Évaluation des systèmes éducatifs

Dominique Lafontaine
Université de Liège

Marielle Simon
Université d'Ottawa

Mots clés: Enquêtes comparatives, rendement scolaire, politiques éducatives, indicateurs, équivalence culturelle

Au cours des trente dernières années, l'évaluation des systèmes éducatifs a connu d'importants changements tant sur le plan des finalités que sur le plan méthodologique. Les développements majeurs qu'a connus le champ ne peuvent être retracés au travers des articles que lui a consacrés la revue Mesure et évaluation en éducation. *Ce thème n'y occupe en effet qu'une place marginale C'est donc au travers d'autres sources que les enjeux et développements majeurs du domaine sont abordés. Les principaux défis à relever s'inscrivent dans la perspective d'une démarche plus active de la revue pour solliciter des contributions sur cette problématique en plein développement.*

Key words: Comparative surveys, student achievement, educational policy, indicators, cultural equivalence

During the last 30 years, assessment of educational systems has gone through several tremendous changes both from a policy and a methodological point of view. The major developments in the field cannot be tracked through the papers appearing in Mesure et évaluation en éducation. *Indeed, very few papers have been dedicated to that topic. Main stakes and trends will therefore be studied through other sources. Major challenges to address in the near future will be presented and suggestions made for a broader coverage of the domain in the journal* Mesure et évaluation en éducation.

Palavras-chave: Estudos comparativos, rendimento escolar, políticas educativas, indicadores, equivalência cultural

Ao longo dos últimos trinta anos, a avaliação dos sistemas educativos conheceu importantes mudanças, quer no plano das finalidades, quer no plano metodológico. Os maiores desenvolvimentos neste domínio não podem, no entanto, ser sinalizados através dos artigos que lhe consagrou a revista Mesure et évaluation en éducation, na qual, com efeito, este tema tem tido um lugar marginal. Daí que os maiores desafios e desenvolvimentos da avaliação dos sistemas educativos sejam abordados através do recurso a outras fontes. Os principais desafios a considerar inscrevem-se na perspectiva de um processo mais activo da revista, no sentido de solicitar contributos sobre esta problemática em pleno desenvolvimento.

Note des auteures – Nous tenons à remercier vivement Christian Monseur pour sa relecture attentive du texte et ses commentaires pertinents. Toute correspondance peut être adressée par courriel aux adresses suivantes : [dlafontaine@ulg.ac.be] ou [msimon@uottawa.ca].

Cadrage du texte

Depuis 1978, rares sont les articles publiés dans la revue *Mesure et évaluation en éducation* qui relèvent du domaine de l'évaluation des systèmes éducatifs[1]. Le présent article se tournera donc vers d'autres sources pour analyser les enjeux et développements majeurs encourus dans le domaine au cours des trente dernières années. Avant d'entreprendre cette synthèse, nous nous sommes interrogées sur le sens à donner à l'évaluation des systèmes éducatifs. Pour la recension des articles parus dans la revue *Mesure et évaluation en éducation*, nous avons pris en compte à la fois les évaluations nationales et internationales. En revanche, pour l'analyse des enjeux et développements majeurs dans le domaine, nous avons jugé pertinent de limiter l'objet aux enquêtes internationales, pour différents motifs. Un premier motif, pragmatique, était celui de l'espace disponible, trop restreint pour traiter des développements qui nous apparaissent comme distincts sur le plan national et international, même si les liens d'enrichissement mutuel ne manquent pas. Par ailleurs, les développements dans les évaluations nationales sont passablement divergents, très dépendants des contextes éducatifs nationaux et bien plus traversés par des enjeux de type politico-administratif que par des enjeux proprement scientifiques auxquels nous pensons dans cet article devoir donner la priorité. Pour plus d'information sur les enjeux liés aux évaluations nationales des systèmes, on se reportera à Crahay (à paraître), Hurteau (à paraître), Lafontaine, Soussi et Nidegger (à paraître), Maroy (à paraître). Pour ces différents motifs, l'analyse se centrera donc 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.oecd.pisa.org]). Ces enquêtes concernent principalement les pays économiquement les plus avancés ; nous n'aborderons pas ici les enquêtes menées par d'autres instances, comme le PASEC (Programme d'analyse des systèmes éducatifs de la CONFEMEN) et le SAMEQ (Southern and Eastern Africa Consortium for Monitoring Educational Quality), qui s'adressent aux pays en voie de développement[2]. Dans un premier temps, les enjeux et développements majeurs encourus au cours des trente dernières années seront analysés. Ensuite, nous tenterons d'esquisser les défis à relever dans ce domaine pour les prochaines années.

Évaluation des systèmes éducatifs : quelle évaluation ?

L'évaluation des systèmes éducatifs peut porter sur différents acteurs de ce système – les élèves, les enseignants, les établissements – ou sur différentes composantes, telles que le coût, l'efficacité, la qualité, les ressources (Nacuzon Sall & De Ketele, 1997). Il ne fait guère de doute que les grandes enquêtes internationales ont, au cours de leur développement, privilégié certaines dimensions des systèmes. C'est ce que nous allons, dans un premier temps, tenter de cerner en examinant l'ensemble des enquêtes réalisées.

Tableau 1
Liste des enquêtes internationales par discipline

	1960	1970	1980	1990	2000
Langue maternelle		Lecture-compréhension, Littérature	Composition écrite	Reading (RLS) Literacy	RLS_R PIRLS 2001 PIRLS 2006 PISA2000 PISA2003 PISA2006 PISA2009
Mathématiques	FIMS		SIMS	TIMSS TIMSS_R	PISA2000 PISA2003 PISA2006 PISA2009 TIMSS 2003 TIMSS 2007 TIMSS Ad. 08
Sciences		FISS	SISS	TIMSS, TIMSS_R	PISA2000 PISA2003 PISA2006 PISA2009 TIMSS 2003 TIMSS 2007 TIMSS Ad. 08
Autres		Anglais, Français comme langue étrangère, Éducation civique	Comped (Computer Education)	Civics Sites PPP (Preprimary Project) Sites	Sites (Information Technology) Teacher Education Study (Teds)

L'examen du tableau fait rapidement apparaître que la majorité des études se sont centrées sur le **rendement scolaire** des élèves. Les grandes enquêtes internationales évaluent les acquis cognitifs des élèves dans un ou plusieurs domaines et mettent ces acquis en relation avec différents éléments de contexte établis principalement via des questionnaires à l'élève, au chef d'établissement et, plus rarement, aux enseignants. L'évaluation est parfois élargie à des dimensions socio-affectives (attitudes, motivation, intérêt, etc.). Les études ne portent pas véritablement sur la mise en relation des coûts (investissement) et des résultats. Il est plus exact de dire, en référence au modèle de l'IEA, que l'ambition première de ces études est d'**évaluer le curriculum réalisé** (ou atteint). Une certaine attention est toutefois accordée au curriculum visé via l'analyse des instructions officielles, comme dans *PIRLS Encyclopedia* (Kennedy, Mullis, Martin, Kennedy & Trong, 2007; Mullis, Martin, Kennedy & Flaherty, 2001), et au curriculum implanté via des questions aux enseignants portant sur les occasions d'apprendre (OTL, *opportunity-to-learn*). Parmi les domaines scolaires abordés, une nette priorité est accordée aux disciplines de base – la lecture-compréhension, les mathématiques et les sciences. Un peu en retrait vient l'éducation civique, à laquelle trois opérations ont été consacrées. Enfin, quelques rares études menées par l'IEA à la fin des années 1980 et dans les années 1990, telles l'étude PPP (*Preprimary Project*), les études *Comped* et *Sites* portant sur l'informatique s'inscrivent clairement en dehors du cadre traditionnel de ces études. Ces dernières développent une perspective davantage descriptive, utilisent des inventaires de pratiques, et n'évaluent pas les acquis des élèves. Mais il s'agit d'études marginales par rapport aux études de rendement scolaire. Nos analyses porteront donc de façon privilégiée sur les évaluations du rendement scolaire.

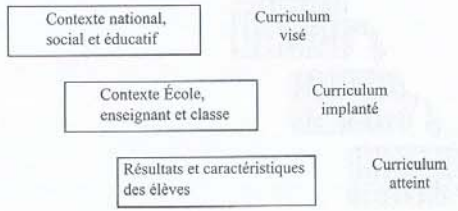

Figure 1. ***Modèle de l'IEA (Mullis et al., 2005, p. 5)***

Comme l'indique le tableau 1, les années 1995-2008 se caractérisent par un fort accroissement de l'offre. Les sociétés modernes sont demandeuses d'évaluation, l'école est sommée de rendre des comptes. Les évaluations de systèmes éducatifs, qui avaient précédé ce mouvement d'*accountability*, vont se retrouver en quelque sorte prises dans cette vague. Ceci va contribuer à redéfinir au moins partiellement leurs finalités, leur mode d'organisation et même leur contenu et leur méthodologie.

Changements de finalité

Quelles intentions et quel objet d'évaluation ?

L'ambition de la plupart des études de l'IEA est d'évaluer le curriculum réalisé, à savoir les acquis des élèves. Mais le curriculum visé diffère évidemment dans les différents pays participants et un arbitrage doit donc avoir lieu pour décider du contenu de l'évaluation. Jusqu'au milieu des années 1990, les évaluations de l'IEA ont été accompagnées d'une étude des curriculums visés ou prescrits (programmes d'études, instructions officielles). L'approche consistait dès lors à comparer les différents curriculums et à faire porter le test sur une forme de «dénominateur» commun; une procédure empirique donc. Classiquement, les évaluations plus anciennes s'accompagnent de questions dites d'*opportunity-to-learn*, visant à déterminer dans quelle mesure le contenu du test correspond aux contenus réellement enseignés dans les classes, une manière de vérifier *a posteriori* si, dans chaque pays considéré, le test correspond de près ou de loin au curriculum enseigné (Beaton, Mullis, Martin, Gonzalez, Kelly & Smith, 1996). En 2000, avec l'enquête PISA (Programme international pour le suivi des acquis des élèves de 15 ans) s'amorce un tournant important. L'OCDE (l'Organisation de coopération et de développement économiques) affiche en effet une autre ambition que les enquêtes de l'IEA. Alors que les études de l'IEA font un bilan de rendement scolaire, la question que pose PISA est davantage tournée vers l'avenir et l'insertion des jeunes dans une société et un monde globalisés. La question du lien du test avec le curriculum s'y pose donc en de tout autres termes. Le choix du contenu du test ne se fait dès lors plus en référence à des contenus enseignés, mais à des compétences jugées essentielles par des experts internationaux ou nationaux pour que les jeunes s'insèrent de manière active dans la société de demain. «Le programme PISA, lancé par les gouvernements de l'OCDE en 1997 vise à évaluer quelques-unes des

compétences-clés qui contribuent à la réussite des individus, sur une base régulière et avec un cadre d'évaluation commun accepté internationalement. PISA cherche à établir un dialogue autour des politiques éducatives et une collaboration dans la définition et l'implantation d'objectifs éducatifs, en suivant une approche innovante qui implique des jugements sur les skills pertinents pour la vie adulte» (Schleicher, 2007, p. 350).

L'analyse des curriculums disparaît donc dans PISA, où elle est remplacée par un cadre de référence dans lequel sont détaillés les références théoriques, les choix concernant la structuration du domaine, les modalités d'évaluation, les échelles destinées à présenter les résultats, proposés par le groupe d'experts international et soumis à l'examen des représentants des pays participants et des gestionnaires nationaux de projet (Campbell, Kelly, Mullis, Martin & Sainsbury, 2000; Mullis, Martin, Ruddock, O'Sullivan, Arora & Erberber, 2005; OCDE, 1999). Le fait que ces documents soient destinés à être rendus publics se traduit par un effort de clarification et de planification. Les études contemporaines et à venir seront ainsi davantage assises sur le plan théorique qu'elles ne l'ont été antérieurement, où une approche empirique prédominait. De manière plus critique, on peut aussi entrevoir, derrière cette absence de relation aux curriculums nationaux, une tentative de l'OCDE d'imposer une forme de curriculum globalisé. C'est ce que pensent, en tout cas, les courants de recherche portant sur la globalisation dans le champ de l'éducation. Pour Spring (2008), «l'OCDE a joué un rôle majeur dans la standardisation globale de l'éducation, au travers de son programme PISA. [...] En devenant un standard international, PISA a le potentiel direct pour déterminer le contenu du curriculum dans les domaines testés, *i.e.* les mathématiques, la lecture et les sciences» (p. 346).

Analyses primaires et secondaires

Un autre changement important à la fin des années 1990 est le développement des analyses secondaires. Les enquêtes à grande échelle produisent une foule de données qui sont analysées d'abord afin de fournir un profil des connaissances, des compétences et des attitudes des élèves et ensuite mises en relation avec diverses variables contextuelles recueillies via les questionnaires à l'élève, à l'enseignant, au chef d'établissement, ou, de façon plus occasionnelle, aux parents.

Les analyses et rapports consacrés aux premières enquêtes internationales menées par l'IEA sont signés par de grands noms de la pédagogie – Carroll, Husen, Neville Postlethwaite, Purves, Thorndike. Si l'analyse des résultats

obtenus par les pays est, aux yeux des standards actuels, relativement limitée, les analyses statistiques menées pour tenter d'expliquer les résultats sont déjà d'un niveau de sophistication élevé. Dès le départ, les modèles de régression multiple sont utilisés ; les enquêtes internationales de l'IEA représenteront à cet égard, dans les années 1970-1980, une école, un lieu de formation où plusieurs experts mondialement reconnus sont venus construire leur expérience pour ensuite la redistribuer dans leurs équipes de recherche. Certes, l'arrivée des modèles multiniveaux dans le courant des années 1990 représentera un plus incontestable, permettant une meilleure estimation des effets, et surtout une appréhension plus exacte de la réalité des effets de contexte (effets école, effets de pairs).

Dès ses débuts, l'IEA a mis ses données brutes à la disposition du public afin d'encourager l'application d'analyses secondaires aux fins politiques, administratives ou académiques. Sur demande écrite, les données étaient envoyées sur des rubans ou disquettes magnétiques, accompagnées de documents précisant les codes et les libellés de chacune des variables. À la fin des années 1980, les organismes responsables des enquêtes internationales fournissent sur leurs sites web respectifs des variables composites, des mini-programmes d'analyses ciblées et toute information complémentaire. Depuis 2000, aussi bien l'IEA que l'OCDE multiplient les initiatives destinées à encourager le développement et le partage des études d'analyses secondaires.

Seules les études de Dumay et Dupriez (2005) et de Simon, Turcotte, Ferne et Forgette-Giroux (2007) parues dans la revue *Mesure et évaluation en éducation* se fondent sur des analyses secondaires des données des enquêtes internationales. La rareté de tels articles peut probablement s'expliquer en partie par l'accès électronique relativement récent et simplifié aux données brutes. L'application de modèles d'analyses multiniveaux aux données des systèmes éducatifs de l'Union européenne en 2004 permet à Dumay et Dupriez d'associer la variance totale des scores en mathématiques des élèves à la différence entre les établissements scolaires selon leur composition (ou dimension contextuelle), leurs processus internes (comme les politiques de redoublement) et leur interaction. Par ailleurs, Simon et al. (2007) appliquent le V de Cramer comme mesure d'association aux données du *Progress in International Reading Literacy Study* (PIRLS 2001) afin d'examiner les tendances entre les pratiques d'enseignement et d'évaluation de la lecture dans trois populations variant selon le contexte majoritaire ou minoritaire (enseignement en langue anglaise ou française). Les deux articles rapportent

des limites associées aux bases de données respectives telles que les modalités d'échantillon (classes ou élèves), la taille relativement faible des populations en contexte minoritaire, surtout lorsqu'il s'agit de la population enseignante, ce qui limite l'application de modèles davantage appropriés, et le taux élevé de données manquantes.

Communication des résultats

Quels que soient la qualité et l'intérêt des premiers rapports de l'IEA, il ne fait guère de doute que ceux-ci n'étaient lus que par une minorité de *happy few*. Leurs finalités, leur mode d'organisation, leur temporalité étaient ceux de la recherche. Depuis, les choses ont radicalement changé. Au milieu des années 1990, les enquêtes internationales se sont retrouvées inscrites dans le mouvement plus vaste d'exigence de rendre des comptes pour les politiques publiques (*accountability*), avec d'importantes conséquences. La « pression publique » a imposé aux centres responsables une vitesse dans les analyses et la publication sans commune mesure avec ce qui se passait auparavant. Par ailleurs, les évaluations internationales sont peu à peu sorties de la confidentialité. La presse a commencé à s'y intéresser à partir des années 1980. Cette publicité des résultats fait considérablement monter les enjeux et la pression, non seulement au sein des équipes responsables des enquêtes, mais aussi au sein des systèmes éducatifs eux-mêmes, qui sont ainsi régulièrement mis en concurrence. En ce qui concerne les rapports de recherche les plus récents – rapports successifs des différents cycles de PISA d'une part, des différentes études TIMSS et PIRLS conduites sous l'égide de l'IEA, d'autre part –, une évolution distincte se dessine. De part et d'autre est publié un rapport général présentant de manière globale les résultats pour un large public, et un rapport technique destiné aux experts. Au-delà du palmarès des pays, l'OCDE[3] déploie des efforts considérables pour comprendre ce qui explique les différences de résultats entre pays et à l'intérieur des pays. Pour ce faire, des modèles d'analyse sophistiqués sont déployés (analyses multiniveaux). Il en résulte un rapport dense et assez complexe à lire par des non-experts, difficulté que l'OCDE pallie en publiant des documents plus accessibles du type résumé (*Executive summary*). Cette stratégie d'analyse et de communication s'explique à nos yeux par la volonté – non dissimulée – de l'OCDE d'influencer les politiques éducatives des États membres.

L'IEA, de son côté, semble développer une autre stratégie. Le rapport international[4] est en apparence beaucoup plus abordable, puisqu'au-delà du palmarès, seules des analyses bivariées sont présentées, croisant une par une

les variables de contexte et les performances au test. Ce rapport est essentiellement descriptif; son apparente simplicité présente à nos yeux un danger potentiel d'interprétations erronées, dans la mesure où il ne tient aucunement compte des interactions entre variables et des effets de composition. Par ailleurs, l'IEA, conforme à sa tradition, continue à investir dans l'analyse du curriculum. Le rapport international se limite en quelque sorte à la description et laisse la place aux pays pour les analyses plus explicatives. En forçant quelque peu le trait, pour les enquêtes les plus récentes, on pourrait dire qu'on navigue entre une ambition de tirer un maximum d'enseignements des résultats en termes de politiques éducatives – du côté de l'OCDE –, et une attitude réservée, du côté de l'IEA, où les rapports se limitent à des constats et à des analyses simples pour lesquels un risque d'interprétations erronées ne peut être écarté.

La montée en puissance des indicateurs d'équité

À la fin des années 1990, l'arrivée de l'OCDE sur le marché a bouleversé le paysage des enquêtes comparatives. Comme l'indique Bottani (2004), l'influence du programme OCDE/PISA sur les recherches internationales en éducation présente trois composantes: reconnaissance de la présence des gouvernements, qui sont à présent de réels partenaires dans la mise en œuvre de ces programmes, renforcement de la façon dont ces programmes sont subventionnés, collecte de données récurrentes avec l'introduction d'une perspective diachronique dans la planification de ces études.

Depuis son avènement, le programme OCDE/PISA est conçu pour produire trois types d'indicateurs. Un premier type d'indicateurs, classique, porte sur le niveau de performances des élèves; il constitue l'objet même des évaluations internationales, dès leur fondement. Un deuxième type d'indicateurs – les indicateurs de contexte – est déjà plus innovant. Si, de tout temps, les enquêtes internationales se sont préoccupées des résultats enregistrés en fonction du sexe, de l'origine sociale ou ethnique des élèves, on note dans les publications et les indicateurs publiés par l'OCDE[5] une préoccupation beaucoup plus affirmée, et donnant lieu à publication systématique, pour les indicateurs d'équité. Aux yeux de l'OCDE, il ne fait aucun doute qu'il faut non seulement encourager la performance, mais aussi l'équité des systèmes éducatifs. Au regard de la compétitivité économique et des théories du capital humain, il est en effet essentiel, non seulement qu'un pays affiche de bonnes performances moyennes, mais aussi que la proportion d'élèves disposant des habiletés jugées nécessaires pour une insertion réussie dans la société soit la plus large possible, et que ceci se fasse de préférence en minimisant les discriminations liées aux caractéristiques démographiques de l'individu.

Quant au troisième type d'indicateurs, les *indicateurs de tendance*, ils répondent au besoin de dépasser les évaluations ponctuelles et de répondre à la question – cruciale s'il en est –, et politiquement sensible, de l'évolution du niveau absolu et relatif des performances des élèves. Avec TIMSS, à la fin des années 1990, et encore plus avec PISA, les grandes enquêtes internationales récentes s'inscrivent toutes dans une perspective cyclique et ont l'ambition de mesurer non seulement les acquis au temps t, mais aussi les progrès entre le temps t et le temps t + 3, + 5 ou + 9 ans, selon la temporalité choisie.

Il s'agit là, à l'évidence, d'une avancée majeure. On conçoit aisément toute l'importance et tout l'intérêt qu'il y a à mesurer d'une manière la plus rigoureuse possible si un système éducatif a vu s'accroître ou diminuer ses performances dans le temps, par rapport à lui-même et par rapport à d'autres systèmes éducatifs. À voir aussi comment cette évolution s'est produite, via une amélioration des performances des meilleurs élèves, ou via une réduction de la proportion des élèves les moins performants. Enfin, en se dotant d'indicateurs de tendance, les évaluations internationales voient s'accroître sensiblement la puissance de leur apport aux recherches en matière de réformes et d'évaluation des politiques éducatives. Ainsi, lorsque la Pologne instaure un tronc commun jusqu'à 16 ans, on constate une élévation très significative des performances entre PISA 2000 et 2003, mais surtout une réduction importante de la dispersion des résultats, de la proportion d'élèves très faibles et de la variance liée à l'établissement fréquenté (OCDE, 2004): ce résultat constitue une évidence de recherche qui vaut son pesant d'or.

Avancée importante donc, qui va de pair avec de nouveaux défis méthodologiques. Car la recherche et l'affinement des procédures méthodologiques pour construire et interpréter ces indicateurs de tendance se font en marchant. Certes, les consortiums en charge des enquêtes contemporaines et les *Technical Advisory Groups* comptent dans leurs rangs quelques-uns des meilleurs édumétriciens du moment. Mais ces enquêtes, comme on l'a souligné, n'échappent pas aux pressions politiques de toutes sortes. Ainsi, pour ne prendre qu'un exemple, alors que l'OCDE, au début, annonçait et croyait pouvoir proposer une comparaison tous les trois ans dans les trois domaines évalués, il est de plus en plus apparu qu'il fallait se montrer davantage prudent – éviter

les comparaisons entre un domaine évalué comme majeur, puis trois et six ans plus tard comme mineur – et, pour évaluer les progrès, se fonder sur les comparaisons de domaine majeur à domaine majeur, soit tous les neuf ans.

Pour assurer la validité de ces indicateurs de tendance, l'OCDE a conçu une méthodologie qui minimise les changements d'une collecte de données à l'autre : continuité dans la définition de la population cible et des procédures d'échantillonnage, les cadres de l'évaluation et les conditions d'administration des épreuves, pour ne citer que les principales composantes méthodologiques. Néanmoins, entre deux collectes de données, des changements méthodologiques considérés comme mineurs et sans conséquences sur la validité des indicateurs de tendance ont parfois été introduits.

Dans un article paru récemment, Monseur et Lafontaine (2006) analysent l'impact que peuvent avoir différents choix méthodologiques sur les indicateurs de tendance, en particulier, pour la lecture, le choix des textes et des items retenus pour constituer l'ancrage. Ils démontrent que la sélection opérée n'est pas sans incidence sur les conclusions que l'on peut tirer quant à l'évolution des performances d'un système éducatif. Selon le choix opéré, la conclusion pourrait être, dans un cas, que les performances ont légèrement décru et dans l'autre, qu'ils ont légèrement augmenté. L'augmentation constatée entre 2000 et 2003 reste valide, pour autant qu'on limite l'interprétation de ce changement aux huit unités d'ancrage qui ont été choisies. Les auteurs concluent en insistant sur la nécessité de « privilégier l'interprétation relative de ces indicateurs de tendance au détriment de l'interprétation absolue » (Monseur & Lafontaine, 2006, p. 359).

Changements méthodologiques

Des changements méthodologiques nombreux et d'ordres divers ont affecté les enquêtes internationales au cours des trente dernières années. La place nous manque ici pour les aborder tous. Nous avons choisi de privilégier les questions qui peuvent aussi concerner les évaluations de systèmes nationales et qui, par ailleurs, ne nous entraînent pas dans des débats d'une trop haute technicité. Pour plus d'information sur d'autres évolutions méthodologiques, on se reportera à Monseur (2006) et aux articles parus dans le volume 31(2) 2008 de la revue, en particulier les contributions de Blais (2008), Ramseier (2008) et Rocher (2008).

Design des évaluations et échelles utilisées pour rendre compte des résultats

Jusqu'au début des années 1990, les résultats des pays sont présentés sous forme de scores standardisés classiques (moyenne et écart type). C'est à l'occasion de la publication du rapport de l'étude IEA-Reading Literacy (Elley, 1994) que sont utilisés pour la première fois les modèles de réponses aux items (MRI). Ils seront ensuite systématiquement utilisés dans les rapports TIMSS et PISA. L'utilisation des analyses fondées sur le MRI (Glas, 2008 ; Valois et Martin (2008) dans les enquêtes comparatives représente l'une des avancées méthodologiques les plus importantes de ces vingt dernières années. Dans la perspective de telles analyses, il convient d'anticiper, de préciser dès le début ce que chaque item est supposé mesurer et de prévoir des équilibres entre catégories. Le nombre et la nature des échelles sur lesquelles on envisage de présenter les résultats doit être fixé au départ. L'arrivée des échelles MRI force ainsi, de son côté, le passage de l'intuition, ou du bricolage savant, à une approche plus théorique, et plus complexe techniquement, de recherche d'équilibres *a priori*.

En lecture, pour ne prendre qu'un exemple, à côté de l'échelle globale tous items confondus, il a été décidé 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 couvrant l'étendue de l'échelle, du plus simple au plus complexe. Dans la construction du test, il a fallu être particulièrement attentif à construire des items de réflexion qui soient assez simples et des items de localisation de l'information qui soient suffisamment complexes. Si on laissait les choses se faire « naturellement », les items de réflexion auraient tendance à être plus complexes et les items de localisation de l'information plus simples.

Par ailleurs, le recours aux MRI rend possible l'utilisation de plans d'évaluation complexes avec rotation de carnets. Davantage d'items peuvent être utilisés, davantage de compétences évaluées sans que chaque individu testé doive pour autant subir une charge accrue. Le domaine peut être mieux couvert et les résultats peuvent être présentés sur des échelles de performances distinctes. Les différents items sont répartis dans des carnets qui ont une partie commune (par exemple 25 items sur 60) et une partie variable (les 35 autres

items varient d'un carnet à l'autre). La technique d'ancrage rend possible la comparaison des scores d'élèves qui n'ont pas passé exactement le même ensemble d'items.

Le principe de ces modèles MRI est, il faut le rappeler, qu'ils permettent d'exprimer, sur une même échelle, le niveau de difficulté d'un item et le niveau de compétence d'un individu. Un individu qui obtient un score d'une valeur n sur une telle échelle a une probabilité supérieure à un seuil fixé (on utilise habituellement 50%) de réussir les items d'une difficulté inférieure à n et une probabilité inférieure au seuil fixé de réussir les items d'une difficulté supérieure à n. Comme l'indiquent Kirsch, Mosenthal et Jungeblut (1998), «les échelles MRI permettent de prendre en considération ensemble les sujets et les tâches» (p. 105). Celles-ci permettent d'aller au-delà d'une perspective strictement normative consistant à ordonner les individus ou les pays, dans les enquêtes internationales, du plus performant au moins performant. Avant les échelles MRI, il ne restait la place, au-delà du classement, que pour une analyse qualitative. Avec l'arrivée de ces modèles, l'information se précise et devient plus diagnostique. Non seulement tel pays sait qu'il est plus ou moins performant que tel autre, mais il sait aussi en quoi: face à quel type de texte par exemple, quelle proportion des individus évalués se sont révélés capables, ou non, de mettre en œuvre telle démarche de lecture ou d'effectuer telle tâche. Ces modèles réussissent en quelque sorte à combiner de façon élégante une approche normative (qui reste au cœur de toute étude comparative) et la finesse diagnostique d'une approche critériée qui ne peut s'appliquer comme telle dans une évaluation internationale.

Modalités d'évaluation : instruments, format des questions

La modalité d'évaluation privilégiée par les enquêtes internationales reste le questionnaire papier-crayon (pour les tests cognitifs) et le questionnaire pour recueillir les données de contexte auprès des différents partenaires éducatifs (élèves, chefs d'établissement, enseignants, parents). À de rares occasions, d'autres modalités d'évaluation ont été utilisées, telles que l'observation des pratiques de classe avec grilles (IEA-PPP), ou le recours à la vidéo dans TIMSS 1995. Quelques situations de *testing* plus innovantes – situations pratiques où les élèves manipulent du matériel concret – ont également été proposées aux élèves dans le cadre de l'étude TIMSS. Dans les enquêtes les plus récentes, PISA 2006 (à titre d'option) et PISA 2009, des parties du test

ont été ou pourraient être administrées sous forme électronique. L'évaluation en ligne constitue un défi qui devra à court terme être relevé par les évaluations des systèmes éducatifs.

En attendant, l'une des questions qui a régulièrement animé les groupes d'experts ou de représentants nationaux dans les études internationales est celle du format des questions: questions ouvertes ou à choix multiple, questions fermées ou à réponse construite? En très schématiquement résumé, les questions fermées ou les QCM ont pour elles l'avantage d'être rapidement corrigées, de façon fiable et peu coûteuse; elles se prêtent mal, en revanche, à l'évaluation de certains processus (la réflexion critique notamment). Les questions ouvertes à réponse construite et les exercices pratiques encore davantage sont jugés plus authentiques, s'imposent pour l'évaluation de certains aspects, mais elles nécessitent l'élaboration de guides de correction standardisés dont la mise en œuvre est coûteuse. La comparabilité des résultats peut dans ce dernier cas se révéler plus incertaine (fidélité inter-correcteurs). Par ailleurs, si une question exige une réponse élaborée, l'évaluation des compétences se confond en partie avec les compétences en expression écrite. Voyons comment les évaluations successives ont tranché cette épineuse question dans le domaine de la lecture[6].

Les deux premières évaluations de la lecture (1971 et 1991) ne comportent que des questions à choix multiples (QCM) ou des questions ouvertes à réponse brève (un chiffre, un mot). Ce choix guidé par la prudence est cohérent avec le modèle de la lecture sous-jacent à ces évaluations. Aux différents questions posées, il n'existe en effet qu'une et une seule réponse correcte, et rien ne s'oppose donc, sur le plan théorique, à ce que des QCM et rien que des QCM soient proposées. En revanche, dans PISA et PIRLS, à côté des questions à choix multiple, on trouve une proportion importante de réponses ouvertes simples et de réponses ouvertes construites. Ce choix est, dans les deux cas, fondé non seulement sur des raisons psychologiques ou de «*face validity*» – rendre les tâches plus «authentiques» –, il est également guidé par des raisons théoriques. Certains des processus évalués supposent en effet que le lecteur mette en résonance le texte et ses connaissances antérieures. Plusieurs «bonnes» réponses sont donc acceptables et il devient difficile de couler ce type de question complexe dans le moule des choix multiples. Le passage d'une évaluation fondée essentiellement sur un questionnement à choix multiple à la nécessité, pour l'individu évalué, de rédiger ses réponses change clairement la perspective. La question de la motivation des individus

testés se pose, en relation avec celle des contextes d'administration – contexte culturel au sens large et contexte particulier à telle ou telle classe. Des recherches sont sans doute nécessaires sur ces questions qui sont loin d'avoir été résolues par les travaux antérieurs. Lafontaine et Monseur (sous presse) ont à cet égard montré qu'il existait notamment une interaction entre le sexe de l'élève et le format de question : en lecture, l'écart de performances garçons-filles est plus important pour les questions ouvertes que pour les QCM. D'une manière plus générale, les élèves les moins performants dans un domaine réussissent mieux les QCM que les questions à réponse ouverte.

Équivalence culturelle et linguistique, comparabilité

La question de la comparabilité des évaluations est clairement un enjeu majeur, en particulier dans le domaine de la lecture. En lecture, plus qu'ailleurs, la difficulté des passages et des questions apparaît liée au matériel écrit, qui peut être altérée par la traduction. L'épreuve du PIRLS 2006, par exemple, fut traduite en 44 langues et 15 systèmes éducatifs l'ont administrée en 11 langues. Celle du PISA 2000 fut administrée dans 31 systèmes éducatifs et traduite en 20 langues (Grisay, 2003, p. 223). L'enquête PISA 2003 a produit 55 versions nationales couvrant 33 langues qui ont été administrées dans 41 systèmes éducatifs. L'inégale familiarité culturelle des élèves avec les modalités d'évaluation en général et certains types de textes ou de contenus est également un sujet sensible. Sur ces questions essentielles, il est instructif de se pencher sur les dispositifs qui ont été mis en place pour assurer au mieux la comparabilité des données.

Dès 1971, un appel a été lancé aux différents pays participants afin qu'ils contribuent à la constitution d'un matériel d'évaluation d'origines linguistique et culturelle diverses, le principe étant de fournir un matériel aussi «universel» que possible, en évitant les biais et les particularismes sexuels, ethniques ou linguistiques. Les responsables nationaux, encadrés de leurs comités de référence, ont en outre été invités à réagir de façon systématique aux stimuli proposés pour l'évaluation et à estimer dans quelle mesure ceux-ci ne posent pas de problèmes particuliers dans le contexte national. Dans l'enquête IEA Reading Literacy (1991), 20 pays soumettent ainsi du matériel pour l'évaluation mais on ne sait pas combien de ces propositions ont été finalement retenues. Dans PISA, le nombre de pays contributeurs est plus étendu puisque davantage de pays sont engagés dans l'étude et la diversité d'origine est plus grande, mais les textes d'origine anglo-saxonne restent nettement majoritaires : seules 12 unités sur les 54 que comporte l'étude PISA 2000 (soit

22 %) ne sont pas à l'origine de langue anglaise. Vingt-cinq unités au moins sont en anglais (mais peuvent venir de pays différents), auxquelles il faut ajouter une partie des unités IALS reprises dans Pisa (17 unités). Les 12 unités non anglaises appartiennent à sept langues différentes. La dominante anglo-saxonne reste une réalité. En outre, dans tous les cas, sauf dans PISA, les textes proposés, même s'ils ne sont pas en anglais à l'origine, ont dû être soumis en anglais. Dans PISA, les textes peuvent théoriquement être soumis en anglais ou en français. En 2000, 5 % des unités soumises par 18 systèmes étaient soumis en langue française (Grisay, 2003, p. 230). Lors de l'évaluation PISA 2003, 15 instances ont soumis des textes dans leur langue nationale respective ou en anglais et la version source française fut élaborée entièrement à partir de la version anglaise à l'aide de l'approche à double traduction et à réconciliation (OECD, 2005, p. 71).

Plusieurs procédures de contrôle ont été mises en place pour garantir une équivalence optimale du matériel de lecture. En 1971, la première étude de l'IEA portant sur la compréhension ne s'embarrasse guère de procédures. Le *steering committee* s'en remet au bon sens et espère qu'avec un peu de soin, les tâches du test pourront être maintenues suffisamment proches pour rendre les comparaisons entre pays intéressantes et fécondes. L'étude IEARL de 1991 fournit un document de consignes pour les traductions qui doivent être entreprises par les responsables nationaux de projet. Ce document recommande que deux traductions soient effectuées par deux personnes indépendantes bilingues et qu'une *back translation* soit envoyée, ainsi qu'un rapport sur le processus de traduction. Des contrôles statistiques *post hoc* sont effectués (corrélations de l'indice de difficulté moyen des items de chaque pays avec l'indice de difficulté moyen international et l'indice de difficulté moyen des pays anglophones, la langue source du test).

Sur ce plan, c'est à l'occasion de l'étude IALS[7] que s'amorce un tournant d'importance. Non pas que l'étude IALS se soit montrée particulièrement innovatrice en la matière : celle-ci prévoit, comme IEARL, quelques consignes à respecter pour la traduction et les carnets sont «examinés avec soin pour les erreurs d'adaptation» (Murray, Kirsch & Jenkins, 1998, p. 77). Cette procédure est toutefois suffisamment lâche pour que les pays ne se sentent pas tenus de corriger les erreurs. Et la traduction, comme d'autres aspects méthodologiques, se retrouve dans l'œil du cyclone, à l'occasion de «l'incident français» (pour rappel, la France, après avoir participé aux différentes phases d'élaboration d'IALS, a mis en question plusieurs aspects de la qualité

de l'étude et s'en est retirée, en sorte que les résultats de la France ne figurent pas dans les rapports internationaux). S'ensuivra, tant du côté de l'IALS que de ses détracteurs, un intense processus de réflexion sur toutes les facettes destinées à assurer la comparabilité des données, dont les enquêtes ultérieures tireront un bénéfice majeur.

Pour ne prendre que l'exemple de PISA[8], le contrôle de l'équivalence des traductions y prend une importance névralgique, tandis que les exigences de qualité se font plus contraignantes. Originalité due à l'OCDE (qui a deux langues officielles, l'anglais et le français), le test est donc disponible en deux langues sources[9], qu'il est recommandé aux pays d'utiliser (une double traduction étant dans tous les cas nécessaire). La double traduction est soumise à la vérification de traducteurs professionnels informés de consignes précises à respecter. Un expert ressource arbitre les débats. Toutes les adaptations nationales sont soumises à une vérification serrée. Enfin, un panel de révision culturel arbitre l'ensemble des questions d'équivalence. Bref, les mailles du filet se sont resserrées et tout porte à croire qu'en matière d'équivalence des traductions, rien ne sera plus jamais comme avant... IALS. Dans la même perspective, un document édité par l'IEA (Martin, Rust & Adams, 1999) définit des standards de contrôle de la qualité des traductions particulièrement stricts : double traduction suivie d'une conciliation, vérification par le centre international de la qualité de la traduction, respect de règles précises en matière d'adaptation et surtout contrôle du respect par les pays des recommandations faites.

L'étude des méthodes de traduction utilisées lors de PISA 2000 par Grisay (2003) a démontré que la double traduction à partir de l'élaboration de deux versions sources, suivie d'une contre-vérification et d'une réconciliation, ainsi que la double traduction à partir d'une seule version source (mais avec amples contre-vérifications à l'aide de la version source de l'autre langue), menaient à significativement moins d'erreurs que la double traduction au départ de la seule version source anglaise ou la simple traduction simple. Les enquêtes subséquentes de PISA ont vu une plus grande proportion de systèmes éducatifs adhérer à ces approches (Grisay, de Jong, Gebhardt, Berezner & Halleux-Monseur, 2007).

Au-delà de l'équivalence linguistique, les enquêtes à grande échelle tentent également de réduire les biais culturels. Les idéologies et les valeurs fondamentales des systèmes éducatifs peuvent donner lieu à des différences culturelles en termes de contenus, de pratiques d'enseignement et d'apprentissage, ou encore de stratégies d'évaluation (Normand, 2004, p. 82). Ainsi, Crowne, Cytermann, Koch et Nardi (2002) expliquent qu'« [E]n France, "justifier" correspond à une démonstration : lorsque l'élève ne sait pas démontrer, il ne dit rien. En revanche, dans d'autres pays comme l'Allemagne ou la Grande-Bretagne, "justifier" consiste à laisser les traces écrites de la démarche utilisée » (p. 20). À cet égard, certaines études empiriques ont démontré des structures psychométriques internes inhérentes à différents systèmes d'éducation signifiant la présence de dimensions multiples (Ercikan & Koh, 2005 ; Grisay, 2003 ; Grisay et al., 2007). Malgré la mise en place de mécanismes rigoureux de révision intuitive et psychométrique aux fins d'établissement d'équivalence, des différences systématiques dues aux systèmes éducatifs semblent donc encore y échapper. Si ces différences sont systématiques, elles n'expliquent cependant qu'une portion modeste de la variance des performances.

En trente ans, la revue *Mesure et évaluation en éducation* n'a produit que deux textes traitant de la problématique du biais linguistique et culturel dans les évaluations de système. L'article de Ramseier (2008) examine empiriquement la validité d'un modèle de compétences à la base de standards de rendement établis aux fins d'harmonisation de la scolarité obligatoire dans les trois régions linguistiques de la Suisse (projet HarmoS). Il vérifie, à l'aide d'analyses MRI, l'adéquation du modèle dans les deux principales régions linguistiques en analysant le fonctionnement différentiel de chacun des items. « L'analyse met en évidence que la difficulté relative des items peut en général être considérée comme similaire dans les deux régions. Cette unité est perceptible au travers de la corrélation de 0,91 entre les difficultés relatives pour les deux régions linguistiques » (Ramseier, 2008, p. 15). Globalement, la compétence présente une structure similaire à travers les régions, le nombre d'items « sensibles » aux différences linguistiques ou culturelles est marginal et n'affecte pas le modèle sous-jacent.

Par ailleurs, Simon, Turcotte, Ferne et Forgette-Giroux (2007) comparent les pratiques pédagogiques et évaluatives des enseignants des écoles de langues française et anglaise en contexte minoritaire et majoritaire de l'Ontario et du Québec, en vue de mieux comprendre dans quelle mesure elles expliquent les résultats significativement inférieurs des élèves des écoles de langue française en situation minoritaire en Ontario. Leurs analyses portent sur les réponses aux questionnaires à l'intention des enseignants qui ont participé au PIRLS 2001. Puisque les pratiques des enseignants des écoles de langue

française de l'Ontario et du Québec se rapprochent davantage l'une de l'autre que celles des enseignants des écoles de langue française et anglaise de l'Ontario qui souscrivent au même programme d'études, les auteures terminent leur article en formulant l'hypothèse d'un biais culturel potentiel et se demandent si la présence d'une structure psychométrique multidimensionnelle aurait échappé aux mesures d'équivalence appliquées.

Trente ans d'évaluation des systèmes : bilan et défis à relever

Un bilan…

Ce bilan de trente années d'évaluation des systèmes éducatifs a permis de mettre au jour quelques axes d'évolution majeurs. Depuis trente ans, le défi premier des évaluations comparatives reste inchangé : il s'agit d'évaluer les acquis des élèves dans un domaine en assurant la comparabilité des données et de produire des indicateurs de rendement qui aient du crédit et de la valeur aux yeux des responsables et acteurs éducatifs des différents pays. Ce combat n'est pas gagné d'avance et les résistances, dans les pays, sont encore nombreuses. Cependant, les réponses et les solutions techniques pour relever ce défi ont sensiblement évolué depuis les années 1960. La problématique de l'équivalence culturelle et du contrôle des traductions est devenue de plus en plus sensible et a débouché sur la mise en place de *dispositifs* complexes, plus contraignants que par le passé, *de gestion et de contrôle de la qualité et de l'équivalence des traductions.*

D'autres évolutions sont par ailleurs intervenues, qui ont considérablement modifié les finalités et les retombées des enquêtes internationales. Les évaluations se sont faites plus visibles, plus ambitieuses en matière de retombées sur les politiques éducatives, ce qui fait des analyses et de la communication des résultats des enjeux majeurs. Une préoccupation de plus en plus claire est apparue pour les *indicateurs d'équité*, ce qui a conduit à élargir le débat et à dépasser le palmarès unique centré uniquement sur les indicateurs de performances. Par ailleurs, toutes les grandes enquêtes internationales proposent dorénavant, à côté des indicateurs de performances ponctuels, des *indicateurs de tendance, destinés à estimer l'ampleur des progrès.*

Sur un plan plus méthodologique, le *travail préalable d'élaboration théorique* s'est sensiblement développé au fil du temps. La démarche empirique caractéristique des premières études a progressivement cédé le pas à une approche structurée, planifiée, où les équilibres entre différentes contraintes et critères sont définis *a priori*. L'usage de carnets de tests identiques pour tous les sujets évalués a été remplacé par des *rotations de carnets avec ancrage*. Dans le même temps, la proportion de questions à choix multiple a régressé tandis que se développait l'usage de *questions à réponse construite* et de guides destinés à assurer une correction aussi objective que possible des réponses enregistrées face à des *tâches qui se sont faites plus complexes*. Enfin, la manière de présenter les résultats a été bouleversée par l'apparition des *échelles MRI*, débouchant sur des données plus diagnostiques et pertinentes aux yeux des enseignants.

Si, dès l'origine, l'IEA a choisi de *présenter des résultats de base et d'encourager l'accès aux données brutes* afin de laisser aux chercheurs chevronnés et aux décideurs politiques le soin d'exploiter plus avant les analyses secondaires, l'OCDE tend plutôt à *contrôler les analyses primaires et secondaires.* L'avènement de l'Internet facilite grandement l'accès aux données, ce qui suscite une plus grande percée, dans les années 2000, d'analyses secondaires indépendantes et le développement parallèle de modèles psychométriques appropriés aux diverses conditions méthodologiques des enquêtes.

Les évolutions intervenues se caractérisent ainsi par un double mouvement : évolutions techniques et méthodologiques, d'une part, qui se traduisent dans un rehaussement des standards de qualité et des procédures de contrôle, et évolution politique, d'autre part, qui débouche sur une visibilité nettement accrue des résultats de ces enquêtes. Il ne fait pas de doute que la puissance des résultats de recherche issus des enquêtes internationales a fait un bond en avant, en particulier au cours de la dernière décennie. Quant à la publicité faite aujourd'hui aux résultats de ces enquêtes, elle est à la fois la meilleure et la pire des choses : d'un côté, elle a contribué à ouvrir plus largement le débat sur l'école – certains systèmes éducatifs ne se sont pas encore remis du choc causé par les résultats de la première enquête PISA ; d'un autre côté, la pression médiatique qui entoure ces enquêtes fait en sorte qu'elles représentent, aux yeux de certains, l'unique étalon de mesure de la qualité d'un système éducatif. Dérive qu'il importe sans doute d'éviter, de même d'ailleurs que les interprétations hâtives ou trop confiantes dans des résultats qui, comme toute mesure, ont leurs limites. Ces résultats auront d'autant plus de valeur si on les considère en gardant la tête froide, et le sens de la mesure.

Quels défis pour l'avenir ?

Le domaine de l'évaluation des systèmes éducatifs constitue sans aucun doute un terrain fertile de recherches auquel la revue *Mesure et évaluation en éducation* pourrait s'ouvrir davantage à l'avenir. On a vu que les contributions dans la revue sont rares, récentes, et touchent à différents aspects du champ : les concepts théoriques (Nacuzon Sall & De Ketele, 1997), les analyses secondaires (Dumay & Dupriez, 2005), l'équivalence culturelle et linguistique (Ramseier, 2008 ; Simon, Turcotte, Ferne & Forgette-Giroux, 2007), les modèles MRI (Glas, 2008).

Les défis qui nous paraissent à ce jour les plus pertinents à relever par la revue touchent en particulier le débat sur le concept de compétence, la comparabilité, les modèles d'analyse, la communication des résultats et l'impact des enquêtes sur l'apprentissage des élèves.

Concept de compétence

Dans le présent article, nous avons, faute de place, limité la discussion autour du concept de compétence, ce qui ne diminue pas pour autant son importance. Le passage de l'évaluation du « bagage d'acquis scolaires » au « bagage d'acquis durables » suscite tout un débat sur le concept de compétence et sur les modèles d'évaluation appropriés. Certains systèmes éducatifs se dirigent vers des pratiques d'enseignement, d'apprentissage et d'évaluation centrées sur le développement de compétences complexes tandis que d'autres continuent à adopter une pédagogie par objectifs et des modèles de mesure traditionnellement cognitifs. Étant donné ces réalités, on peut dès lors se demander dans quelle mesure l'approche théorique actuelle de détermination des compétences jugées essentielles se démarque d'une définition de compétence réduite au « dénominateur » commun. De plus, on peut se demander en quoi l'adoption d'indices de tendances qui découlent de la nature cyclique des enquêtes oblige à conserver les modèles de mesure cognitifs aux dépens de modèles innovants, davantage basés sur une approche socioconstructiviste. Ce ne sont que quelques défis qui pourraient faire l'objet de débats et d'études empiriques à paraître dans la revue.

Comparabilité

L'équivalence de mesures adaptées en diverses langues n'est pas assurée uniquement par des processus de double traduction et de réconciliation. Pour l'heure, les enquêtes internationales présentent les résultats de chaque population sur une échelle commune déterminée à partir de procédures

psychométriques de parallélisation (*equating, scaling* ou *linking*) (Bottani & Vrignaud, 2005) et privilégient l'application de MRI pour identifier le fonctionnement différentiel d'items. Le domaine d'équivalence des épreuves présente plusieurs pistes de recherche à explorer et à débattre dans la revue dont :

a) l'examen de l'efficacité de modèles complémentaires de fonctionnement différentiel d'items provenant de populations ayant passé la même version linguistique de l'épreuve (Bottani & Vrignaud, 2005) ;
b) l'étude du fonctionnement de tests ou de sous-groupes d'items (*subtests* ou *testlests*), modèle de questionnement typique des épreuves de lecture ;
c) la comparaison des résultats de divers regroupements de populations qui se partagent la même langue (p. ex., germanique, latin, anglo-saxon, arabe) ;
d) l'application d'approches multidimensionnelles (ou modèles d'équations structurales) comme l'utilisation de la régression PLS (*Partial Least Squares*) ou les analyses en composantes principales (ACP) dans l'identification de biais linguistiques, culturels ou de genre (Bottani & Vrignaud, 2005 ; De Ketele & Gerard, 2005 ; Ercikan & Koh, 2005 ; Grisay et al., 2007) ;
e) l'utilisation des protocoles à voix haute auprès des élèves afin de mieux comprendre leur raisonnement en contexte de *testing* ;
f) l'investigation des initiatives d'équivalence des programmes d'évaluation nationaux qui s'inspirent des enquêtes internationales mais qui ne disposent pas nécessairement des moyens requis.

Modèles d'analyses des données

Le choix de modèles d'analyses des données s'applique lors du traitement primaire et secondaire des données issues des enquêtes. L'avènement des MRI a révolutionné les analyses mais a également éclipsé la discussion autour de modèles potentiellement plus puissants et sensibles à la nature et au contexte multidimensionnel des données recueillies. En effet, il existe plusieurs autres modèles statistiques qui devraient faire l'objet de comparaisons empiriques afin de déterminer leur efficacité relative dans l'établissement de l'équivalence linguistique ou culturelle (p. ex. modèle de Stout, Sibtest, etc.). Par ailleurs, en ce qui concerne les analyses secondaires, l'article de Simon, Roberts, Tierney et Forgette-Giroux, 2007 regroupe des défis d'ordre pragmatique, conceptuel et psychométrique reliés aux analyses secondaires. Ceux-ci comprennent

notamment les difficultés associées à l'accès aux données, aux conditions d'échantillon, au traitement des données manquantes, à la pertinence du design d'évaluation, à la pénurie d'experts dans ce domaine et aux critères d'application des modèles statistiques (MRI, FDI, SEM, HLM). Toutes ces difficultés devraient intéresser la revue.

Communication des résultats

Les choix méthodologiques reliés à la communication des résultats des enquêtes internationales et leur influence sur les politiques et les pratiques en salle de classe font de plus en plus l'objet d'études dans le domaine de la mesure et de l'évaluation. Goodman et Hambleton (2004) fournissent plusieurs pistes pertinentes à cet égard. Celles-ci portent essentiellement sur l'examen approfondi de l'impact de différents formats, type, fréquence et qualité de l'information fournie et sur la qualité et la nature de l'interprétation par les diverses parties intéressées, soient les cadres ministériels, les médias, les parents, les élèves, les membres du personnel enseignant.

Impact sur les pratiques en salle de classe

Enfin, la revue devrait encourager la publication d'études portant sur l'impact des enquêtes internationales sur les systèmes éducatifs, et plus précisément sur les pratiques en salle de classe. Simon, Turcotte et Forgette-Giroux (2006) ont mené une enquête sur les perceptions des représentants de divers systèmes éducatifs à l'égard de l'impact du programme pancanadien d'indicateurs de rendement sur les diverses parties intéressées. Les avis étaient partagés en fonction de l'interprétation que faisaient les participants des finalités, de l'objet d'évaluation, du modèle d'échantillon, du type de questionnement et des modalités de diffusion. Par exemple, certains des enseignants de l'étude admettaient subir l'influence de l'effet contraignant et compétitif des enquêtes et avoir tendance à modéliser leurs pratiques sur celles des enquêtes, tandis que d'autres étaient d'avis que les enquêtes offraient des sources riches et complémentaires de ressources pédagogiques et évaluatives. Parallèlement, les systèmes éducatifs variaient considérablement dans leur niveau d'intégration des éléments novateurs de l'enquête pancanadienne dans leurs programmes d'études respectifs. La revue aurait intérêt à recevoir et à publier davantage d'études de l'impact que ces variations dans l'interprétation et la réaction des divers praticiens concernés ont sur les pratiques en salle de classe.

Au cours des trente dernières années, les enquêtes comparatives, on l'a vu, ont connu des évolutions considérables tant sur le plan des finalités et de l'impact sur les politiques éducatives que sur un plan méthodologique. En raison de l'importance grandissante de ces enquêtes comparatives dans un monde globalisé où les évaluations des politiques publiques sont désormais la règle, les défis à relever pour les années à venir sont aussi nombreux que difficiles. Il serait dommage qu'une revue dont l'objet principal constitue la mesure et l'évaluation ne cherche pas davantage à solliciter des publications relevant d'un domaine qui est tout, sauf marginal.

NOTES

1. Si l'on excepte le numéro 31(2) de la revue paru en fin de 2008, dont plusieurs contributions portent sur les aspects méthodologiques de l'évaluation des systèmes. Ce numéro est hélas paru trop tard pour que nous puissions l'intégrer dans nos analyses.
2. Un article de Dolata (2008) paru dans la revue porte sur la construction de l'indice SES dans l'enquête du SAMEQ.
3. Depuis le départ, PISA est géré directement par le secrétariat de l'OCDE (direction Andreas Schleicher), mais toute l'exécution du programme est confiée sur appel d'offre à un consortium de centres de recherche. En 2000, 2003 et 2006, c'est le consortium dirigé par l'*Australian Council of Educational Research* (ACER) qui a été seul à la manœuvre. Pour 2009, la responsabilité est partagée entre le consortium dirigé par ACER, qui s'occupe de la partie cognitive, et un consortium dirigé par le CITO (compagnie de *testing* internationale située aux Pays-Bas; voir [http://www.cito.com]), qui a en charge la partie non cognitive (questionnaires).
4. Il est important de savoir que depuis TIMSS 1995, toutes les études IEA portant sur la lecture, les mathématiques et les sciences sont confiées à l'équipe de Boston College, dont on reconnaît la «marque de fabrique» dans les rapports successifs.
5. Il ne faut pas oublier qu'à l'origine du programme PISA se trouve le réseau INES A de l'OCDE, centré sur les acquis des élèves et responsable de la publication *Regards sur l'éducation*. Pendant des années, cette publication a fait fond sur les données recueillies via les enquêtes de l'IEA principalement, avant de décider, à la fin des années 1990, d'organiser sa propre collecte de données.
6. Cette question d'apparence technique véhicule celles, plus fondamentales, du lien aux pratiques de la classe et de l'effet en retour (*washback effect*) que peut exercer une évaluation dont le mode de questionnement n'est pas conforme aux usages en vigueur ou aux usages que les responsables éducatifs voudraient voir adopter par les enseignants.
7. L'étude IALS – *International Adult Literacy Study* – n'est ni une étude de l'IEA, ni une étude de l'OCDE. Mais nous l'abordons ici au tournant parce qu'elle a joué un rôle majeur dans l'évolution de la réflexion en matière d'équivalence culturelle.
8. En témoigne par exemple le document *Report on the implementation of the Pisa translation procedures,* Deelsa/Pisa/BPC(99)16, destiné à la rencontre du BPC (Board of Participating Countries) des 4 et 5 octobre 1999.
9. La comparabilité des deux versions sources est contrôlée par plusieurs experts.

RÉFÉRENCES

Beaton, A.E., Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., Kelly, D.L., & Smith, T.A. (1996). *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: Boston College.

Blais, J.-G. (2008). Les standards de performance en éducation. *Mesure et évaluation en éducation, 31*(2) 93-105.

Bottani, N. (2004). Les évaluations internationales des acquis des élèves et leur impact sur les politiques d'éducation. *Politiques d'éducation et de formation, 2*(11), 9-20.

Bottani, N., & Vrignaud, P. (dir) (2005). *La France et les évaluations internationales*. Rapport établi à la demande du Haut conseil d'évaluation de l'école. Paris: Haut conseil de l'évaluation de l'école.

[http://cisad.adc.education.fr/hcee/documents/rapport_Bottani_Vrignaud.pdf]

Campbell, J.R., Kelly, D.L., Mullis, I.V.S., Martin, M.O., & Sainsbury, M. (2000). *Framework and specifications for PIRLS Assessment 2001*. Amsterdam: IEA.

Crahay, M. (2009). Articuler l'évaluation en classe et le pilotage des systèmes, est-ce possible? In L. Mottier Lopez & M. Crahay (éds), *Évaluations en tension. Entre la régulation des apprentissages et le pilotage des systèmes* (pp. 160-172). Bruxelles: De Boeck.

Crowne, S., Cytermann, J.-R., Koch, H.K., & Nardi, E. (2002). *Les rencontres de la DESCO* (2003). *Conférence-débat « Évaluation des connaissances et des compétences des élèves de 15 ans: questions et hypothèses formulées à partir de l'étude de l'OCDE »*. [http://eduscol.education.fr/D0122/evaluation_accueil.htm, récupéré le 28 juillet 2008].

De Ketele, J.-M., & Gerard, F.M. (2005). La validation des épreuves d'évaluation selon l'approche par les compétences. *Mesure et évaluation en éducation, 28*(3), 1-26.

Dolata, S. (2008). Indice du statut socioéconomique du milieu familial des élèves du SAMEQ: construction avec le modèle Rasch et analyses. *Mesure et évaluation en éducation, 31*(1), 121-149.

Dumay, X., & Dupriez, V. (2005). Effet établissement: quelles relations entre composition sociale des établissements et processus internes? *Mesure et évaluation en éducation, 28*(2), 67-92.

Elley, W.B. (1994). *The IEA Study of reading literacy: Achievement and instruction in thirty-two school systems*. Oxford: Pergamon.

Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing, 20*(2), 225-240.

Glas, K. (2008). Item response theory in educational assessment and evaluation. *Mesure et évaluation en éducation, 31*(2).

Goodman, D.P., & Hambleton, R.K. (2004). Student test scores and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17*(2), 145-220.

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) 2007, 249-266.

Hurteau, M. (2009). Évaluation des programmes: ses visées? Qui la pilote? Qui y participe? In L. Mottier Lopez & M. Crahay (éds), *Évaluations en tension. Entre la régulation des apprentissages et le pilotage des systèmes* (pp. 77-85). Bruxelles: De Boeck.

Kennedy, A.M., Mullis, I.V.S., Martin, M.O., Kennedy, A.M., & Trong, L.T. (2007). *PIRLS 2006 Encyclopedia. A guide to reading education in the forty PIRLS 2006 countries.* Chestnut Hill, MA: Boston College.

Kirsch, I.S., Mosenthal, P.B., & Jungeblut, M. (1998). The measurement of adult literacy. In T.S. Murray, I.S. Kirsch & L.B. Jenkins (éds), *Adult literacy in OECD countries. Technical report on the First International Adult Literacy Survey* (pp. 105-134). Washington: National Center for Education Statistics, US Department of Education, Office of Educational Research and Improvement.

Lafontaine, D., & Monseur, C. (sous presse). Impact of test characteristics on gender equity indicators in the assessment of reading comprehension. *European Educational Research Journal. Special issue on PISA and gender.*

Lafontaine, D., Soussi, A., & Nidegger, C. (à paraître). Évaluations internationales et/ou épreuves nationales: tensions et changements de pratiques. In L. Mottier Lopez & M. Crahay (éds), *Évaluations en tension. Entre la régulation des apprentissages et le pilotage des systèmes* (pp. 40-55). Bruxelles: De Boeck.

Maroy, C. (2009). Régulation post-bureaucratique des systèmes d'enseignement et travail enseignant. In L. Mottier Lopez & M. Crahay (éds), *Évaluations en tension. Entre la régulation des apprentissages et le pilotage des systèmes* (pp. 55-68). Bruxelles: De Boeck.

Martin, M.O., Rust, K., & Adams, R.J. (éds) (1999). *Technical Standards for IEA Studies*. Delft, Netherlands: Eburon Publishers.

Monseur, C. (2006). *The relative character of indicators and the diachronic perspective of current international surveys in education: potential incompatibilities that call for an enhancement of the survey methodology*. Thèse de doctorat non publiée, Université de Liège.

Monseur, C., & Lafontaine, D. (2006). Le caractère relatif des indicateurs de tendance. *Revue suisse des sciences de l'éducation, 3*, 353-371.

Mottier Lopez, L., & Crahay, M. (éds) (2009). *Évaluations en tension. Entre la régulation des apprentissages et le pilotage des systèmes*. Bruxelles: De Boeck.

Mullis, I.V.S., Martin, M.O., Kennedy, A.M., & Flaherty, C.L. (2001). *PIRLS 2001 Encyclopedia. A reference guide to reading education in the countries participating in IEA's Progress in International Reading Literacy*. Chestnut Hill, MA: 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.

Murray, T.S., Kirsch, I.S., & Jenkins, L.B. (1998). *Adult literacy in OECD countries. Technical report on the First International Adult Literacy Survey*. Washington: National Center for Education Statistics, US Department of Education, Office of Educational Research and Improvement.

Nacuzon Sall, H., & De Ketele, J.-M. (1997). L'évaluation du rendement des systèmes éducatifs: apports des concepts d'efficacité, d'efficience et d'équité. *Mesure et évaluation en éducation, 19*(3), 119-142.

Normand, R. (2004). Les comparaisons internationales de résultats: problèmes épistémologiques et questions de justice. *Éducation et sociétés: revue internationale de sociologie de l'éducation, 2*, 73-89. [P 6200]

[http://www.cairn.info/article.php?ID_REVUE=ES&ID_NUMPUBLIE=ES_012&ID_ARTICLE=ES_012_0073]

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

OCDE (2001). *Connaissances et compétences pour la vie. Premiers résultats de PISA 2000.* Paris: OCDE.

OCDE (2004). *Apprendre pour le monde de demain. Premiers résultats de PISA 2003.* Paris: OCDE.

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

Ramseier, E. (2008). Validation of competence models for developing education standards: methodological choices and their consequences. *Mesure et évaluation en éducation, 31*(2), 35-53.

Rocher, T. (2008). La détermination de standards minimaux dans le cadre d'indicateurs de résultats: méthodologie, intérêt, utilité. *Mesure et évaluation en éducation, 31*(2) 75-91.

Schleicher, A. (2007). Can competencies assessed by PISA be considered the fundamental school knowledge 15-year-olds should possess? *Journal of Educational Change, 8*, 349-357.

Simon, M., Roberts, N., Tierney, R., & Forgette-Giroux, R. (2007). Secondary Analysis with Minority Group Data: A Research Team's Account of the Challenges. *Canadian Journal of Program Evaluation, 23*(3), 73-97.

Simon, M., Turcotte, C., Ferne, T., & Forgette-Giroux, R. (2007). Pratiques pédagogiques dans les écoles de langue française de l'Ontario selon les données contextuelles du PIRLS 2001. *Mesure et évaluation en éducation, 30*(3), 59-80.

Simon, M., Turcotte, C., & Forgette-Giroux, R. (2006). Impacts du PIRS en milieu scolaire. *Revue canadienne de l'évaluation de programme, 21*(1), 155-174.

Spring, J. (2008). Research on globalization and education. *Review of Educational Research, 2*(78), 330-363.

Valois, P., & Martin, R. (2008). Les modèles de mesure en éducation: enjeux, développements et orientations. *Mesure et évaluation en éducation, 31*(3), 125-153.