Principes méthodologiques et techniques des enquêtes internationales

Dominique Lafontaine
Université de Liège

Gilles Raîche
Université du Québec à Montréal

Mots clés : Cadre d'évaluation, élaboration des items, population et échantillon, plan d'évaluation, théorie de la réponse à l'item, échantillonnage matriciel, méthodes d'estimation, qualité des mesures

Cet article vise à expliquer les principes méthodologiques sous-jacents à l'élaboration et à la mise en œuvre des enquêtes internationales dans le domaine de l'éducation. Plus spécifiquement, ce sont les épreuves de rendement disciplinaires qui sont abordées et non pas les questionnaires contextuels. À cette fin, le développement du cadre de référence, le choix des tâches et des items, l'essai de terrain, le plan d'évaluation ainsi que la définition de la population de référence et le plan d'échantillonnage sont décrits. De plus, les aspects psychométriques de base sont abordés : modélisations issues de la théorie de la réponse à l'item utilisées ; méthodes d'estimation des paramètres et particularités propres à ces enquêtes ; qualité des mesures obtenues.

Key words: Framework, test development, population definition and sampling, item response theory, matrix sampling, estimation methods, measurement quality

This paper explains methodological principles specific to the elaboration and application on international educational surveys. More specifically, performance test construction is considered but not contextual questionnaires. To this end, the development of the reference context, the choice of tasks and items, the population definition and the sample design are described. Also, psychometric aspects are explained: item response theory models; parameter estimation and specific characteristics of these surveys; measurement quality.

PALAVRAS-CHAVE: Quadro de avaliação, elaboração de itens, população e amostra, plano de avaliação, teoria da resposta ao item, amostragem matricial, métodos de estimativa, qualidade da medição

Este artigo visa explicar os princípios metodológicos subjacentes à elaboração e à realização de inquéritos internacionais no domínio da educação. Mais especificamente, são abordadas as provas de rendimento disciplina e não os questionários contextuais. Neste sentido, são descritos os seguintes aspetos: o desenvolvimento do quadro de referência, a escolha das tarefas e dos itens, o plano de avaliação e a definição da população de referência e o plano de amostragem. Além disso, os aspetos psicométricos de base são abordados: modelizações da teoria de resposta ao item utilizada; métodos de estimativa de parâmetros e particularidades próprias destes inquéritos; qualidade da medição obtida.

Note des auteurs – Toute correspondance peut être adressée comme suit: Dominique Lafontaine, Université de Liège, Département Éducation et Formation, Sart Tilman, B32, B4000 Liège, Belgique, téléphone: ++32 4 366 20 97, télécopieur: ++32 4 366 28 55, ou Gilles Raîche, Université du Québec à Montréal, Faculté des sciences de l'éducation, Département d'éducation et de pédagogie, Pavillon Paul-Gérin-Lajoie, bureau N-6250, C.P. 8888, succ. Centre-ville, Montréal, QC, Canada, H3C 3P8, téléphone: (514) 987-3000 poste 1712, télécopieur: (514) 987-4608 ou par courriel à aux adresses suivantes: [dlafontaine@ulg.ac.be] ou [raiche.gilles@uqam.ca].

Introduction

L'article se centrera sur les enquêtes comparatives internationales menées à l'initiative de l'IEA (Association internationale pour l'évaluation du rendement scolaire [http://www.iea.nl]) et de l'Organisation de coopération et de développement économiques (OCDE [http://www.pisa/oecd.org]). Dans un premier volet, l'article se centrera sur quelques grands principes méthodologiques sous-jacents à l'élaboration et à la mise en œuvre des enquêtes internationales dans le domaine de l'éducation. Le développement du cadre de référence, le choix des tâches et des items, l'essai de terrain, le plan d'évaluation ainsi que la définition de la population de référence et le plan d'échantillonnage seront ainsi décrits. Dans un second volet, plus technique, l'article abordera les modélisations issues de la théorie de la réponse à l'item utilisées dans les enquêtes internationales ainsi que la question de la qualité des mesures obtenues. Tout au long de l'article, l'accent sera mis sur les évolutions qui sont intervenues dans le domaine, surtout depuis qu'on recourt aux modèles de la réponse à l'item. Quelques pistes d'amélioration seront enfin suggérées en guise de conclusion.

Comment est élaborée une épreuve internationale ?

Élaboration d'un cadre de référence

De nos jours, préalablement à toute épreuve internationale, un cadre d'évaluation *(framework)* est défini (Mullis et al., 2005 ; Mullis et al., 2009 ; OCDE, 2009b). Il n'en a pas toujours été ainsi. Les études internationales les plus anciennes, menées sous les auspices de l'IEA (voir [http://www.iea.nl]), entre les années 1970 et 1995, ne comportaient pas de cadre de référence formalisé. À cet égard, les enquêtes TIMSS (Robitaille et al., 1993) et PISA (OCDE, 1999) marquent un véritable tournant.

Depuis lors, toute épreuve internationale comporte ce document « fondateur » qu'est le cadre de référence pour l'évaluation. Il revient généralement au groupe d'experts internationaux du domaine évalué de produire une première version de ce cadre de référence, qui est largement diffusée et mise en débat auprès des représentants des systèmes éducatifs ou des gestionnaires

nationaux de l'enquête, ensuite révisée pour tenir compte des remarques avant d'arriver à une version définitive qui recueille le consensus des pays participant à l'étude.

Typiquement, un cadre de référence envisage les aspects suivants:

- Quel est l'objet de l'évaluation? Quelle définition se donne-t-on de l'objet? À quels modèles théoriques fait-on référence pour évaluer ce domaine?
- Comment se structure le domaine que l'on veut évaluer? Quels aspects, quelles compétences, quels contenus l'évaluation va-t-elle prendre en compte?
- Quelles modalités d'évaluation seront utilisées?
- Comment présenter et interpréter les résultats de l'évaluation? Existe-t-il une compétence représentée par un score global sur une échelle ou les compétences sont-elles multidimensionnelles, appelant plusieurs scores distincts, plusieurs sous-échelles?

Les études antérieures à TIMSS 1995 et PISA 2000 ont sans doute élaboré de tels documents à usage interne; le fait que ces documents soient aujourd'hui destinés à être rendus publics et à faire l'objet de débats a incontestablement entraîné un effort de clarification et de planification accru. Les études contemporaines sont ainsi davantage assises sur le plan théorique qu'elles ne l'ont été antérieurement lorsqu'une approche plus empirique prédominait[1].

En lecture, pour ne prendre qu'un exemple dans PISA, à côté de l'échelle globale définie tous items confondus, le cadre de référence prévoit de rapporter les résultats sur trois sous-échelles spécifiques, *retrouver l'information, interpréter le texte et réfléchir sur le texte.* Cette décision impose non seulement de disposer d'un nombre suffisant d'items pour chacune de ces sous-échelles, mais aussi d'un éventail d'items de difficulté variée couvrant l'étendue de l'échelle, du plus simple au plus complexe.

On pourrait se demander, étant donné les corrélations très élevées entre les sous-échelles (corrélations égales ou supérieures à 0,90[2]), quel est l'intérêt de rapporter ainsi les résultats sur plusieurs sous-échelles[3]. Plusieurs raisons peuvent être invoquées, dont la principale touche à la finesse du diagnostic posé. En dépit des corrélations élevées, certains systèmes éducatifs peuvent présenter des points forts ou des faiblesses spécifiques sur certaines des sous-échelles. Ainsi, dans PISA, pour ce qui est de la lecture, on observe généralement que les pays de tradition latine (Espagne, Portugal, Grèce, pays d'Amérique du Sud) sont relativement plus performants sur l'échelle *Réfléchir et évaluer* que sur l'échelle *Retrouver de l'information*. Ces points forts et faiblesses peuvent, à l'échelle des pays, être mis en relation avec les curriculums et déboucher sur des pistes d'amélioration plus spécifiques que ne le permettrait une échelle globale de lecture.

Choix des tâches et des items

La question du choix des tâches et des items sera envisagée successivement sous deux angles: un angle général (quels types de tâches ou d'items pour ce genre d'épreuves?) et l'angle plus particulier de la production et de l'origine des tâches et des items (qui élabore ceux-ci? d'où proviennent-ils?).

Les principes généraux présidant au choix des tâches et des items sont d'emblée fixés, on l'a vu, par le cadre de référence. Ainsi, si l'on prend l'exemple des épreuves PISA, la définition qui est donnée de la culture mathématique ou scientifique conditionne largement la nature des tâches. Ainsi, on ne trouvera pas dans PISA, à la différence de ce que l'on peut observer dans TIMSS, de questions où l'on demande d'effectuer des calculs ou d'appliquer des formules, en dehors d'un contexte. Les tâches dans PISA sont toujours présentées dans un contexte et relèvent donc de la résolution de problèmes au sens large[4].

Le cadre de référence définit aussi le type de support – papier-crayon ou support électronique – sur lequel les items seront administrés[5]. Ainsi, dans PISA 2009, un ensemble de tâches et d'items évaluent la manière dont les élèves lisent « en ligne »; une vingtaine de pays ont décidé de participer à ce volet optionnel de l'étude.

Les genres et types de textes qui servent de support dans le domaine de la lecture sont également définis dans le cadre de référence, qui précise les équilibres à respecter: combien de textes continus (prose) et non continus (mêlant de la prose et des graphiques, tableaux, listes, etc.), combien de textes narratifs, argumentatifs, informatifs, etc.

Enfin, le format des questions (question à choix multiple, questions ouvertes à réponse courte et à réponse élaborée) et les proportions à respecter figurent également dans le cadre de référence, dans une table de spécifications où sont croisées les différentes dimensions à prendre en considération et les équilibres à respecter pour éviter tout biais de construction. Si, par exemple, un processus était systématiquement associé à un format de question, il en résulterait une interaction processus-format peu opportune[6].

À titre d'exemple, voici comment dans PIRLS 2006 se distribuent les items selon l'objectif de lecture, les processus de compréhension et le format de question.

Tableau 1
PIRLS 2006 : Répartition des items selon l'objectif de lecture, les processus de compréhension et le format de question

Objectif de lecture	*Nombre total d'items*	*Nombre de questions à choix multiple*	*Nombre de questions à réponse construite*	*Nombre maximum de points*
Lire pour l'expérience littéraire	64	34	30	85
Lire pour acquérir et utiliser des informations	62	30	32	82
Total	**126**	**64**	**62**	**167**

	Processus de compréhension	*Pourcentage d'items*	*Nombre total d'items*	*Nombre de questions à choix multiple*	*Nombre de questions à réponse construite*	*Nombre maximum de point*
Processus 1	Retrouver et prélever des informations explicites	22	31	19	12	36
Processus 1	Faire des inférences simples (directes)	28	43	29	14	47
Processus 2	Interpréter et intégrer des idées et des informations	37	34	6	28	61
Processus 2	Examiner et évaluer le contenu, la langue et les éléments textuels	14	18	10	8	23
	Total	**100**	**126**	**64**	**62**	**167**

Une question souvent sensible dans les enquêtes internationales concerne la provenance linguistique et culturelle des items. D'aucuns pensent que les tests sont un pur produit de la culture anglo-saxonne, susceptible de biaiser les résultats. Il est indéniable que les enquêtes internationales contemporaines les plus connues (PISA, PIRLS et TIMSS) sont dirigées par des consortiums (PISA) ou un centre (PIRLS et TIMSS) situé dans le monde anglo-saxon :

ACER (*Australian Council of Educational Research)* pour PISA et *Boston College* pour PIRLS. Une différence sépare cependant les deux études. Par exemple, pour PISA 2012, le consortium dirigé par ACER regroupe plusieurs centres de recherche, européens notamment (UL - Université du Luxembourg, IPN - *Leibniz Institute for Science Education* en Allemagne, ETS - Educational Testing Service et Westat aux États-Unis, DIPF - *Deutsches Institut für Internationale Pädagogische Forschung,* Service d'Analyse des systèmes et des pratiques d'enseignement de l'Université de Liège en Belgique, ainsi que NIER - *National Institute for Educational Policy Research* au Japon). Au-delà de leur composition, plus multiculturelle que certains ne l'imaginent, les consortiums ou centres responsables de la mise en œuvre des enquêtes contemporaines ont mis en place des dispositifs pour recueillir des tâches provenant de différentes langues et cultures.

À titre d'exemple, voici les principales étapes du processus mis en œuvre pour recueillir des items dans le cadre de PISA 2006 pour le domaine majeur des sciences, telles que décrites dans le *PISA 2006 Technical Report* (OCDE, 2009a)[7]. Le processus s'appuie sur un ensemble de lignes directrices préparées au début de projet : le cadre d'évaluation, une table spécifiant les caractéristiques des items, une présentation des aspects influençant la difficulté des items et des exemples d'items.

Première phase : développement local

Les équipes membres du consortium développent un ensemble d'items (accompagnés de leur guide de correction) dans leur langue nationale (français, anglais, allemand, japonais, etc.). Ces items font l'objet de laboratoires cognitifs individuels ou par petits groupes (les étudiants sont invités à réfléchir à voix haute en répondant aux items ou sont interviewés après). Des ensembles d'items plus larges sont ensuite prétestés dans différentes classes d'élèves de 15 ans dans des établissements du pays où ils ont été développés. À chacune des étapes, les items sont révisés en profondeur, ou éliminés le cas échéant.

Deuxième phase : développement international

À ce stade, chaque unité (stimulus + items) est examinée par une des équipes du consortium autre que celle responsable du développement initial. On examine ici avec un soin particulier comment les items se comportent dans des cultures et contextes nationaux différents. Ceci conduit à écarter des items ou même des unités. Les unités que l'on conserve feront ensuite l'objet d'un test pilote à plus large échelle (principalement en Australie) et circuleront dans les centres nationaux des pays pour révision.

Proposition d'items par les pays

En vue d'assurer une diversité culturelle et contextuelle, la contribution des pays à l'élaboration du test est également sollicitée. Ceux-ci reçoivent un ensemble de lignes directrices, comme au début du processus. Les unités peuvent être soumises dans une dizaine de langues différentes. Pour PISA 2006, pas moins de 155 unités ont ainsi été fournies par les pays ; 40 de ces unités ont été, après examen et révision, incluses dans les livrets d'items qui seront soumis aux centres nationaux pour révision. À l'issue de la sélection, 37 % des unités retenues pour l'essai de terrain sont issues des soumissions nationales. « Au total, 11 des 29 unités retenues pour le test définitif proviennent des soumissions nationales faites par huit pays. Au final, les 29 unités ont été développées dans 12 pays, dans huit langues différentes ; huit unités ont été à l'origine développées en langue anglaise » (OCDE, 2009a, p. 43).

Révision des items par les pays

À ce stade, chaque centre national reçoit un ensemble de plusieurs centaines d'items (492 items pour PISA 2006) qu'ils doivent évaluer – en s'entourant de panels d'experts – sur différents critères, tels que leur pertinence et leur intérêt pour des élèves de 15 ans, leur lien avec le curriculum national ou leur authenticité. Les gestionnaires nationaux doivent veiller à relever tout problème d'ordre culturel ou linguistique (possibles difficultés de traduction par exemple).

Les unités et items sont aussi évalués, en parallèle, par les groupes internationaux d'experts (ici de sciences) qui se réunissent régulièrement. À l'issue de ces divers cycles de développement, consultations et révisions, deux versions – dites sources (une en anglais et une en français) des unités – sont préparées et distribuées aux pays pour traduction et adaptation locale. Ce processus de traduction/adaptation est aussi étroitement cadré. Les unités ainsi traduites sont ensuite prétestées à large échelle dans tous les pays participants lors de l'essai de terrain (voir point « Rôle de l'essai de terrain », ci-après).

Le panorama dressé à larges traits ci-dessus illustre bien en quoi l'élaboration d'un test international est une entreprise de coopération. Plusieurs types d'acteurs y sont impliqués – développeurs de tests membres du consortium et nationaux, experts du domaine évalué, représentants des pays participants, gestionnaires nationaux du projet PISA et panels d'experts nationaux – et s'associent pour contribuer à produire un test qui soit culturellement diversifié,

conforme à la fois aux attentes du cadre de référence et aux exigences de qualité et qui ne heurte pas les sensibilités culturelles ou nationales d'autres pays participants.

Durant le processus de révision des items, les panels d'experts des pays sont invités à indiquer par une note chiffrée quelle priorité ils donneraient à chacun des items pour leur inclusion dans le test. On peut supposer que ces items « prioritaires » représentent pour chaque pays les items les plus proches de leurs préoccupations curriculaires ou autres. Adams, Berezner et Jakubowski (2010) ont étudié quel serait le classement des pays dans PISA 2006 si, pour chaque pays, le test était composé de « leurs » items préférés. L'analyse montre que les variations de rang sont légères, ce qui permet d'exclure l'hypothèse d'un biais culturel massif en faveur des pays anglo-saxons.

> Les résultats montrent une cohérence remarquable du rang occupé par les pays quand on compare la position moyenne de chaque pays fondée sur différents ensembles d'items préférés par les pays avec le rang qu'occupe ce même pays dans le rapport initial PISA 2006 [...]. En général, très peu de pays améliorent ou diminuent systématiquement leur classement. Les changements de classement pour ces pays sont relativement mineurs et leur ampleur comparable à l'erreur d'échantillonnage. Pour la plupart des pays, de telles modifications de classement ne sont pas observées. (Adams, Berezner, & Jakubowski, 2010, p. 12).

Une autre question sensible des enquêtes internationales est la question des traductions et des adaptations tolérées. Faute de place, nous ne l'aborderons pas ici. Nous nous contenterons de souligner que ce processus est très étroitement cadré dans les enquêtes contemporaines et que les standards de vérification sont considérablement plus sévères qu'ils ne l'étaient par le passé (Lafontaine & Simon, 2008). Nous renvoyons le lecteur intéressé par ces questions aux rapports techniques des enquêtes (par exemple Grisay, 2003 ; Grisay et al., 2007 ; Martin, Mullis, & Kennedy, 2007 ; OCDE, 2009a).

Rôle de l'essai de terrain

Toute enquête internationale est nécessairement précédée, un an auparavant, d'un essai de terrain. Il s'agit là d'une étape aussi essentielle qu'incontournable, qui consiste à répéter, dans tous ses aspects et dans tous les pays participants, l'enquête définitive. Ainsi, il s'agit bien entendu d'administrer à un nombre important d'élèves toutes les unités et items du test retenus à ce stade, mais aussi les questionnaires aux élèves et chefs d'établissement. Il s'agit aussi de tester les procédures de recrutement des administrateurs de

tests dans les écoles et de les former à cette fin. Des correcteurs expérimentés doivent être recrutés, formés, les questions ouvertes doivent être corrigées, la fidélité des correcteurs évaluée, etc. Bref, on l'aura compris, l'essai de terrain est une répétition générale de l'opération définitive, dont le but est de vérifier que tout fonctionne bien et surtout d'apporter les améliorations, précisions, modifications qui se révéleraient nécessaires pour tous ou pour certains des pays en particulier.

La place nous manque pour détailler tous les aspects concernés par cet essai de terrain. Nous nous contenterons de rappeler, pour ce qui concerne plus particulièrement le test cognitif, que les items sont soumis à une série d'analyses destinées à vérifier leurs qualités psychométriques dont le détail se trouve à la section 3 de ce texte. Les items qui ne satisfont pas aux exigences de qualité sont, à la suite de ces analyses, révisés ou éliminés. Le temps moyen nécessaire pour répondre aux items est aussi estimé ; sur cette base, il est possible de déterminer le nombre d'items qu'il est raisonnable d'inclure dans chaque bloc (ensemble d'unités comportant un stimulus et des items).

Plan d'évaluation

Toutes les enquêtes internationales récentes (depuis l'enquête IEA-PIRLS de 1991) ont recours aux modélisations issues de la théorie de la réponse à l'item (voir la section « Modèles de mesure utilisés », ci-après). Ceci rend possible l'utilisation de plans d'évaluation complexes avec rotation de livrets (*booklets*). Concrètement, cela signifie que tous les élèves d'une même classe (dans les études IEA) ou d'un même établissement (dans PISA) n'ont pas sous les yeux le même livret[8] ; il en existe différentes formes (à titre indicatif, 13 dans PIRLS 2011, 13 dans PISA 2006 et jusqu'à 21 dans PISA 2009), élaborées selon des plans stricts qui permettent, au départ de livrets distincts, mais présentant suffisamment d'items communs, de produire des scores comparables. Au total, davantage d'items peuvent ainsi être utilisés, davantage de compétences évaluées sans que chaque individu testé ne doive pour autant subir une charge cognitive accrue. Le domaine peut être mieux couvert et les résultats peuvent être présentés non seulement de manière globale, mais aussi sur des sous-échelles distinctes. Les différents items sont répartis dans des livrets qui ont une partie commune (par exemple 15 items sur 60) et une partie variable (les 45 autres items varient d'un carnet à l'autre). La technique d'ancrage (voir la section « Modèles de mesure utilisés », ci-après) rend possible la comparaison des scores d'élèves qui n'ont pas passé exactement le même ensemble d'items.

Pour élaborer le plan d'évaluation d'une épreuve internationale, il est tenu compte de différents éléments :

- Chaque bloc d'items doit apparaître dans les différentes positions possibles (début, milieu, fin de séance), ceci afin de neutraliser l'effet de fatigue, et un même nombre de fois. Dans PISA 2000, par exemple, les items de lecture qui ont été retenus pour l'ancrage en 2003 apparaissaient quasi tous en 1re, 2e ou 3e position dans les livrets, ce qui est loin d'être une situation optimale. Ces items ont en effet une probabilité de réussite plus élevée que les items apparaissant plus loin dans les carnets ; en outre, pour que l'ancrage s'effectue dans de bonnes conditions, il conviendrait que les items d'ancrage apparaissent dans les mêmes positions que lors des cycles précédents, ce qui fait peser une contrainte importante sur le design du cycle en cours ;
- Chaque bloc doit prendre à peu près le même temps de réponse, en sorte que la proportion d'items non atteints soit à peu près identique d'un bloc à l'autre. Ceci est important vu que certains blocs sont amenés à servir d'ancrage d'un cycle à l'autre (dans PISA, PIRLS, TIMSS, etc.) ;
- Chaque paire de blocs consécutifs n'apparaît qu'une seule fois (c'est le principe de l'échantillonnage matriciel) ;
- Il importe de contrôler les effets de contexte et de minimiser, dans la mesure du possible, les changements de domaine à l'intérieur d'un même livret (ceci concerne surtout PISA). Dès lors que PISA évalue trois domaines (un majeur et deux mineurs par cycle), certains livrets comportent des blocs de lecture, de mathématique ou de sciences. Passer d'un domaine à l'autre à l'intérieur d'un même test mobilise des ressources cognitives et peut influencer les résultats des élèves. Ainsi, il est établi que ce type de transition peut varier en fonction de l'aptitude de l'élève et de son sexe notamment (en interaction avec le domaine) (Monseur, communication personnelle, 20 juin 2010).

Population de référence et échantillonnage

Population de référence

Une enquête internationale doit, avant toute chose, préciser la population de référence à laquelle elle va s'intéresser. Plusieurs possibilités existent à cet égard, dont les deux principales sont:

- une population définie exclusivement par son âge (population dite à âge constant); c'est le cas de PISA, qui s'intéresse aux élèves âgés de 15 ans, sans prendre en considération l'année d'études qu'ils fréquentent; en Communauté française de Belgique, par exemple, on trouve dans l'échantillon des élèves de la 1re à la 6e année du secondaire, le mode étant la 4e secondaire où se trouvent les élèves à l'heure dans leur parcours scolaire;
- une population définie principalement par son niveau d'études (population dite à niveau d'études constant), avec des indications ou limites relatives à l'âge des élèves, variables selon les études (il en existe de nombreuses variantes); dans PIRLS par exemple, la population de référence est définie comme le grade où se trouvent les élèves ayant reçu quatre années d'enseignement, le point de départ étant la première année où débute l'enseignement formel de la lecture, de l'écriture et des mathématiques. Au moment de l'évaluation, l'âge moyen des élèves ne peut être inférieur à 9,5 ans. Aucune limite supérieure d'âge n'est fixée. À l'intérieur des groupes ainsi définis, des variations d'âge sont inévitables. Elles sont notamment dues aux politiques de redoublement et au fait que dans certains systèmes, il faut avoir 6 ou 7 ans accomplis pour entrer en 1re primaire.

Une solution élégante pour réduire ce type d'inconvénients est de tester, comme l'a fait TIMSS en 1995, deux années d'études adjacentes (grades 7 et 8 en l'occurrence).

Il importe de souligner que si le choix de la population de référence comporte des aspects techniques, c'est d'abord en référence aux objectifs de l'enquête qu'il est posé. Ainsi, si PISA a opté pour une définition à âge constant, c'est parce que l'objectif de PISA est d'évaluer le niveau d'acquis des jeunes au moment où ils approchent de l'âge de fin de scolarité obligatoire dans la majorité des pays (16 ans). Tester une année d'études n'aurait pas été pertinent dans cette perspective.

Pour les systèmes éducatifs qui pratiquent peu le redoublement, le choix d'une approche à âge ou à niveau d'études constant change en réalité peu de choses, puisque les élèves d'un même niveau ont le même âge. Pour les systèmes qui recourent fréquemment au redoublement, comme la Belgique, la France, le Luxembourg, l'approche à âge constant présente des inconvénients de divers ordres. Les élèves étant répartis sur plusieurs niveaux d'étude, la dispersion des résultats a tendance à être plus élevée, et les indicateurs d'équité – ceux liés au genre notamment – peuvent s'en trouver affectés, du fait que les garçons redoublent davantage que les filles. D'un point de vue politique, il est généralement plus difficile à faire admettre que l'on puisse comparer des élèves qui ne sont pas dans le même niveau d'études.

Échantillon et couverture de la population de référence

Dans les enquêtes internationales, des règles strictes sont à respecter en matière d'échantillon, de participation et de couverture de la population de référence. La sélection de l'*échantillon* répond à un certain nombre de règles, en termes de taille, de stratification et de procédure de sélection, afin d'en assurer la représentativité. Chaque pays est tenu de fournir des documents et des données statistiques reprenant les caractéristiques de son système éducatif et de sa population scolaire auprès du centre international chargé de sélectionner les échantillons des pays participants (Westat pour PISA, DPC Hambourg pour les études de l'IEA). De la même façon, les taux de participation des écoles sélectionnées sont contrôlés et doivent être strictement respectés. Des procédures d'adjudication des données valident *a posteriori* la qualité des échantillons. Dans PISA, les pays qui ne respectent pas l'ensemble des critères ne sont pas présentés dans les tableaux de résultats des rapports internationaux: ce fut notamment le cas des Pays-Bas en 2000. Pour les enquêtes de l'IEA, les éventuels problèmes sont signalés dans les tableaux de résultats internationaux: dans PIRLS 2006, par exemple, la Norvège avait un taux de participation après remplacement au-dessous du seuil requis (71 %) et ceci est mentionné dans les tableaux.

Des exigences doivent également être satisfaites en matière de couverture de la population de référence. Même si la définition de la population de référence est généralement large (et inclusive), tous les élèves ne peuvent pas participer à une évaluation telle que PIRLS ou PISA. Il est donc toléré d'exclure un nombre limité d'écoles ou d'enfants à l'intérieur des écoles, pour autant que ces exclusions ne dépassent pas un certain seuil (le seuil toléré est généralement de 5%) et répondent à certains critères. Les critères en usage

pour exclure des écoles ou des élèves dans les études de l'IEA sont : les élèves à besoins spéciaux (relevant de l'enseignement spécialisé)[9], les petites écoles, comptant moins de cinq élèves dans le grade testé et les écoles difficiles d'accès.

Le tirage de l'échantillon s'effectue toujours en deux étapes (échantillonnage par degrés[10])[11] :

- sélection d'établissements, en tenant compte de la taille des établissements, en sorte que chaque individu de la population a une chance égale de participer à l'enquête;
- sélection d'une classe entière (dans de rares cas, deux classes) dans les enquêtes de l'IEA ; dans PISA, sélection aléatoire de 35 élèves sur la liste de l'ensemble des élèves de 15 ans fréquentant l'établissement.

La sélection de classes entières ou la sélection d'élèves issus de différentes classes (qui peuvent même fréquenter des années d'études et filières d'études distinctes) a une influence non négligeable sur plusieurs aspects :

- dans un cas de figure comme PISA, les élèves évoluent dans des contextes de classe différents et il est exclu de mettre en relation certaines pratiques enseignantes avec les résultats – ce qui est possible dans le cas de classes entières. Par ailleurs, les effets des pairs sur l'apprentissage ne sont appréhendés qu'au niveau de l'école ; il s'agit davantage d'effets de composition (effet du recrutement) que d'effets de pairs liés à une socialisation commune dans le quotidien du groupe-classe ;
- dans le cas d'une classe entière retenue par établissement, la classe retenue peut par hasard se trouver être la classe la meilleure ou la plus faible de l'établissement. Celui-ci peut s'estimer « injustement » représenté, surtout si la classe donne une mauvaise image de l'établissement. Sur un plan plus technique, la variance entre écoles est confondue avec la variance entre classes. C'est en réalité bien plus entre les classes que l'on mesure les différences qu'entre les établissements. Dans le cas d'une sélection d'élèves à travers classes, les établissements sont mieux représentés dans leur diversité et la variance entre écoles est mieux estimée.

Modèles de mesure utilisés

Modèles de mesure

À l'intérieur des enquêtes à grande échelle, le nombre d'items disponibles est assez important et il est impossible de les administrer tous à chacun de ces derniers. C'est pourquoi, comme indiqué plus haut, seul un sous-ensemble des items comportant des items d'ancrage, soit des livrets (*booklets*), est présenté aux répondants. Une forme matricielle de l'échantillonnage engendrée par l'utilisation de ces livrets, qui diffèrent alors d'un répondant à un autre, exige toutefois l'utilisation de modélisations de la réponse aux items qui ne peuvent plus se limiter simplement à la somme des scores à chacun des items administrés. Puisqu'elles reposent sur des modèles strictement probabilistes, ces modélisations permettent d'ailleurs de pallier la présence de données manquantes par design (livrets) ou par omission de la part des répondants. Les items d'ancrage servent à l'estimation des paramètres des items et des répondants ; les autres items, n'étant pas administrés, correspondent alors à des données manquantes. De plus, ces modélisations permettent aussi d'assurer un suivi de l'évolution de la performance des élèves à travers le temps, ainsi que d'étudier le fonctionnement différentiel des items. Dans ce contexte, des modélisations qui reposent sur la théorie de la réponse à l'item sont appliquées.

Celles-ci permettent de calculer la probabilité que le répondant j donne une réponse spécifique à un item i, soit x_{ji}, conditionnellement à son niveau d'habileté θ_j et aux paramètres ζ_i propres au choix de réponse à l'item, soit $P(x_i \mid \theta_j, \zeta_i)$. De par le postulat d'indépendance locale, la probabilité d'un patron de réponse à n item est pour sa part égale à $\prod_{i}^{n} P(x_i \mid \theta_j, \zeta_i)$. Il est à noter que le niveau d'habileté peut être multidimensionnel et être ainsi représenté par un vecteur. C'est d'ailleurs le cas dans le contexte des enquêtes internationales.

Le paramètre de difficulté correspond, comme son libellé l'indique, à un indice qui indique le niveau de difficulté d'un item. Il varie de $+\infty$ à $-\infty$ et sa moyenne est souvent fixée à 0. Toutefois, à l'intérieur des enquêtes à grande échelle, on a plutôt tendance à fixer cette moyenne à 100 ou 500 pour éviter d'utiliser des valeurs négatives et ainsi faciliter la transmission des résultats.

De plus, un des avantages des modélisations issues de la théorie de la réponse à l'item est de faire en sorte que le niveau de difficulté de l'item soit sur la même échelle de mesure que le niveau d'habileté du répondant.

$$P(x_i \mid \theta_j, b_i) = \frac{1}{1+e^{-(\theta_j - b_i)}}$$

La modélisation précédente est appropriée aux items à réponse choisie et, plus généralement, aux items où il n'y a que deux possibilités de répondre, correcte ou incorrecte. Toutefois, dans plusieurs situations, on a besoin de modélisations plus flexibles qui permettent de calculer la probabilité qu'a un répondant de donner plus de deux réponses (correcte/incorrecte) et surtout de graduer la qualité de la réponse (modélisations polytomiques ordonnées). C'est le cas notamment lorsqu'on utilise une échelle descriptive ou une échelle d'appréciation de Likert pour noter la réponse du répondant. À cette fin, plusieurs modélisations ont été proposées : graduée (Samejima, 1969, 1997) et graduée généralisée (Muraki, 1990), crédit partiel (Masters, 1982) et crédit partiel généralisé (Muraki, 1992), échelle d'appréciation (*rating scale* d'Andrich, 1978). Dans le contexte des enquêtes internationales, c'est généralement la modélisation à crédit partiel qui est utilisée, ceci pour deux raisons. La première de ces raisons est théorique. En fait cette modélisation est appropriée avec des items à réponse construite dont la réalisation est effectuée par étape et dont la notation dépend de la réussite de l'étape précédente. On note alors la réponse à l'item à partir de crédits partiels. La seconde raison est pratique. Les modélisations pour réponses polytomiques ordonnées sont soutenues par des applications logicielles facilement disponibles. Toutefois, dans le contexte d'un échantillonnage matriciel des enquêtes internationales, seules les modélisations à crédit partiel ou à crédit partiel généralisé et par échelle d'appréciation sont soutenues par des logiciels spécialisés : par exemple, Conquest (Wu, Adams, & Wilson, 1998) ou ConstructMap (Kennedy, Wilson, Draney, Tutunciyan, & Vorp, 2008).

L'équation 2 représente la modélisation par crédit partiel où b_{ik} correspond au paramètre d'étape (*step*) de l'item i au score associé à la réussite de l'étape k. On remarquera au numérateur que la probabilité que la réponse du répondant soit associée au score k est une sommation et est donc tributaire du fait que les étapes précédentes ont été réussies.

$$P(x_i \mid \theta_j, b_{ik}) = \frac{\sum_{k=0}^{x} e^{-(\theta_j - b_k)}}{\sum_{h=0}^{m}\sum_{k=0}^{h} e^{-(\theta_j - b_k)}}$$

La figure 1 illustre la courbe caractéristique d'un item calibré selon une modélisation à crédit partiel. Il s'agit d'un exemple d'item en mathématiques (M124Q03T) tiré des échantillons fournis publiquement sur le site du PISA. Cet exemple est reproduit en annexe, ainsi que les paramètres qui lui sont associés selon le rapport technique du PISA (2005, p. 412). On peut ainsi noter à quel type de réponses correspondent un crédit complet (3), un crédit partiel (2) et pas de crédit (1). Contrairement aux courbes caractéristiques d'item des modélisations à réponse dichotomique, les modélisations polytomiques nécessitent toutefois plusieurs courbes, soient des courbes caractéristiques de réponse pour chacun des scores associés à chacune des k étapes. On peut y remarquer que le score le plus élevé, soit 3, est représenté par une courbe de probabilité strictement ascendante tandis que le score le plus faible, soit 1, est représenté par une courbe strictement descendante. Le score intermédiaire, pour sa part, affiche une courbe qui possède un maximum à son centre, ce qui indique que les répondants dont le niveau d'habileté est très faible ou très élevé ont peu de chances d'obtenir le score 2.

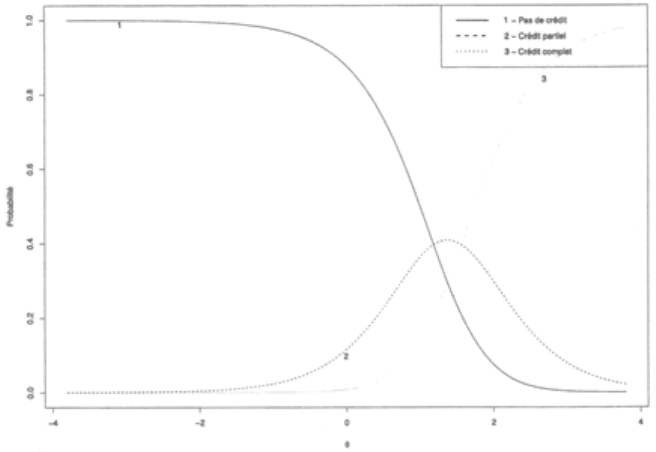


Figure 1. ***Courbe caractéristique d'un item selon la modélisation polytomique ordonnée par crédit partiel***

Estimation des paramètres

L'utilisation des modélisations issues de la théorie de la réponse à l'item nécessite l'estimation des paramètres de sujets et d'items. Cependant, comme on peut le remarquer aux équations 1 et 2, pour obtenir une estimation des paramètres d'items, il faut fixer les valeurs du niveau d'habileté. De manière similaire, pour obtenir une estimation du niveau d'habileté des répondants, il faut fixer les valeurs des paramètres d'items. On ne peut donc pas estimer les paramètres d'items et de sujets simultanément. Pour pallier ce problème, différentes solutions ont été proposées : maximum de vraisemblance conjointe (Birnbaum, 1968), maximum de vraisemblance marginale (Bock & Aitkin, 1981) et maximum de vraisemblance conditionnelle (Andersen, 1970). Puisqu'elle conduit à un biais important dans l'estimation des paramètres d'items, la méthode d'estimation par maximum de vraisemblance conjointe est rarement utilisée. Dans le contexte plus particulier des enquêtes à grande échelle, vu le caractère multidimensionnel du niveau d'habileté et la lourdeur des calculs associés à la méthode d'estimation par maximum de vraisemblance conditionnelle, on lui préférera généralement la méthode d'estimation par vraisemblance maximale marginale.

Valeurs plausibles

Dans le contexte particulier des enquêtes à grande échelle, on tient compte aussi d'autres aspects. Le premier concerne la structure matricielle de l'échantillonnage propre à l'utilisation de livrets qui varient d'un répondant un autre (Mislevy, 1991 ; Mislevy, Beaton, Kaplan, & Sheehan, 1992). Cette caractéristique implique que des données sont manquantes pour chacun des répondants (puisqu'aucun répondant n'a répondu à l'ensemble des items du test) et que de ce fait le niveau d'habileté de ceux-ci doit être imputé. Dans la plupart des enquêtes à grande échelle, cinq valeurs plausibles (*plausible values*) sont ainsi produites pour chacun des niveaux d'habileté (sous-échelles). Ensuite, dans la plupart de ces enquêtes, plusieurs niveaux d'habiletés sont estimés simultanément : par exemple, dans PISA 2003, on s'intéresse simultanément au rendement en lecture, en mathématique et en sciences. Enfin, en lien avec une troisième caractéristique, dans ces enquêtes, pour améliorer l'estimation des niveaux d'habileté, on tente de pallier le nombre restreint d'items administrés à chacun des répondants par l'utilisation de données collatérales (Mislevy & Sheehan, 1989) : statut socioéconomique, caractéristiques du milieu familial, perception du contexte scolaire, etc. Selon l'enquête, autour

de 300 variables collatérales diverses peuvent être utilisées à cette fin. Les valeurs plausibles sont ainsi produites à partir d'un modèle probabiliste en fonction non seulement des réponses aux items, mais aussi de ces données collatérales. Un algorithme de type EM (*Expectation – Maximisation*) est généralement appliqué à cette fin. Cette stratégie de remplacement des valeurs manquantes offre l'avantage de permettre d'ajuster le calcul de la précision des estimateurs en fonction du nombre de valeurs plausibles générées. Ce sont d'ailleurs tous ces aspects que les applications logicielles précitées permettent de prendre en compte : Conquest, ConstructMap et aussi BGROUP (Sinhary & von Davier, 2005) – cette dernière ne permet toutefois pas d'estimer les paramètres d'items. Cette stratégie d'imputation multiple pourrait d'ailleurs aussi expliquer le phénomène souligné plus haut (section sur l'élaboration d'un cadre de référence), soit qu'on observe une corrélation importante entre les sous-échelles des enquêtes allant jusqu'à 0,90. Plus précisément, puisque le niveau d'habileté relié à chacun des sous-échelles est estimé en utilisant de l'information relative à chacune des autres sous-échelles, une corrélation entre celles-ci s'ajoute à la corrélation naturelle entre elles. On peut alors se demander si cette corrélation élevée n'est que factice. Le nombre d'items restreints administrés à chacune des sous-échelles renforce alors cet impact sur l'estimation des coefficients de corrélation. Dans ce contexte, il est possible que l'ajustement global à un modèle multidimensionnel ne soit pas beaucoup plus important que celui à un modèle unidimensionnel. On pourrait alors se demander l'avantage théorique d'utiliser un modèle multidimensionnel, plutôt qu'un modèle unidimensionnel.

Pour terminer cette section, il faut souligner que ces enquêtes à grande échelle ne sont pas conçues pour estimer le niveau d'habileté de chacun des répondants, mais plutôt pour estimer des valeurs moyennes par groupe de répondants. Toutes les analyses qui sont effectuées à partir des données issues de ces enquêtes ne devraient d'ailleurs s'appliquer qu'à des groupements de répondants. De plus, puisque les niveaux d'habiletés sont plutôt imputés et ainsi fournis sous forme de valeurs plausibles, les analyses statistiques basées sur ceux-ci doivent tenir compte de cette particularité. Plus spécifiquement, il est nécessaire de répéter les analyses autant de fois qu'il y a de valeurs plausibles, généralement cinq. Les résultats correspondent alors à la moyenne des cinq analyses et l'erreur type des statistiques associées doit alors être ajustée en conséquence.

Qualité des mesures

La qualité des mesures obtenues à l'intérieur des enquêtes internationales concerne aussi bien l'ensemble des données à l'étude que chacun des paramètres d'items. Considérant ce qui vient d'être souligné au regard du fait que les niveaux d'habileté doivent être abordés pour des groupes de répondants et non pas de manière individuelle, l'analyse de l'ajustement des patrons de réponses par répondant (*person fit*) ne semble pas appropriée. Ainsi, en ce qui a trait aux niveaux d'habileté, ce sera presque uniquement l'erreur type par groupe qui sera présentée. Par exemple, l'erreur type dans PISA 2003, avec moyenne du niveau d'habileté en mathématiques de 532, varie entre 1,80 et 4,70 dans les provinces canadiennes (moyenne de 1,80). Dans ce contexte, on s'attardera plus spécifiquement à l'ajustement global, à l'ajustement des paramètres d'items et à l'étude du fonctionnement différentiel des items.

Ajustement global et des paramètres

Les indices d'ajustement global usuels sont généralement associés à une statistique du χ^2 et calculés soit à partir d'un carré résiduel, soit d'un rapport de vraisemblance. Ainsi, par exemple, à l'intérieur de l'application logicielle PARSCALE (du Toit, 2003), on utilise le logarithme naturel du rapport de vraisemblance pondéré associé à l'indice G^2 qui se distribue selon une loi du χ^2.

Beaton (2003 : voir Sinhary, Guo, & von Davier, 2010), pour sa part, dans le contexte de l'enquête du NAEP, préfère utiliser une adaptation des indices d'ajustement de Wright et Stone (1979) pondérée par les poids associés à chacun des répondants de l'enquête W_i. Ces indices se distribuent aussi selon une loi du χ^2, mais sont plutôt basés sur la somme des résidus standardisés au carré.

L'indice proposé par Beaton présente l'avantage de tenir compte du poids accordé à chacun des répondants de l'enquête. Il serait toutefois assez facile de proposer une adaptation de l'indice G^2 à cet effet.

L'indice suggéré par Beaton présente un désavantage important, car il est calculé à partir du niveau d'habileté estimé, ce qui peut fausser grandement la valeur obtenue. Théoriquement, il devrait être calculé à partir de la valeur réelle du niveau d'habileté du répondant, valeur qui bien sûr est inconnue. Pour pallier ce problème, Wu (1997; Wu, et al., 1998) suggère une version de cet indice qui est calculé par marginalisation sur l'ensemble des valeurs possibles du niveau d'habileté. Sans aucun doute, cette stratégie est bien plus

appropriée et devrait être utilisée plus largement. Les calculs impliquent toutefois l'utilisation d'intégrales multiples complexes. De plus, elles sont disponibles, pour le moment, uniquement à partir d'applications logicielles spécialisées telles que Conquest et ConstructMap. De ce fait, malheureusement, les modélisations de la réponse à l'item sont actuellement limitées à celles qui sont issues des modèles de Rasch (Rasch, 1960).

Fonctionnement différentiel d'item

Un dernier aspect de la qualité des mesures qui mérite d'être abordé ici concerne le fonctionnement différentiel des items en fonction du groupe à l'intérieur duquel ces paramètres sont estimés. Pour que les mesures soient comparables, par exemple d'un pays à un autre ou chez les hommes et les femmes, il est nécessaire que tous les items d'un même test partagent les mêmes paramètres. Si ce n'est pas le cas, on peut se douter que la probabilité d'obtention d'un score donné à certains items est affectée par d'autres facteurs que le niveau d'habileté des répondants. Des différences culturelles peuvent affecter cette probabilité ou encore des processus cognitifs différenciés. Lorsque cette situation se produit, on dit que ces items présentent un fonctionnement différentiel. Il faut donc les relever et les retirer du test. Plusieurs stratégies et indices ont été proposés pour détecter des items qui présentent un fonctionnement différentiel (Bertrand & Blais, 2004). Il serait trop long ici de les présenter tous, mais on peut noter que les méthodes de Mantel-Haenszel (Dorans & Holland, 1994) et de la régression logistique (Swaminathan & Rogers, 1990) sont parmi les plus fréquemment employées lorsque les réponses aux items sont dichotomiques. Lorsqu'il s'agit de réponses polytomiques ordonnées, toutefois, beaucoup de travaux restent encore à faire pour obtenir des indices satisfaisants dont on connaît bien le comportement.

Conclusion

Cet article a tenté de jeter un éclairage sur des aspects des enquêtes internationales peu connus en dehors des cercles de spécialistes. Comment s'élabore une enquête internationale, quels choix méthodologiques sont posés aux diverses étapes du processus avec quelles conséquences pour l'interprétation des résultats, quels contrôles pour garantir la comparabilité et la rigueur dans la collecte des données, quel choix de population et quelles procédures pour garantir la représentativité de l'échantillon, quels modèles de mesure sont utilisés ?

Nous espérons, au travers des exemples fournis, avoir pu montrer à quel point ces enquêtes, dont on ne connaît souvent que la face immergée – le palmarès des pays – allient à la fois rigueur méthodologique et expertise technique, sans négliger une réelle implication des systèmes éducatifs concernés dans le processus d'élaboration des tâches d'évaluation notamment.

Comme nous l'avons montré ailleurs (Lafontaine & Simon, 2008), les enquêtes internationales ont connu, depuis une quinzaine d'années, d'importantes avancées. Les procédures de contrôle de qualité, de vérification, notamment des échantillons et des traductions, sont aujourd'hui beaucoup plus strictes qu'elles ne l'étaient il y a 20 ans. La qualité et la comparabilité des données s'en sont trouvées considérablement accrues.

Par ailleurs, l'utilisation des modélisations issues de la théorie de la réponse à l'item a eu un impact qui dépasse de loin les possibilités techniques qu'autorisent ces modèles. Concrètement, c'est grâce à ces modélisations que l'on peut mettre en œuvre des plans d'évaluation complexes avec rotation de livrets – ce qui permet d'évaluer plus en profondeur le domaine sans surcharger les élèves testés. C'est aussi grâce à ces modélisations que l'on peut mesurer de façon rigoureuse les évolutions dans le temps (indicateurs de tendance) ou encore décrire, d'une manière précise, les compétences d'élèves situés à différents niveaux sur les échelles de résultats.

Les modélisations issues de la théorie de la réponse à l'item posent toutefois des défis qui n'ont pas toujours reçu de solutions complètement satisfaisantes. Par exemple, comme on a pu le constater à la section précédente qui abordait la qualité des mesures, d'importants travaux sont à effectuer en ce qui a trait à l'application des indices d'ajustement des modélisations et à la détection du fonctionnement différentiel d'items dans le contexte de ces enquêtes internationales. Au regard de l'ajustement, c'est sûrement la détection

des patrons de réponses inappropriés (*person fit*) qui exigerait le plus d'attention. Des travaux sont actuellement à leur début pour proposer des modélisations multidimensionnelles qui permettraient non seulement de détecter les patrons de réponses inappropriés dans le contexte de ces enquêtes, mais aussi de corriger éventuellement l'estimation des niveaux d'habileté des répondants en fonction de paramètres de personne supplémentaires : fluctuation du niveau d'habileté dans le test, inattention et pseudo-chance personnelle (Blais, Raîche, & Magis, 2009 ; Raîche, Magis, & Blais, 2009). Il est d'ailleurs actuellement possible d'estimer les paramètres de ces modélisations à l'aide de la librairie *irtProb* du logiciel *R* (Raîche & Magis, 2009).

Il faut aussi souligner que l'imputation des niveaux d'habileté exigerait d'être plus étudiée, car le nombre de celles-ci est presque toujours relativement faible (cinq). La disponibilité des logiciels à la fois adaptés aux particularités des enquêtes internationales et offrant une meilleure panoplie de modélisations de la réponse à l'item est encore relativement restreinte. Enfin, à l'intérieur des grandes enquêtes internationales, comme on l'a expliqué précédemment, beaucoup d'efforts sont fournis pour élaborer des items de grande qualité afin d'estimer au mieux les niveaux d'habileté des répondants. Les items dits contextuels gagneraient à recevoir la même considération pour faire en sorte que leurs caractéristiques psychométriques soient optimisées et que leur pouvoir de discrimination soit maximisé. Ce serait fort important pour deux raisons. La première se réfère au fait que les réponses à ces items sont utilisées dans les calculs destinés à la génération des valeurs imputées des niveaux d'habileté. Il est donc primordial que l'information provenant de ces items contextuels permette la meilleure prédiction possible. La seconde raison est que ces items contextuels sont disponibles pour les chercheurs et utilisateurs qui ont accès aux données de ces enquêtes. Ces derniers effectuent différentes analyses pour comparer les pays et expliquer le rendement des répondants en fonction de ces items contextuels. Ils rendent alors publics les résultats de ces analyses et il est important que les données disponibles puissent soutenir correctement les interprétations. Plusieurs de ces items contextuels, ou les indices qui en découlent, permettent d'expliquer jusqu'à environ 30 % de la variance du niveau d'habileté : dans le cas du PISA 2003, notamment les indices associés au sentiment d'efficacité personnelle en mathématiques. Un grand nombre de ces items contextuels, toutefois, n'expliquent pas plus de 2 % de la variance du niveau d'habileté : toujours dans PISA 2003, on peut donner en exemple le temps alloué par le répondant à la réalisation de ses devoirs ou encore la fréquence d'utilisation de logiciels éducatifs. Malgré ce fait,

considérant le nombre important d'observations disponibles dans ces enquêtes, les résultats sont très fréquemment statistiquement significatifs, mais peuvent tout de même s'avérer peu significatifs sur le plan des pratiques pédagogiques ou des politiques éducatives. Cette situation est malheureusement rarement commentée lorsque les résultats des analyses sont rendus publics.

NOTES

1. Pour une étude de l'évolution historique des enquêtes internationales consacrées à la lecture, on se reportera à Lafontaine (2004).
2. Rappelons qu'il s'agit de corrélations latentes (épurées de l'erreur de mesure), qui sont donc plus élevées que des corrélations classiques.
3. Plus loin dans ce texte, à la section relative aux valeurs plausibles, une note plus technique à cet effet sera apportée.
4. L'intention n'est pas d'ouvrir ici un débat sur la nature des tâches proposées dans les enquêtes internationales, mais d'insister sur la cohérence entre la définition des domaines à évaluer et la nature des tâches.
5. Ici non plus, il ne s'agit pas de discuter l'impact que peut avoir le type de support sur les processus évalués et les résultats. Pour une discussion de cette question, nous renvoyons le lecteur à Blais (2008).
6. La question de l'impact du format de question sur les performances a notamment été étudiée par Lafontaine et Monseur (2009).
7. Les paragraphes qui suivent sont largement empruntés au *Technical report* de PISA 2006 (OCDE, 2009a).
8. Le livret est la forme matérielle du test que les élèves reçoivent. Les «blocs» représentent des ensembles de tâches et d'items : dans PISA 2009, par exemple, on compte sept blocs pour la lecture, trois blocs pour les mathématiques et trois blocs pour les sciences.
9. Dans PISA, une partie des élèves fréquentant l'enseignement spécialisé est concernée.
10. L'un des avantages de travailler avec des échantillons par degré (par rapport à un échantillon aléatoire et simple, par exemple) est qu'il n'est pas nécessaire de posséder *a priori* la liste complète et sans répétition de tous les élèves qui fréquentent le système éducatif d'un pays (ce qui est en général impossible à obtenir). Il est simplement nécessaire de disposer d'une liste complète et sans répétition des établissements scolaires d'un pays et des classes des établissements sélectionnés à la première étape. Bien entendu, il permet en outre de concentrer géographiquement la prise de données mais complexifie les estimations des variances et des erreurs standards (d'où la quasi-nécessité d'utiliser des logiciels comme WESTAT pour faciliter le calcul des estimations).
11. Autrement dit, on ne tire pas directement au sort des classes ou des élèves, on passe d'abord par une sélection des établissements.

RÉFÉRENCES

Adams, R., Berezner, A., & Jakubowski, M. (2010). *Analysis of PISA 2006 preferred items ranking using the percent correct method*. OECD Education Working Paper n° 46.

Andersen, E. M. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society, Series B, 32*, 283-301.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Bertrand, R., & Blais, J-G. (2004). *Modèles de mesure. L'apport de la théorie des réponses aux items*. Ste-Foy, Québec : Presses de l'Université du Québec.

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (éds), *Statistical theories of mental test scores*. Reading, MT : Addison-Wesley.

Blais, J.-G. (dir.) (2008). *Évaluation des apprentissages et technologies de l'information et de la communication. Enjeux, applications et modèles de mesure*. Ste-Foy, QC : Presses de l'Université Laval.

Blais, J.-G., Raîche, G., & Magis, D. (2009). La détection des patrons de réponses problématiques dans le contexte des tests informatisés. In J.-G. Blais (dir.), *Évaluation des apprentissages et technologies de l'information et de la communication : enjeux, applications et modèles de mesure* (p. 275-291). Ste-Foy, QC : Presses de l'Université Laval.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika, 46*, 443-459.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (éds), *Differential item functioning*. Mahwah, NJ: LEA.

du Toit, M. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific software international.

Grisay, A. (2003). Translator procedures in OECD/PISA 2000 international assessments. *Language Testing, 20*, 228-240.

Grisay, A., de Jong, J. H. A. L., Gebhardt, E., Berezner, A., & Halleux-Monseur, B. (2007). Translation Equivalence across PISA Countries. *Journal of Applied Measurement, 8*(3), 249-266.

Kennedy, C. A., Wilson, M. R., Draney, K., Tutunciyan, S., & Vorp, R. (2008). *Construct Map v4.4.4.0. Quick start guide*. Berkeley, CA: University of Berkeley.

Lafontaine, D. (2004). From comprehension to literacy: thirty years of reading assessment. In J. Moskowitz & M. Stephens (éds), *Comparing learning outcomes: international assessment and education policy*, (p. 29-46). London: Routledge Falmer.

Lafontaine, D., & Monseur, C. (2009). Gender Gap in Comparative Studies of Reading Comprehension: to what extent do the test characteristics make a difference? *European Educational Research Journal. Special issue on PISA and gender, 8*(1), 69-79.

Lafontaine, D., & Simon, M. (2008). L'évaluation des systèmes éducatifs. *Mesure et évaluation en éducation, 31*(3), 95-125.

Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (2007). *PIRLS 2006 Technical Report.* Lynch School of Education: Boston College.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*(2), 177-196.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133-161.

Mislevy, R. J., & Sheehan, K. M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika, 54*(4), 661-679.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14*, 59-71.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., Trong, K. L., & Sainsbury, M. (2009). *PIRLS 2011 Assessment framework*. Lynch School of Education: Boston College.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., Arora, A., & Erberber, E. (2005). *TIMSS 2007 Assessment framework*. Chestnut Hill, MA: Boston College.

OCDE (1999). *Mesurer les compétences et les connaissances des élèves : un nouveau cadre d'évaluation*. Paris : OCDE.

OCDE (2005). *PISA 2003 Technical Report*. Paris: OCDE.

OCDE (2009a). *PISA 2006 Technical Report.* Paris: OCDE.

OCDE (2009b). *Le cadre d'évaluation de PISA 2009 : les compétences en compréhension de l'écrit, en mathématiques et sciences*. Paris : OCDE.

Raîche, G., & Magis, D. (2009). *irtProb* 1.0 - *Utilities and probability distributions related to multidimensional person item response models (IRT)*. R package.

Raîche, G., Magis, D., & Blais, J.-G. (2009, juillet). *Multidimensional item response theory models integrating additional inattention, pseudo-guessing, and discrimination person parameters*. Communication présentée au congrès annuel de la Psychometric Society, Durham, NH.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Robitaille, D. F., Schmidt, W. H., Raizen, S., Mc Knight, C., Britton, E., & Nicol, C. (éds) (1993). *Curriculum frameworks for mathematics and science.* Vancouver: Pacific Educational Press.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometric Monographs*, No. 17.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (éds), *Handbook of item response theory* (p. 85-100). Mahwah, NJ: LEA.

Sinhary, S., Guo, Z., & von Davier, M. (2010). Assessing the fit of latent regression models. In *IERI monograph series 3, Issues and methodologies in large-scale assessments, volume 3*. Princeton, NJ: Educational testing service.

Sinhary, S., & von Davier, M. (2005). *Extension of the NAEP BGROUP program to higher dimensions*. Princeton, NJ: Educational testing service.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.

Wright, B. D., & Stone, M. (1979). *Best test design: Rasch measurement*. Chicago, IL: MESA Press.

Wu, M. L. (1997). *The development and application of a fit test for use with marginal maximum likelihood estimation and generalized item response models*. Mémoire de maîtrise non publié, Université de Melbourne.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ACER Conquest. Generalised item response modelling software*. Melbourne, Victoria, Australia: ACER Press.

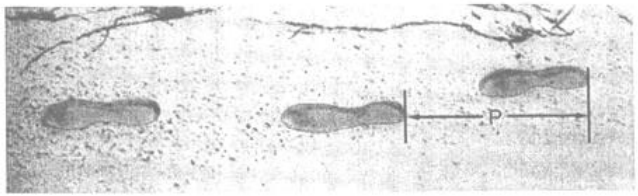
Date de réception : 4 mars 2011

Date de réception de la version finale : 21 juin 2011

Date d'acceptation : 30 juin 2011

ANNEXE

Marche à pied



L'image montre les traces de pas d'un homme en train de marcher. La longueur de pas *P* est la distance entre l'arrière de deux traces de pas consécutives.

Pour les hommes, la formule donne un rapport approximatif entre *n* et *P*, où:

n = nombre de pas par minute,

P = longueur de pas en mètres.

Question M124Q3T : MARCHE À PIED

Bernard sait que la longueur de son pas est de 0,80 mètre. La formule s'applique à sa façon de marcher.

Calculez la vitesse à laquelle marche Bernard en mètres par minute et en kilomètres par heure. Montrez vos calculs.

Marche À pied: CONSIGNES DE CORRECTION

Crédit complet

Code 31: Réponses correctes fournies à la fois pour les mètres par minute et les km par heure (les unités ne sont pas exigées):

n = 140 x 0,80 = 112.

En une minute, il parcourt 112 × 0,80 mètre = 89,6 mètres.

Sa vitesse est donc de 89,6 mètres par minute.

Par conséquent, sa vitesse est de 5,38 km/h ou 5,4 km/h.

Coder 31 si les deux réponses correctes sont fournies (89,6 et 5,4), que l'élève ait montré ou non son travail. Noter que les erreurs d'arrondi sont acceptables. Par exemple, 90 mètres par minute et 5,3 km/h (89 × 60) sont acceptables.

89,6 ; 5,4.

90 et 5,376km/h.

89,8 et 5376 m/h [à noter que si le second chiffre n'avait pas été fourni avec les unités, cette réponse aurait été codée 22].

Crédit partiel (2 points)

Code 21 : Comme pour le code 31, mais oublie de multiplier par 0,80 pour convertir les pas par minute en mètres par minute. Par exemple, sa vitesse est de 112 mètres par minute et 6,72 km/h.

112 et 6,72 km/h.

Code 22 : La vitesse en mètres par minute est correcte (89,6 mètres par minute) mais la conversion en kilomètres/heure est incorrecte ou omise.

89,6 mètres par minute, 8 960 km/h.

89,6 et 5376.

89,6 et 53,76.

89,6 ; 0,087 km/h

89,6 et 1,49 km/h.

Code 23 : Méthode correcte (explicitement montrée), mais erreur(s) de calcul mineure(s), non couverte(s) par les codes 21 et 22. Aucune des deux réponses n'est correcte.

n = 140 × 0,8 = 1120 ; 1120 × 0,8 = 896. Il marche à une vitesse de 896m/min, soit 53,76 km/h.

n =140 x 0,8 = 116 ; 116 x 0,8 = 92,8. 92,8 m/min -> 5,57km/h.

Code 24: Fournit seulement la réponse 5,4 km/h, et non 89,6 m/min (les calculs intermédiaires ne sont pas montrés).

5,4

5,376 km/h.

5 376 m/h.

Crédit partiel (1 point)

Code 11: n = 140 x 0,80 = 112. Pas d'autre calcul montré, ou calcul erroné après ceci.

112.

n = 112; 0,112 km/h.

n = 112; 1120 km/h.

112 m/min ; 504 km/h.

Pas de crédit

Code 00 : Autres réponses.

Code 99 : Omission.

Note a: Tiré des items échantillons en mathématiques fournis publiquement sur le site du PISA : [http://www.oecd.org/document/38/0,3746,en_32252351_32236173_34993126_1_1_1_1,00.html].

Note b: Paramètres d'items selon la modélisation à crédit partiel: difficulté (1,488) et taus (τ = -0,301, 0,076, 0,225).