Comment peut-on utiliser les études comparatives internationales pour doter les politiques éducatives d'informations fiables ?

Harvey Goldstein

INTRODUCTION

Les études comparatives sur les acquis des élèves sont des programmes nombreux, coûteux, qui concernent un grand nombre de pays et sont désormais largement repris par les médias et utilisés par les décideurs politiques. Il existe deux instances rivales : l'Organisation de coopération et de développement économiques (OCDE) et l'IEA (*Association for the Evaluation of Educational Achievement*, l'Association pour l'évaluation des résultats scolaires). Elles sont largement financées par les gouvernements des pays participants. Parmi les études les plus connues, nous trouvons la série d'études PISA de l'OCDE et les études TIMSS (*Trends in International Mathématics and Science Study*) et PIRLS (*Progress in International Reading Literacy Study*) de l'IEA. Les études PISA (*Programme for International Student Assessment*) essaient d'évaluer les résultats scolaires des jeunes de 15 ans en compréhension de l'écrit, et en culture mathématique et scientifique ; les deux études TIMSS se concentrent sur les résultats en sciences et en mathématiques de deux échantillons d'élèves ayant poursuivi quatre années et huit années de scolarité. Les études PIRLS se concentrent sur les résultats en lecture des élèves en quatrième année de scolarité. Les résultats de ces études sont fréquemment utilisés pour élaborer les politiques éducatives à la fois directement pour changer les programmes scolaires ou indirectement, par la publication des résultats, ce qui exerce une certaine pression pour faire changer les choses (1).

Dans cet article, je me propose d'examiner quelques-uns des problèmes scientifiques que posent ces études, les différentes façons dont on en rend compte et leurs utilisations possibles.

PROBLÈMES SOULEVÉS PAR CES ÉTUDES

Les problèmes majeurs peuvent se formuler comme suit :

- comment choisir les contenus de ces études pour obtenir des résultats utiles ?
- comment faire l'échantillonnage des élèves pour obtenir des comparaisons valables ?
- comment construire des études pertinentes qui permettent de comparer dans le temps des systèmes éducatifs différents, ancrés dans des cultures multiples ?
- comment les données de ces études doivent-elles être analysées ?

Dans les parties qui vont suivre, je m'efforcerai de porter un regard nouveau sur ces problèmes. Je ferai référence à certains articles déjà parus sur les fondements méthodologiques de ces études dont un, celui de A. Porter et A. Gamoran (2002), qui est particulièrement utile.

CHOISIR LES QUESTIONS ET LES ITEMS DES TESTS

Ces études soulèvent tout d'abord une question très technique. Le modèle statistique prédominant utilisé dans le processus de définition et de choix des items à inclure dans les tests d'évaluation est le modèle de réponse à l'item (MRI), souvent aussi appelé théorie de réponse à l'item (TRI). Une des procédures les plus couramment choisies est celle connue sous le nom de *Rasch model* et qui n'est en fait qu'un simple modèle d'analyse factorielle dans lequel les mesures ou indicateurs sont binaires plutôt que continus. D'après les partisans de ces modèles, la sélection des items doit se faire principalement avec l'hypothèse sous-jacente qu'il n'y a qu'une seule « dimension » de réussite. Autrement dit, la réussite en compréhension de l'écrit, par exemple, implique un seul facteur ou trait individuel. La conséquence de cette hypothèse est que tous les items d'un test censés évaluer une compétence précise sont perçus comme relevant du même degré de difficulté pour tous les élèves. Le grand avantage de cette hypothèse, pour peu qu'elle soit admise, est qu'elle permet que soient adoptées toutes sortes de procédures claires de gradation statistique afin que les pays puissent être classés sur une seule échelle de valeur. Cette hypothèse est destinée à une cohérence universelle, par-delà les cultures et les systèmes éducatifs, avec des dispersions de résultats qui se manifestent seulement en termes de différences de moyennes sur une même « dimension ».

De prime abord, une telle hypothèse ne semble pas réaliste, ce qui n'est pas très difficile à démontrer empiriquement. H. Goldstein (2004) a montré comment un modèle « bidimensionnel » (c'est-à-dire construit sur deux facteurs) révèle des différences entre la France et l'Angleterre en ce qui concerne l'habitude d'utiliser les questions à choix multiples (voir aussi H. Goldstein *et al.*, 2007). Le premier facteur est un facteur général et le second tend à faire la différence entre les items à choix multiples et les items à réponses ouvertes.

Néanmoins, parce que cette méthode permet des comparaisons simples entre pays, l'envie de dépasser cette hypothèse d'« unidimensionnalité » s'avère faible et les tentatives de la tester de façon rigoureuse demeurent peu nombreuses. En effet, une grande partie de l'activité d'analyse des items lors de l'étape de conception des tests est centrée sur le rejet d'items jugés non conformes à cette hypothèse, créant ainsi une structure de test qui, pour ce qui est de la « dimensionnalité », se suffit largement à elle-même.

A. Blum *et al.* (2001) critiquent ces procédures qui, selon eux, peuvent induire des biais subtils selon les pays impliqués dans l'étude et peuvent atténuer d'importantes différences existant entre les pays. Ainsi, les rapports de PISA ne font pas référence au débat sur ces questions mais notent simplement que « les items qui fonctionnent différemment dans certains pays sont soupçonnés d'être culturellement biaisés. En conséquence, certains items sont écartés » (Kirsch *et al.*, 2002, p. 21). De tels items sont qualifiés de « douteux » et sont écartés si un minimum de huit pays présente des analyses différentes. En fait, ces items peuvent être informatifs et montrer des différences intéressantes dans les réponses faites dans les différents pays. Le fait d'exclure de tels items illustre bien l'accent mis sur des comparaisons simples entre pays au lieu d'entrer dans toute la complexité des différences.

Dans certaines situations, par exemple la certification des élèves, il peut se révéler nécessaire de rassembler les résultats à de multiples échelles en un score « unidimensionnel ». En général, ceci ne convient pas aux enquêtes comparatives internationales dont le but devrait être de mieux comprendre les différences sous-jacentes entre les pays par une analyse plus poussée. Si des comparaisons doivent être faites entre pays, alors l'existence de dimensions multiples doit se retrouver dans ces comparaisons.

ÉCHANTILLONNER LES ÉLÈVES

Les élèves des études comparatives internationales sont habituellement échantillonnés soit en fonction de leur âge, soit en fonction de la classe dans laquelle ils sont inscrits (par exemple 4e ou 8e année de scolarité). Ces deux choix comportent des inconvénients et il faut donc que les résultats en tiennent compte. Pour rendre les choses plus claires, voici un exemple de comparaisons entre la France et l'Angleterre (Goldstein *et al.*, 2007).

Un problème qui apparaît en comparant la France et l'Angleterre (tout comme pour toute autre comparaison entre pays) est que les élèves des deux systèmes progressent d'une classe à l'autre de façon différente. Par exemple, PISA 2000 est constitué à partir d'un échantillon d'enfants nés en 1984. En Angleterre, la plupart des enfants commencent leur scolarité primaire au mois de septembre de l'année scolaire correspondant à leur cinquième année. Ils ne redoublent presque jamais. Donc un jeune anglais qui a 15 ans au moment

où l'enquête PISA a été effectuée (avril/mai 2000) et qui est né en août 1984, a commencé l'école en septembre 1988 et se trouve dans sa onzième année de scolarité (l'équivalent de la classe de première du lycée français) lors du test. Dans sa classe, il y a un certain nombre d'élèves plus vieux que lui (qui ne sont pas pris en compte par PISA) qui sont nés entre septembre et décembre 1983. Cependant, la première année d'école qui s'appelle *reception* doit aussi être intégrée dans le calcul. Donc, en fait, cet enfant aura eu officiellement une scolarité de 12 ans. Un enfant né en septembre 1984 aura commencé sa scolarité un an plus tard et sera dans sa dixième année (l'équivalent français de la classe de seconde du lycée). Cet enfant-là est pratiquement du même âge que le précédent mais il a passé un an de moins à l'école.

En France, en revanche, les élèves commencent l'école au mois de septembre de l'année civile au cours de laquelle ils auront 6 ans : ils rentrent en classe de CP, qui est considérée comme la première année de scolarité. Ainsi, un enfant né en août 1984 qui n'a jamais redoublé sera en seconde (dixième année de scolarité), tout comme celui qui est né en septembre. Tous les deux auront eu le même nombre d'années d'école. Tout élève qui a redoublé une année (soit approximativement un tiers des enfants de 15 ans) se retrouvera à l'âge de 15 ans en troisième (neuvième année de scolarité). Puisque le passage du collège au lycée se fait après la troisième, ces enfants seront au collège avec des enfants qui n'ont pas redoublé, c'est-à-dire ceux nés en 1985. Ainsi, les enfants nés entre septembre et décembre 1984, qu'ils soient français ou anglais, ont le même nombre d'années de scolarité (même si en intégrant l'année anglaise de *reception*, les élèves anglais ont passé un an de plus à l'école). Au contraire, pour les enfants nés entre janvier et août 1984, la France présente une année de scolarité en moins, sans compter les conséquences possibles du redoublement.

La différence d'exposition au préscolaire varie également entre les deux pays. Quasiment 100 % des élèves français suivent l'école maternelle pendant trois ans, alors qu'en Angleterre, 80 % des élèves de 3 ans sont encore à cet âge en crèche à mi-temps. Cet exemple long et un peu technique montre combien les comparaisons internationales sont périlleuses entre des systèmes éducatifs par nature peu comparables. Ces différences d'âge d'entrée dans les cycles scolaires, d'organisation des liaisons entre les niveaux d'enseignement et de pratiques de redoublement ont des conséquences directes sur les données de PISA. Ainsi, sauf à prendre en compte la structure des systèmes scolaires, les statistiques relatives aux différences de performance entre les établissements scolaires sont surévaluées pour la France, du fait à la fois des forts taux de redoublement et de la césure entre le collège et le lycée, à l'âge stratégique de 15 ans, qui est aussi l'âge du test de PISA. Quand on évalue dans le test PISA les variations inter-établissements, pour le cas français, on ne compare pas seulement des établissements différents mais aussi la catégorie des collèges et celle des lycées entre elles, ce qui amplifie mécaniquement ces variations.

Il existe un moyen pour remédier à ce problème précis : il faut incorporer des informations longitudinales pour que les résultats antérieurs soient pris en compte. Les études transversales ne peuvent dire que très peu de chose sur les effets de l'école en tant que tels. Les écarts observés reflètent sans doute des différences entre les systèmes éducatifs, mais elles reflètent aussi, entre autres, des différences sociales qui ne pourront jamais être prises en compte complètement. Pour pouvoir comparer les effets des systèmes éducatifs, il faut (mais cela n'est pas suffisant) avoir des données longitudinales. Peu d'efforts sont faits dans ce sens ; cela reste un point faible des grandes enquêtes internationales sur les acquis des élèves. Par exemple, le rapport fondé sur PISA 2000 (OCDE, 2001) affirme que le niveau de lecture « a un effet direct clair sur les revenus bruts, l'emploi, la santé ». De telles relations de cause à effet peuvent effectivement exister mais ne peuvent pas être déduites d'une seule étude transversale.

COMPARER LES CULTURES

Beaucoup de défenseurs des études comparatives internationales mettent en avant le fait que les grandes études quantitatives internationales répondent à l'impératif de neutralité de la science (Porter & Gamoran, 2002). Si cet argument paraît plausible, il en découle des hypothèses fortes selon lesquelles la « neutralité du chercheur » et des jugements culturellement objectifs peuvent coexister. De tels concepts peuvent paraître séduisants mais sont hautement contestables. En effet, du fait des sources de financement occidentales, des modèles psychométriques occidentaux dominants et de la position centrale de la langue anglaise comme moyen de communication et comme langue de développement des items, nous pouvons faire l'hypothèse de l'existence d'un biais culturel pro-occidental dans ces études (voir Goldstein, 1995, pour en savoir plus).

Il y a peu de critiques systématiques de ces grandes études internationales. L'une des plus documentées est la seconde analyse de l'IALS *(International Adult Literacy Survey* c'est-à-dire l'Étude internationale sur l'alphabétisation des adultes) qui fut financée, dans une seconde étape, par la Commission européenne. Elle a duré plusieurs années et a impliqué des interactions entre l'équipe de départ qui a conçu l'étude pour l'OCDE et un groupe de chercheurs extérieurs. C'est pour cela que beaucoup d'idées et de critiques sur les études comparatives internationales peuvent utilement être illustrées en se référant à cette seconde analyse. A. Blum *et al.* (2001) nous livrent beaucoup de ces conclusions. Le rapport complet est aussi disponible (Carey, 2000). Cette étude a montré qu'un large ensemble de facteurs culturels peut influencer les performances des élèves.

A. Blum *et al.* (2001) donnent un exemple portant sur les différences linguistiques : il s'agit d'un questionnaire visant à évaluer une compréhension de texte. Dans ce questionnaire, les termes employés dans les questions ressemblent fortement à ceux du texte quand le questionnaire est écrit en anglais, contrairement au questionnaire rédigé en français. Par exemple, dans le questionnaire français, une question fait référence à l'emploi de « couches jetables », alors que la réponse dans le texte qui sert de base à l'exercice emploie les mots « changes complets ». Dans le questionnaire anglo-canadien, les mots « *disposable diapers* » sont répétés, de même que « *disposable nappies* » dans la version britannique. Le répondant est guidé plus facilement en anglais vers la phrase contenant la réponse et donc vers la bonne réponse, alors qu'en français, le lecteur doit comprendre que ces termes sont équivalents avant de pouvoir répondre. Il en résulte une augmentation considérable de la difficulté de la question en français.

De même, les termes anglais sont souvent plus précis et, en règle générale, les questions sont écrites de manière plus précise en anglais. Par exemple, en anglais, la question : « *What is the most important thing to keep in mind?* » (littéralement : « Quelle est la chose la plus importante à garder à l'esprit ? ») est traduite en français par « Que doit-on avoir à l'esprit ? » (littéralement : « *What must be kept in mind?* ». Cependant la phrase contenant la réponse dans le texte utilise en anglais les mots « *the most important thing* » et en français « la chose la plus importante », ce qui est la même chose. Le lien est à l'évidence plus facile à voir en anglais. Une autre tâche est définie en anglais par « *List all the rates* » (littéralement : « faire la liste de tous les taux »), Elle est traduite en français par « Quels taux » (littéralement : « *What rates?* »), en oubliant de demander « tous les taux ». Cette omission amène fréquemment les interviewés français de l'IALS à ne donner qu'un seul taux au lieu de la liste requise pour que la réponse soit considérée comme correcte.

Il existe aussi des erreurs de traduction. Certaines sont relativement peu importantes d'un point de vue strictement linguistique mais elles le deviennent au regard de la compréhension. L'exemple qui suit est caractéristique. Une question formulée en français par « soulignez la phrase indiquant ce que les Australiens ont fait pour... » (littéralement : « *Underline the sentence indicating what the Australians did to...* ») se rapporte au texte suivant : « Une commission fut réunie en Australie » (littéralement : « *A commission was set up in Australia* »). En anglais la question devient : «*What the Australians did to help decide...*», les mots du texte étant «*The Australians set up a commission.*». La réponse est ambigüe en français parce qu'il y a confusion entre le pays et ses habitants.

Ces auteurs concluent qu'il existe en fait des différences importantes de degré de difficulté des items dans des pays supposés être linguistiquement équivalents. Ils donnent aussi des exemples dans lesquels les contextes de vie réelle peuvent affecter la compréhension des items. Dans le même esprit, Wuttke (2007) soutient que PISA et par extension les études du même type, ne se préoccupent que très peu de la manière dont les élèves répondent aux items des tests et dont ils les interprètent. Il aborde aussi de nombreux problèmes liés aux stratégies d'échantillonnage choisies pour ces études. Les critères d'acceptabilité ont tendance à être fondés sur les niveaux de réussite des réponses par rapport aux objectifs fixés. Pourtant, il serait aussi intéressant de se demander si les personnes qui ne répondent pas bien sont atypiques. Peu d'énergie est déployée en ce sens. Ces conclusions sont partagées par de nombreux critiques. Ainsi, Grisay et Monseur (2007) concluent que « les équivalents des instruments de test se perdent toujours à la traduction dans une autre langue » (p. 73), et ceci quelle que soit la qualité de la traduction.

Bonnet (2002), lui, critique la qualité des données contextuelles obtenues dans PISA, surtout celles relatives au milieu socioéconomique des parents des enfants testés. Il jette ainsi un doute sur les analyses qui utilisent ces données pour mesurer le degré de relation entre les CSP des parents et la réussite scolaire. En définitive, ces critiques formulées à l'encon-

tre de l'IALS et des autres enquêtes montrent qu'un intérêt majeur de ces études pourrait résider dans la compréhension des différences culturelles exprimées à travers les réponses aux tests.

COMPARAISON DANS LE TEMPS

En 1972, deux chercheurs de la *National Foundation for Educational Research*, conduisirent une étude en changeant les critères d'évaluation portant sur la lecture du début des années soixante avec ceux de la fin des années quarante. Ils se sont servis des résultats qui avaient été obtenus par la passation répétée de ce même test sur cette période et ont montré que les programmes scolaires et l'emploi de la langue durant cette période ayant changé, c'était le test qui était devenu plus difficile et non le niveau des élèves qui s'était affaibli.

Cette dualité dans l'interprétation est connue depuis longtemps : en général, sans aller plus avant dans le travail d'investigation, on ne peut pas savoir si, par exemple, les personnes qui ont passé un test ou un examen sont dans un sens devenues « meilleures » ou si c'est le test qui est devenu « plus facile » parce que le contexte social, culturel ou scolaire a changé. Les mêmes considérations sont valables pour la comparaison dans le temps des résultats des études internationales. Cette comparaison s'appuie sur des procédures de mise en parallèle des tests (ou *test equating*) ; l'idée de base étant que l'on administre deux tests différents à deux dates différentes. Il y a plusieurs variantes mais je n'en décrirai que deux, une procédure qui repose sur la reprise d'items communs (procédure qui sous-tend beaucoup de projets concrets) et une procédure d'échantillonnage.

Dans la première approche, chaque test contient un petit nombre de questions identiques, (environ 15 % du total), c'est-à-dire un nombre suffisamment petit pour éviter qu'elles soient repérées mais suffisamment grand pour permettre des comparaisons satisfaisantes. C'est la raison principale qui fait que les études comparatives internationales conservent une série d'items non publiés. L'hypothèse est que ces items sont invariants, c'est-à-dire qu'on peut supposer qu'ils ont le même sens lors des deux administrations du test, alors que les autres items peuvent refléter des changements par rapport aux programmes scolaires, au contexte général, etc. Les items communs sont ainsi utilisés comme outil d'étalonnage afin de créer une échelle commune à tous les items des tests. Cette échelle commune est ensuite utilisée pour mesurer les changements. Concrètement, les procédures utilisées pour élaborer l'échelle varient en termes de complexité mais très souvent c'est le modèle de réponse à l'item décrit précédemment qui est utilisé.

Toutefois, ce type d'opération pose un double problème. D'abord il est nécessaire d'accepter l'hypothèse d'invariance pour les items communs et cela est inévitablement un problème de jugement. Ensuite, même si cette hypothèse est acceptée, étant donné que les items non communs peuvent refléter les changements contextuels, il faut s'attendre à ce que la relation entre la série d'items communs et les items non communs varie d'un test à l'autre ; pourtant il est nécessaire de considérer que cette relation reste constante. Cette deuxième hypothèse est donc contestable et fait apparaître en plus un problème de jugement.

Ces problèmes de comparaison dans le temps sont illustrés par le test standardisé américain, le *National Assessment of Educational Progress* (NAEP) qui connut une chute des résultats très importante sur une période de deux ans durant les années quatre-vingt (Beaton & Zwick, 1990). Une évaluation de grande ampleur avait alors conclu essentiellement que la procédure *test equating* par des items communs n'était pas fiable pour une quantité de raisons, dont le changement d'« environnement », puisqu'ils étaient associés à des items différents dans les deux instruments de test. Des problèmes similaires de comparaison utilisant des items communs sont clairement présents dans les études comparatives internationales.

Pour assurer la comparaison dans le temps, la deuxième approche consiste à fabriquer une très grande banque d'items. Pour chaque test, un échantillon d'items, éventuellement stratifié, est tiré au sort. Ceci implique que, à l'exception des erreurs d'échantillonnage, une échelle commune existe et peut être utilisée par déduction. Ces procédures sont appelées « banque d'items » (ou *item banking*), bien que ce terme soit aussi utilisé dans d'autres contextes. La difficulté est que la réserve d'items doit être constituée avant que le premier test ne soit administré et que l'on ne peut pas savoir à l'avance quels items seront démodés ou vont devenir plus difficiles avec le temps... Aussi, une fois encore, on est obligé d'émettre des hypothèses contestables sur le comportement des items des tests.

Je ne dis pas que ces approches sont inutiles, ni que les procédures de *test equating* ne sont pas utiles dans d'autres situations. Je suggère plutôt qu'elles ne sont pas des instruments objectifs simples, qui peuvent sans biais résoudre le problème de la comparaison dans le temps mais qu'en fait, elles impliquent des jugements de valeur cruciaux qui peuvent ou non trouver un consensus. Malheureusement, la plupart des écrits concernant les procédures de *test equating* ne mentionnent que rarement ces limites.

En fait, la situation est encore pire dans certaines études comparatives internationales parce que l'échantillon des pays impliqués dans les cycles successifs d'une même enquête le plus souvent évolue dans le temps, si bien que toutes les échelles et les comparaisons dans le temps qui en découlent ne concernent strictement que le groupe des pays participants.

ANALYSE DES DONNÉES

Une des caractéristiques des études comparatives internationales récentes est qu'elles essaient de prendre en compte les différences entre établissements en utilisant des modèles multi-niveaux. Un modèle multi-niveau cherche essentiellement à analyser l'ensemble des sources qui influent sur les réponses des élèves. Ainsi, par exemple, les résultats d'un élève dépendent en premier lieu de facteurs individuels tels que son sexe, son milieu social et ses résultats scolaires précédents, et en second lieu des caractéristiques de l'école fréquentée ou des écoles fréquentées auparavant. Quand nous incluons l'ensemble de ces facteurs dans un modèle statistique, nous trouvons que certaines variations inter-établissements demeurent régulièrement non expliquées par ces facteurs relatifs aux élèves et aux écoles. Les modèles multi-niveaux fournissent une explication valable de telles situations en mettant en évidence explicitement l'existence de variations résiduelles. De plus, ils peuvent facilement être étendus afin de permettre l'analyse au niveau de l'école de variations selon le sexe des élèves ou d'autres critères (pour une introduction claire et précise de ces modèles, voir T. Snijders & R. Bosker, 1999).

Étudier les variations entre les établissements permet une comparaison de second ordre qui peut être plus intéressante et plus pertinente que les comparaisons sur les simples moyennes (Goldstein, 1995). Ainsi une analyse d'items de géométrie de l'étude SIMS (*Second International Mathematics Study*), la seconde étude internationale sur les mathématiques (Goldstein, 1987, chapitre V) montre que les variations inter-établissements au Japon sont plus limitées que dans la province de Colombie britannique au Canada. Il est aussi désormais établi que les établissements scolaires diffèrent sur un grand nombre de dimensions et que la variation inter-établissements est une fonction incluant des coefficients aléatoires d'autres facteurs tels que le sexe, le milieu social, etc. Si la variation moyenne seule est considérée et qu'il y a des coefficients aléatoires importants, alors des informations importantes sont perdues. Ainsi, par exemple, l'ampleur des variations entre les pays peut varier en fonction du groupe social ou de l'éducation des parents. Si la variance inter-établissements est globalement faible, il se peut néanmoins qu'elle soit plus forte pour ceux appartenant à des groupes sociaux plus élevés ou plus bas.

Un autre problème important réside dans la manière dont les résultats sont interprétés sur les échelles produites par les analyses. Généralement, certaines limites déterminent des niveaux de résultats qui seront ensuite interprétés pour savoir ce que telle ou telle personne peut ou ne peut pas faire en lecture, mathématiques, etc. Ceci permet ensuite de calculer pour chacun des pays les pourcentages agrégés d'élèves « faibles » ou « excellents » dans telle compétence. Une fois encore, en utilisant IALS, A. Blum *et al.* (2001) ont montré comment les différentes façons de définir le découpage des résultats peuvent mener à des conclusions différentes. Ainsi, par exemple, en se servant des évaluations de IALS, 65 % des Français interrogés ont une maîtrise de l'écrit de niveau 1 ou 2 (les deux plus basses catégories), alors qu'avec une autre évaluation basée sur la notion de performance maximale, cette proportion descend à 5 %. Pour le Royaume-Uni, ces mêmes proportions sont respectivement de 48 % avec les évaluations de l'IALS et 3 % avec l'autre procédure. Inutile de dire que les secondes estimations ont moins de risque de faire la une des journaux.

Un des traits communs aux études internationales est le secret qui entoure le contenu des items utilisés. Seuls quelques items sélectionnés sont publiés. Il existe diverses raisons à cela, notamment le besoin supposé de garder des items afin de pouvoir les utiliser dans les comparaisons dans le temps. Le problème est que, à moins que les usagers puissent voir quels items sont en fait utilisés, il devient difficile, sinon impossible, de juger de ce que les tests éva-

luent réellement et si les comparaisons sont valables. L'utilisateur doit donc se fier aux jugements portés par les concepteurs des tests. De fait, cela ferme la porte à toute une longue série de débats productifs.

DOTER LES POLITIQUES ÉDUCATIVES D'INFORMATIONS FIABLES

Les sections précédentes ont présenté certaines limites des études comparatives internationales sur les résultats scolaires. Il est important de reconnaître ces limites et les contraintes qu'elles imposent pour une utilisation sérieuse des résultats. Mais dans les limites imposées, ces études peuvent tout de même donner des informations utiles.

Il existe certaines exigences élémentaires auxquelles doivent se soumettre de telles études :

- D'abord, il est important qu'on tienne compte des spécificités culturelles dans la conception des questions des tests ainsi que dans les analyses qui en découlent ;
- Les modèles statistiques utilisés dans les analyses doivent refléter la complexité de la réalité de façon à ce que la « multi-dimensionnalité » des phénomènes étudiés et les différences culturelles soient conservées au lieu d'être éliminées en faveur d'une « échelle commune » ;
- Il faut insister sur le caractère multi-niveau de toute comparaison. Il y a eu quelques essais timides à ce sujet. C'est un début prometteur. Comparer les pays sur la base de la variabilité dont font preuve les institutions et fournir des explications possibles pour les différences observées permettent de nouveaux niveaux de compréhension pour les études interculturelles ;
- Il est très important que les études comparatives s'inscrivent davantage dans des perspectives longitudinales. En n'ayant que des données transversales, il est difficile, sinon impossible, de tirer des inférences satisfaisantes sur les effets des différents systèmes éducatifs. Suivre un échantillon sur une période (même courte, une année par exemple) ajouterait énormément à la valeur de l'étude ;
- La transparence en ce qui concerne la mise à disposition de tous les items utilisés dans les tests doit être accrue afin que les utilisateurs puissent juger correctement de ce qui est évalué. Malgré les inconvénients possibles, une telle transparence doit devenir une condition essentielle ;
- Pour finir, toutes ces études ne doivent pas être perçues comme de simples instruments de classement des pays, même si elles utilisent des échelles de scores variées. Elles doivent être utilisées pour mieux analyser les différences de culture, de programmes scolaires et d'organisation de l'éducation entre les pays. Tout ceci nécessite une approche différente dans la conception des questionnaires et des items de test et l'intention de faire ressortir la diversité plutôt que d'essayer d'exclure tout ce qui est « atypique ». Ce point est abordé en détail par Langfeldt (2007). Un tel point de vue nécessite, concernant le questionnement, une approche différente de l'analyse de la structure des réponses aux items et, concernant les institutions scolaires, une collecte d'informations locales. Le processus complet, du choix des collaborateurs et des consultants jusqu'à la publication de tous les questionnaires et de tous les items des tests, doit être transparent. Les interprétations naïves des résultats de ces enquêtes réalisées par les gouvernements et les médias peuvent être non seulement sans fondement mais, plus grave encore, contre-productives. Elles peuvent même déboucher, ce qui a déjà été le cas, sur des réactions de panique qui non seulement ne s'appuient pas sur des preuves tangibles mais en plus font perdre du temps et de l'énergie au détriment d'approches plus réfléchies.

Les études comparatives internationales sur les résultats scolaires doivent être conçues comme une occasion d'acquérir des connaissances essentielles sur les raisons des différences qui existent entre les pays et non pas comme une compétition pour savoir qui arrive premier au tableau d'honneur.

Harvey Goldstein
h.goldstein@bristol.ac.uk
Université de Bristol

NOTE

(1) Les détails de ces études et des organisations qui les sponsorisent se trouvent sur leur site Internet respectif : <http://www.oecd.org/home/> et <http://www.iea.nl/> (consultés le 17 octobre 2008).

BIBLIOGRAPHIE

BEATON A. E. & ZWIK R. (1990). *Disentangling the NAEP 1985-1986 reading anomaly*. Princeton : Educational testing service.

BLUM A., GODSTEIN H. AND GUERIN-PACE F. (2001). «International adult literacy». *Assesment in Education* n° 8, p. 225-246.

BONNET G. (2002). « Reflections in a critical eye: on the pitfalls of international assessment ». *Assessment in Education,* vol.n° 9, p. 387-400.

CAREY S. (2000). *Measuring Adult Literacy. The International Adult Literacy Survey in the European Context*. Londres : Office for national statistics.

GOLDSTEIN H. (1987). *Multilevel models in educational and social research*. Londres : Griffin ; New York : Oxford university press.

GOLDSTEIN H. (1995). *Interpreting international comparisons of student achievement*. Paris : UNESCO.

GOLDSTEIN H. (2004). «International comparisons of student attainment: some issues arising from the PISA study». *Assesment in Education*, n° 11, p. 319-330.

GOLDSTEIN H., BONNET G. *et al.* (2007). « Multilevel structural equation models for the analysis of comparative data on educational performance ». *Journal of educational and behavioural statistics,* vol. XXXII, p. 252-286.

GRISAY A. & MONSEUR C. (2007). « Measuring equivalence of item difficulty in the various versions of an international test ». *Studies in educational evaluation*, vol. XXXIII, p. 69-86.

KIRSCH I., LONG J. D., LAFONTAINE D., MCQUEEN J., MENDELOVITS J., MONSEUR C. (2002). *Reading for change: performance and engagement across countries.* Paris : OEDC.

LANGFELDT G. (2007). « PISA – Undressing the truth or dressing up a will to govern? ». In S. T. Hopman, G. Brinek & M. Retzl, *PISA according to PISA*. Wien : Lit Verlag (disponible sur <http://www.univie.ac.at/pisaaccordingtopisa/> (consulté le 17 octobre 2008).

OCDE (2001). « Knowledge and skills for life: first results from Programme for international student assessment ». Paris : OCDE.

PORTER A. & GAMORAN A. (2002). *Methodological advances in cross-nation surveys of educational achievement*. Washington : National academy press.

SNIJDERS T. & BOSKER R. (1999). Londres : Sage.

WUTTKE J. (2007). « Uncertainties and bias in PISA ». In S. T. Hopman, G. Brinek & M. Retzl, *PISA According to PISA*. Wien : Lit Verlag (disponible sur <http://www.univie.ac.at/pisaaccordingtopisa/> (consulté le 17 octobre 2008).