Construct Validity and Constructed-Response Tests. In Bennett, R. E., Ward W. C. (Eds), *Construction Versus Choice in Cognitive Measurement. Issues in Constructed Response, Performace Testing, and Portfolio Assessment.* Hillsdale : Lawrence Erlbaum Associates, IEA.

Suzuki, L. A., Meller, P. J., Ponterotto, J. G. (Eds) (1996). *Handbook of Multicultural Assessment. Clinical, Psychological, and Educational Applications*. San Francisco : Jossey-Bass Publishers.

Thorndike, R. L. (1973). *Reading Comprehension Education in Fifteen Countries*. Stockholm : Almquist and Wiksell.

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., Hagen, E. P. (1991). *Measurement and Evaluation in Psychology and Education.* (5e ed.) New York - Toronto : Maxwell Macmillan International Editions.

Truijnman, A., Bottani, N. (dir.). (1994). *Évaluer l'enseignement. De l'utilité des indicateurs internationaux.* Paris : OCDÉ.

Tuijnman, A.C., Postlethwaite, T. N. (Eds). (1994). *Monitoring the standards of Education.* Oxford : Pergamon.

Von Humblodt, W. (2000). *Sur le caractère national des langues et autre écrits sur le langage.* (trad. par Thouard D.) Paris :Éditions du Seuil, Coll. Points - Essais.

West, A., Edge, A., Stokes, E. (1999). *Secondary education across Europe : curricula and school examination systems.* Londres : Center for Educational Research, DG 12, European commision.

Participation des pays lors d'évaluations internationales

Tableau 1 : *Participation de la Belgique et de la Communauté française aux études de l'IEA.*

				Niveaux scolaires			
Disciplines	***Nom***	***Date***	***Nombre de pays***	***Pré-scolaire***	***Primaire***	***Début secon-daire***	***Fin secon-daire***
Mathématiques	FIMS	1965	12			B	B
	SIMS	1981	20			B	B
	TIMSS	1995	41			CFB	
	TIMSS-R	1999					
Sciences	FISS	1971	19		B	B	B
	SISS	1985	26				
	TIMSS	1995	41			CFB	
	TIMSS-R	1999					
Langue de l'enseignement	Reading 1	1971	15		B	B	B
	Reading 2	1991	32		CFB	CFB	
	Littérature	1971	10			B	B
	Composition	1985	13			B	B
Education civique	Civic 1	1971	10				
	Civic 2	1999	24			CFB	CFB
Langues Etrangères	Anglais	1971	10			B	B
	Français	1971	8				
Enseignement préscolaire	PPP	1987-1992	17	CFB			
Informatique	COMPED	1989	21		CFB	CFB	CFB
	SITES	1998	30			CFB	CFB
Lecture	PIRLS	2001	?		?		

Participation des pays lors d'évaluations internationales

Le tableau suivant reprend, pour chaque évaluation internationale envisagée[84], les pays ou systèmes éducatifs impliqués. Les cases colorées indiquent la participation du pays. Le code de couleur correspond aux domaines évalués.

[84] Nous n'avons pas retenu les trois études suivantes : Classroom Environmental Study, PPP (enseignement préscolaire et Pirls. Pour cette dernière, nous n'avons pas encore les données. Nous avons exclu des deux premières, car elles ne s'adressent pas à des élèves du secondaire. Les études retenues évaluent au moins une population d'élèves du secondaire. Nous n'avons pas non plus envisagé ici les évaluations internationales destinées aux adultes (IALS par exemple).

Participation des pays lors d'évaluations internationales

Pays \ Etude	Pilot	FIMS[85]	FISS	Reading Comprehension	Reading Literature	Anglais[86]	Français	Civic 1	SIMS	SISS	Composition	COMPED[87]	Reading Literacy	TIMSS[88]	Civic 2 (phase 2)[89]	SITES[90]	TIMMSS-R[91]	PISA	Participation /Pays
	1959	1965	1971	1971	1971	1971	1971	1971	1981	1985	1985	1989	1991	1995	1995	1998	1999	2000	
Afrique du Sud																			3/18
Albanie																			1/18
Allemagne									[92]	[93]	[94]								13/18
Angleterre								[95]											12/18
Argentine																			1/18
Australie								[96]											9/18

85 Source : Comber, L.C. et Keeves, John P., *Science Education in Nineteen Countries*, .Stockholm, Almqvist & Wiksell, 1973

86 Pour ces études, les données que nous possédons ne sont pas entièrement fiables. Source : Keeves, John P. *The world of School Learing – Selected Key Findings from 35 years of IEA Reasearch*, IEA, 1994.

87 Source : Pelgrum Willem J. et Plomb Tjeerd (Ed.) *The IEA Study of Computers in Education : Implementation of an Innovation in 21 Education Systems*, IEA, 1993.

88 Source : IEA, *Third International Mathematics and Science Study*, IEA, 1997.

89 Source : www.iea.nl

90 Source : Pelgrum Willem J. et Anderson Ronald E. (Ed.) *SITES – ICT and the Emerging Paradigm for Fife Long Learning*, IEA, 1999.

91 Source : www.iea.nl

92 Deux ans après les autres pays.

93 Jusqu'à cette date, seule la RFA participe aux évaluations.

94 A partir de cette étude, l'Allemagne a deux systèmes éducatifs.

95 Angleterre et Wales.

96 L'Australie effectué cette étude après les autres pays.

Participation des pays lors d'évaluations internationales

Autriche																			3/18
Belgique Fl.																			13/18
Belgique Fr.																			16/18
Botswana																			2/18
Brésil																			1/18
Bulgarie																			4/18
Canada																			9/18
Chili																			8/18
Chine																			4/18
Chypre																			6/18
Colombie																			2/18
Corée (République de)																			4/18
Danemark																			6/18
Ecosse																			7/18

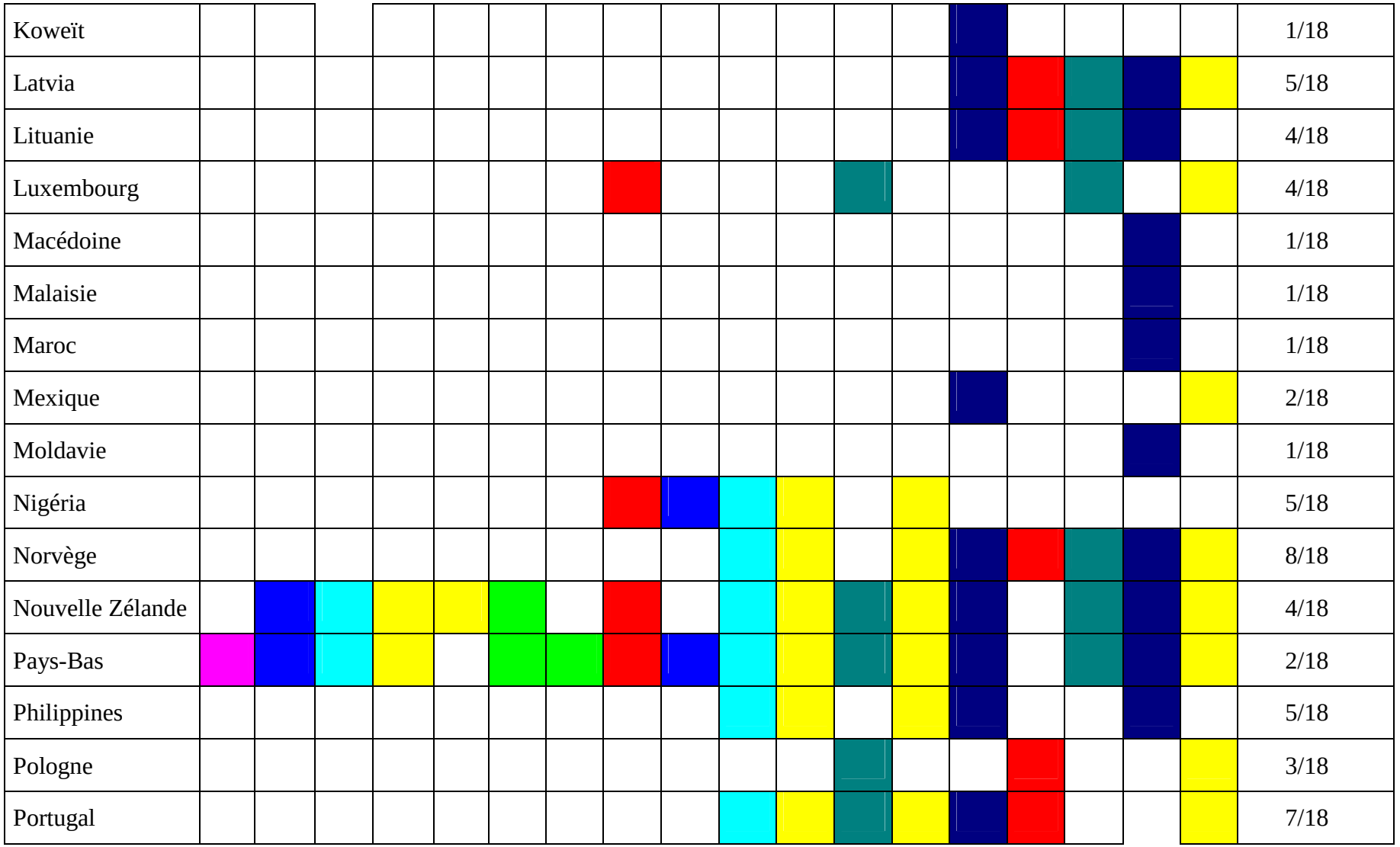
97 Colombie britanique et Ontario.

98 Ontario et Québec.

Participation des pays lors d'évaluations internationales

Espagne																			5/18
Etats-Unis																			16/18
Finlande																			4/18
France																			10/18
Grèce																			6/18
Hongkong																			8/18
Hongrie																			15/18
Inde																			4/18
Indonésie																			5/18
Iran																			7/18
Irlande																			5/18
Islande																			5/18
Israël																			11/18
Italie																			14/18
Japon																			9/18
Jordanie																			1/18

Participation des pays lors d'évaluations internationales



Participation des pays lors d'évaluations internationales

République slovaque																		3/18
République tchèque																		5/18
Roumanie																		3/18
Russie (fédération)																		5/18
Singapour																		6/18
Slovénie																		8/18
Suède																		14/18
Suisse																		5/18
Swaziland																		1/18
Thaïlande																		7/18
Trinidad et Tobago																		2/18
Tunisie																		1/18
Turquie																		1/18

Participation des pays lors d'évaluations internationales

Vénézuela																		2/18
Zimbabwé																		2/18
Total participants	13	19	19	15	10	11	9	20	10	21	30	21	31	46	28	28	40	34	

Les couleurs utilisées renvoient aux différents domaines d'évaluation :

Etude pilote dans différents domaines
Mathématiques
Sciences
Langue d'enseignement
Langue étrangère
Education civique
Informatique
Math et sciences

Participation des pays lors d'évaluations internationales