Les comparaisons internationales de résultats : problèmes épistémologiques et questions de justice

ROMUALD NORMAND
UMR Éducation et Politiques (INRP-Lyon 2)
5, impasse Catelin,
69002 LYON

Au cours du XX^e siècle, les politiques d'éducation dans les pays industrialisés en Europe ou en Amérique du Nord évitèrent les comparaisons internationales (Mc Lean 1992). L'exception fut la panique dont furent pris soudainement les États-Unis en 1957-1958 quand ils découvrirent la supériorité soviétique ayant permis à l'URSS de lancer le premier homme dans l'espace. C'est à cette époque que l'OCDE et l'UNESCO commencent à élaborer des critères et des instruments d'évaluation des systèmes éducatifs à l'échelon international pour comparer l'efficacité de la formation scientifique et technique des différents pays. Avec la crise des années 1970-1980, les causes de la contre-performance relative du Royaume-Uni et des États-Unis comparée à la réussite du Japon et de l'Allemagne ont été imputées à l'éducation. Les bas taux de réussite scolaire ont été tenus pour responsables du manque de compétitivité dans ces pays (*A Nation at Risk* 1983). Progressivement, les comparaisons internationales des systèmes éducatifs sont devenues une référence pour les gouvernements et les organisations internationales à travers le monde, ceux-ci étant de plus en plus convaincus que l'accroissement des connaissances et des compétences de la main-d'œuvre était la clé de la compétitivité économique. Même si cette affirmation est aujourd'hui remise en cause (Robinson 1999), l'un des moyens d'assurer l'efficacité des systèmes éducatifs consiste à établir des standards nationaux et à les comparer avec ceux des autres pays.

À cette fin, de nombreux pays ont mis en œuvre des systèmes nationaux d'évaluation pour produire des données spécifiques. En France, en Australie, au Canada et en Nouvelle-Zélande, ces évaluations nationales ont été conçues pour établir un diagnostic général de la réussite scolaire des élèves. Les données fournissent aux gouvernements et aux organisations internationales des informations concernant l'impact de l'environnement social et des

caractéristiques des écoles sur les résultats. Dans d'autres pays, comme en Angleterre ou aux États-Unis, on considère que le rôle du gouvernement est de fournir aux consommateurs des informations sur les standards atteints par les écoles de façon à favoriser la compétition et élever le niveau d'excellence. Les années 1980 ont connu la mise en œuvre des procédures d'"accountability", avec une préoccupation croissante pour la qualité des systèmes éducatifs, mais c'est dans les années 1990 que la notion d'indicateurs de performance est devenue une référence pour mesurer la qualité de l'éducation. Cette volonté politique s'est appuyée sur une conception rationaliste de la mesure au moyen de standards, d'objectifs à atteindre, d'échelles de comparaison. Cette confiance inébranlable dans la technique des évaluations suscite un intérêt croissant des pouvoirs publics et s'accompagne souvent de l'idée selon laquelle un système de récompenses et de sanctions peut être attaché à des évaluations positives et négatives.

La disponibilité quasi immédiate d'indicateurs de performance grâce à la collecte de données permise par des examens ou des tests a conduit à croire que les études comparatives permettent de mettre facilement en lumière les forces et les faiblesses des systèmes éducatifs. C'est pourquoi l'OCDE s'est engagée pour plusieurs années dans un projet de définition d'une batterie d'indicateurs jugés aptes à fournir des comparaisons statistiques détaillées sur une grande étendue d'inputs, de processus et d'outputs des systèmes nationaux (Bottani & Tuijnman 1994). L'organisme est aussi engagé dans un projet international (le projet INES) pour identifier une série d'indicateurs de réussite scolaire valables pour l'ensemble des pays membres. Ce projet est repris par la Commission européenne. Pourtant, la séduction positiviste opérée par ces comparaisons internationales soulève un certain nombre de débats concernant leur justesse (c'est-à-dire leur validité, leur fiabilité ou leur utilité) mais aussi le sens de la justice qu'elles sont censées incarner. Avant d'aborder les problèmes posés par ces comparaisons, je vais m'intéresser en premier lieu aux évaluations nationales mises en place au cours des années 1980 aux États-Unis, parce qu'elles ont largement influencé les comparaisons entre États au sein des grandes organisations internationales (Papadopoulos 1994). Je montrerai ensuite que ces grandes évaluations internationales font face à un certain nombre de critiques qui en circonscrivent l'étendue et l'intérêt. Puis, j'analyserai la situation paradoxale de ces évaluations qui, alors qu'elles sont destinées à mesurer les inégalités entre élèves afin de les corriger, contribuent en réalité à les entretenir voire à les légitimer.

Une politique du compromis : le "National Assessment Evaluation Progress" (NAEP)

Le NAEP est un dispositif d'évaluation et de comparaison des performances scolaires des élèves américains créé à la fin des années 1960 à la demande du gouvernement américain. L'un des objectifs initiaux était de suivre les progrès réalisés par les élèves selon les États et d'établir leur évolution dans différents domaines d'enseignement. À cette époque, les États-Unis ne possédaient aucun moyen fiable pour juger des résultats de plus de 33 000 districts scolaires très hétérogènes. Dès sa création, le NAEP fut soumis à des pressions politiques pour qu'il ne menace pas l'autorité et l'autonomie des États et des institutions locales en charge de l'éducation. C'est pourquoi la première grande évaluation fut surtout définie par ce qu'elle ne ferait pas : les responsables du dispositif s'engageaient à ne pas collecter de données individualisées concernant les résultats des élèves, des classes, des écoles, des districts, des États afin que l'évaluation ne permette en aucun cas la définition d'un programme scolaire national. Par conséquent, beaucoup des caractéristiques du NAEP furent conçues en tenant compte des pressions politiques de tout bord, l'échantillonnage devant répondre à une technique spécifique évitant de faire passer un examen ou un test grandeur nature à l'ensemble des élèves américains. Elle devait aussi permettre d'obtenir une représentation statistique significative à l'échelon fédéral. Ces évaluations nationales concernaient des disciplines très variées comme l'art, la musique, les sciences sociales mais aussi des matières plus académiques comme la lecture, l'écriture, les mathématiques et les sciences.

Au moment de la publication du rapport *A Nation at Risk* (1983), l'attention se tourna vers le NAEP qui produisait des données nationales ou régionales mais pas de comparaison entre États. Sa structure technique et politique fut complètement réexaminée afin qu'elle permette de mieux juger des performances et des résultats des élèves américains et d'obtenir une vision globale à l'échelle du pays. De plus, afin de limiter les pressions politiques et d'assurer une large représentation, le Congrès nomma un comité (le "National Assessment Governing Board", NAGB) regroupant des gouverneurs, législateurs, responsables de l'éducation aux échelles locale et nationale, enseignants, spécialistes des disciplines scolaires, experts des procédures de tests et acteurs du monde économique et de la société civile. Ce comité devait s'occuper de fonctions comme le développement des normes de réussite scolaire, la spécification des cadres et des items des évaluations nationales, la conception de la méthodologie et des protocoles de tests, la publication des résultats et leur diffusion, enfin l'élaboration des

comparaisons aux échelons fédéral, étatique et local. Il avait aussi à juger de la pertinence cognitive des items pour éviter les biais raciaux, culturels ou de genre.

Dans les années 1990, de nouvelles décisions furent prises concernant le NAEP. Le gouvernement américain cherchait à redéfinir les responsabilités des États dans le domaine de l'éducation tout en fixant un certain nombre d'objectifs prioritaires à l'échelle fédérale : la lecture à l'école, la fixation de taux de passage d'une classe à l'autre, l'amélioration des compétences des élèves dans les disciplines fondamentales, l'objectif de rendre en l'an 2000 les élèves américains les meilleurs du monde en mathématiques et en sciences, le combat contre l'illettrisme des adultes, l'évacuation hors des écoles de la violence et de la drogue. Les évaluations du NAEP ne furent plus seulement considérées comme le baromètre national de la réussite scolaire des élèves américains, mais aussi comme la norme à l'aune de laquelle les États pourraient comparer leurs succès et leurs échecs. Le travail au sein du NAGB conduisait à définir deux grands principes susceptibles de maintenir le consensus le plus large. Le principe d'équilibre exigeait que la forme et le contenu des évaluations nationales maintiennent un compromis entre d'un côté les efforts nécessaires en termes d'instruction, les réformes de certaines disciplines d'enseignement, les résultats de la recherche sur le développement cognitif et les apprentissages et, de l'autre, les nécessités pour la nation américaine d'atteindre à terme de bons résultats scolaires. Le "principe de participation" soulignait le besoin, pour les différents acteurs, de participer activement aux délibérations et que celles-là soient gouvernées par l'équité et la justice dans le respect des différences de sexe, de race, d'ethnicité, de région géographique, de conception pédagogique.

Aujourd'hui, le NAEP fait l'objet d'un large consensus politique aux États-Unis au point qu'il reçoit même le soutien du principal syndicat d'enseignants. Connu maintenant sous le nom de "Nation's Report Card", il a évalué bon nombre d'échantillons nationaux d'élèves par niveaux de scolarité et par âges en montrant la réduction des disparités entre les élèves noirs et blancs, entre les garçons et les filles, ou des disparités à l'encontre des minorités ethniques dans le domaine de la lecture, des mathématiques et des sciences, même si les écarts demeurent importants en termes de réussite. Mais si les causes des différences entre groupes sont évoquées, les résultats eux-mêmes, dans leurs différences ou leurs similitudes, sont considérés uniquement du point de vue de leur validité et de leur fiabilité. La procédure d'évaluation est présumée techniquement neutre et les données sont considérées comme dignes de confiance. Pourtant, quelques chercheurs ont essayé de situer les résultats du NAEP en explorant ce qui pouvait rendre compte des différences ethniques ou de sexe dans la réussite (Gipps & Murphy 1994). Ils concluent que la discontinuité entre le langage des enfants et celui utilisé par

les évaluations nationales a un effet négatif sur la réussite alors que certaines tâches évaluées comportent d'importants biais culturels. La familiarité ou le défaut de familiarité avec le test conduit à augmenter ou à diminuer la probabilité de réussite des élèves selon leur appartenance sociale. Les attitudes et les attentes en termes de réussite varient aussi fortement d'un groupe à l'autre. Il existe ainsi des styles d'apprentissage très différents parmi les élèves. Par exemple, des élèves peuvent échouer à certains items ou être incapables de démontrer leur réussite, mais ils peuvent accomplir d'autres tâches que celles prévues par l'évaluateur, ou fournir des solutions adéquates quoique non prévues, ou encore faire des réponses jugées inopportunes et auxquelles les évaluateurs attribuent de mauvais scores. Des chercheurs ont montré également qu'il existait des écarts dans l'approche des problèmes entre les garçons et les filles. Elles ont tendance à choisir certaines réponses (ex : je ne sais pas) et à abandonner davantage que les garçons, ce qui ne manque pas d'affecter les performances évaluées. Malgré cela, les auteurs des rapports successifs du NAEP, dans leur volonté de comparer et d'interpréter les différences de performances, ne questionnent jamais le caractère approprié des items utilisés ni du modèle d'analyse correspondant. Pourtant, ces évaluations nationales sont utilisées comme une référence pour changer les programmes et la pédagogie dans les écoles américaines (Orfield & Kornhaber 2001). Mais ces omissions sont congruentes avec la perspective scientifique et politique incarnée par le NAEP. Les évaluations nationales ont pour but de rendre compte de résultats, non de mener une réflexion sur la validité des instruments qui les ont produits. Le rôle du NAEP est simplement d'estimer dans quelle mesure les objectifs fixés par les États ont été atteints, notamment l'augmentation des niveaux de réussite des élèves des groupes minoritaires et la réduction des disparités de réussite selon l'ethnie et le sexe. Dès lors, il n'est pas étonnant d'observer une absence totale de critique sur ce que mesurent les tests et les items.

Les grandes études internationales : comparer l'incomparable ?

On sait aujourd'hui que le NAEP a joué un rôle important dans la promotion des comparaisons internationales de résultats au sein de l'OCDE et qu'il a contribué, sous la pression américaine, à développer des travaux concernant la production d'indicateurs d'enseignement articulés à la mesure de la réussite scolaire des élèves. Mais les premières études furent conduites par l'IEA (International Association for the Evaluation of Educational Achievement) dans les années 1960, sous la direction de Torsten Husen.

L'origine de ces comparaisons est liée à l'ambition de l'IEA d'associer la réussite scolaire à un grand nombre de variables éducatives et sociales. La première comparaison internationale (First International Mathematic Study, FIMS) concernait la réussite en mathématiques et fut conduite en 1964 avec douze pays dont la France, l'Angleterre et les États-Unis. La plupart des groupes testés concentrait des sujets âgés de 13 ans ou des niveaux de scolarité ou des classes possédant une forte proportion d'élèves de 13 ans. Les domaines couverts étaient l'arithmétique, l'algèbre et la géométrie. L'étude montrait qu'il existait des écarts importants en faveur des garçons mais que dans certains pays, les filles affichaient de manière inattendue des compétences plus élevées. Les résultats montraient également que les différences entre sexes étaient moins importantes dans les écoles mixtes. S'intéressant à des facteurs explicatifs autres que les capacités intellectuelles pour les mathématiques, comme l'"intérêt des élèves pour les mathématiques" ou la "mixité dans les écoles", les auteurs de l'étude ne prirent pas en compte le fait que les contenus des tests ou leur forme pouvaient à eux seuls générer des différences. La seconde étude de l'IEA (Second International Mathematics Study, SIMS) impliquait vingt pays. Les tests furent administrés entre 1980 et 1982 en reprenant des dispositifs de la première. Les domaines des compétences évaluées comprenaient l'arithmétique, l'algèbre, la géométrie, les statistiques, tous les items correspondant à des questions à choix multiple. Chose étrange, dans certains pays (Belgique, Finlande, Suède, Thaïlande) les filles obtenaient des résultats supérieurs à ceux des garçons. Là aussi, les auteurs reconnurent que ces écarts et ressemblances entre les pays demandaient des explications complémentaires et ils s'intéressèrent aux facteurs internes (part des enseignants masculins et féminins dans chaque pays, comparaison des performances en fonction du type de tâches). Mais aucun argument ne fut avancé pour expliquer la supériorité des filles dans le domaine des fractions ou de l'algèbre.

Les auteurs du FIMS et du SIMS s'attendaient clairement à voir les garçons dépasser les filles dans tous les domaines des mathématiques à l'exception de l'oral (Gipps & Murphy 1994). Comme ce ne fut pas le cas, ils cherchèrent à comprendre les facteurs internes aux écoles pour expliquer ce phénomène. Dans leurs explications des différences de résultats, ils se centrèrent sur les facteurs environnementaux, le degré de mixité des écoles, le sexe des enseignants, l'attitude et l'intérêt pour la matière, etc. Nulle part n'apparaît une discussion sur la forme du test et sa responsabilité dans les écarts observés (Brown 1996, 1998a). De même, le contenu n'est pas critiqué bien que le curriculum varie beaucoup d'un pays à l'autre. Les études de l'IEA sont souvent si détaillées et si complexes qu'un laps de temps important est nécessaire entre la conception des items et l'analyse des données, celles-ci étant généralement dépassées quand le rapport est disponible. C'est pourquoi en

1988, une nouvelle étude internationale fut entreprise : l'International Assessment of Educational Progress (IAEP) qui devait évaluer les élèves âgés de 13 ans en mathématiques et en sciences. L'objectif était d'utiliser les mêmes items que ceux mis en œuvre dans le cadre du NAEP aux États-Unis. L'étude proposait une conception différente des performances des filles et des garçons suggérant qu'elles étaient identiques pour dix pays sur douze. On ne connaît pas la proportion des items intégrant des QCM mais il est sûr que le test n'évaluait aucune activité pratique. Il n'y eut pas davantage de discussion concernant le caractère adapté ou non de ces tests en termes de contenus ou d'items alors qu'ils avaient été conçus pour le contexte américain. L'IAEP réalisa en 1989 une seconde étude sur les mathématiques couvrant cette fois vingt pays.

Á la lecture de ces études, il apparaît que les performances des garçons par rapport à celles des filles se sont fortement réduites en mathématiques. Mais il est impossible d'en déduire quelque chose du point de vue de la réussite scolaire des uns et des autres. D'une part, il est difficile de s'appuyer sur les commentaires accompagnant ces comparaisons internationales de résultats parce que les études menées par l'IEA sont très différentes de celles de l'IAEP et que la seconde étude IAEP est aussi très différente de la première. D'autre part, les items utilisés pour la comparaison restreignent fortement la portée de ces analyses. En effet, il faut noter que les premières études de l'IEA ne donnaient pas d'information sur ces items pour éviter qu'ils ne soient enseignés au préalable. On ne dispose donc pas d'informations sur l'utilisation de ces items en mathématiques ou en sciences**.** Or, des modifications, même mineures, dans le contenu des items ou la position des items les uns par rapport aux autres peuvent changer les réponses (Goldstein 1996). De plus, le contenu ou le contexte de l'item ou de la tâche est particulièrement déterminant sur la performance des élèves, ce qui peut générer d'importants biais. En général, la majorité des items utilisés dans ces comparaisons internationales sont des questions à choix multiple pour des raisons de facilité et de rapidité. Ces tests apparaissent donc très limités parce qu'ils contribuent à accentuer les écarts de performance entre les groupes. Une autre contrainte est que les données collectées ne sont pas suffisamment détaillées pour permettre une analyse en profondeur des différences observées. Par exemple, on est capable de montrer que l'écart entre filles et garçons s'est resserré en mathématiques, mais il est impossible de dire si cela est lié à la nature de l'échantillon, aux items propres au test ou à des changements décisifs dans la réussite scolaire des filles et des garçons. Les comparaisons montrent que les déterminants des écarts sont plus environnementaux que biologiques, pourtant les auteurs ne questionnent guère la construction de l'évaluation elle-même, qui s'appuie du reste sur les tests d'intelligence et les modèles de la psychométrie. Ils n'interrogent pas davantage la contribution des items à la production des différences

ni le statut des sujets ou disciplines enseignées par rapport à l'attitude ou l'intérêt des élèves.

L'espace de justification des comparaisons internationales : un modèle politique pour l'éducation ?

Dans les comparaisons internationales, les mathématiques ont été choisies parce qu'elles apparaissaient comme une discipline universelle et culturellement neutre (Purves 1987, Husen 1987). Mais l'universalité du curriculum en mathématiques n'existe pas. D'abord parce que cette discipline d'enseignement est centrale dans certains pays et secondaire dans d'autres. Ensuite parce que certains élèves sont habitués à être testés en mathématiques toute l'année alors que d'autres sont évalués seulement à certains moments de leur scolarité. Les approches des tests sont elles-mêmes différentes entre l'oral et l'écrit, l'examen continu ou terminal, les questions à choix multiple ou les questions ouvertes (Theissen & alii 1983). Ainsi, il semble difficile d'établir des normes internationales de résultat entre pays quand le contenu de la scolarité, la structure et les formes d'évaluation sont si dissemblables. Des pays à forts taux de réussite comme les Pays-Bas ou le Japon possèdent plus de différences entre leurs pratiques éducatives que des pays très proches comme le Royaume-Uni et les États-Unis, lesquels montrent pourtant des écarts plus importants. Toutefois, malgré la faiblesse de leurs fondements techniques et scientifiques, les gouvernements se sont emparés des résultats des comparaisons internationales. En effet, alors qu'ils accordent beaucoup de crédit à la comparaison des indicateurs économiques comme les taux de croissance ou de chômage, les décideurs politiques considèrent qu'ils disposent d'indicateurs tout aussi objectifs dans le domaine de l'éducation. De fait, les comparaisons internationales de résultats sont utilisées dans de nombreux pays pour critiquer l'enseignement national et faire adopter des réformes jugées indispensables, souvent au nom d'une idéologie libérale (Apple 1989). Cela n'a pas empêché les chercheurs, initialement hostiles à ce type de comparaisons, d'en accepter finalement le principe (Duru-Bellat & Kieffer 1999). Pourtant, certains commentateurs ont pu montrer que les différences entre les moyennes nationales étaient moins importantes que les écarts à l'intérieur de chaque pays (Inkeles 1979, Noah 1987). Les effets d'agrégation dissimulent ainsi des disparités intranationales concernant les niveaux de performance des élèves et leurs causes. Par exemple, l'augmentation des taux de réussite en Grande-Bretagne masquait le fait que l'écart s'était accentué entre les meilleures et les plus mauvaises écoles. De même, lors de la seconde IAEP, les variations de réussite à l'inté-

rieur de pays comme la France et la Grande-Bretagne étaient plus grandes que celles entre les deux pays (Lees 1994). Dans son étude des performances des élèves de Hong-Kong, Winter a montré que le temps d'instruction alloué pour les mathématiques pouvait varier de manière substantielle à l'intérieur des écoles comme entre les pays, avec des effets significatifs sur les résultats (Winter 1998).

De nombreux chercheurs considèrent qu'il est nécessaire d'étudier le contexte des comparaisons internationales et d'analyser de manière approfondie les systèmes éducatifs concernés avant de se lancer dans des conclusions hâtives (Broadfoot & Osborn 1992, Holmes 1981, Neaves 1988). L'approche des comparaisons internationales suppose généralement qu'il existe une explication unilatérale des problèmes mais les différences entre pays relèvent de comportements collectifs inscrits durablement dans une histoire et une culture nationales. De même, les contenus enseignés et les styles pédagogiques expliquent ces écarts, qu'ils participent d'une épistémologie rationaliste comme en France, humaniste et individualiste comme en Grande-Bretagne ou bien pragmatiste comme aux États-Unis. La motivation des élèves et leur attitude envers les tests a été identifiée comme l'un des facteurs possédant un impact important sur les différences de performance dans les comparaisons internationales (Broadfoot & alii 2000). Pour qu'une évaluation soit valide, elle doit avoir un objectif et une signification claire pour les élèves. Or la confiance des élèves et leur intérêt pour la tâche affectent leur motivation et leur engagement. Certains élèves peuvent donner un sens particulier aux questions en inhibant leur capacité à fournir une réponse correcte. Ces problèmes se renforcent quand les évaluations cherchent à donner une signification aux problèmes en empruntant leurs exemples à la vie réelle.

Si les comparaisons internationales de résultats ont attiré l'attention des médias en contribuant à fabriquer des boucs émissaires (Brown 1998b), la recherche d'une explication des résultats a été beaucoup plus faible. Aux États-Unis en particulier, où l'opinion publique a été très préoccupée par les mauvais résultats en mathématiques des élèves, les chercheurs mettent aujourd'hui en garde contre des interprétations trop simplistes en défendant l'idée que ces comparaisons doivent être resituées dans le contexte variable d'un État à l'autre, voire d'un district à l'autre. Au Royaume-Uni, le projet Kassel (Burghes 1999), en étudiant les approches de l'enseignement des mathématiques dans treize pays, a conduit à relativiser certaines affirmations alimentant le débat public sur une faiblesse des élèves britanniques vis-à-vis de leurs pairs asiatiques révélée par l'étude récente de l'IEA (Third International Mathematics and Science Study, TIMSS). La mauvaise performance des élèves anglais s'explique par le fait qu'ils ne perçoivent pas la nécessité d'un travail soutenu en mathématiques. Des constats similaires ont été faits concernant le succès relatif des élèves japonais en mathématiques, en

remarquant qu'il existait un lien culturel fort entre l'éducation et le reste de la société : l'apprentissage et les normes de réussite y possèdent un statut élevé (Hughes 1997). D'autres auteurs ont montré que cette performance des élèves japonais résultait d'une combinaison de facteurs : une attention et une aide parentale beaucoup plus soutenue, une haute motivation des élèves, des attentes élevées des enseignants en termes de réussite, la mise en œuvre de stratégies pédagogiques très actives (Green 1999).

En Grande-Bretagne, à la fin des années 1990, le projet QUEST (Quality of Primary Education : Children's Experiences of Schooling in England and France) a montré la nécessité de réévaluer la façon dont les études comparatives sur la qualité des systèmes éducatifs peuvent être conduites (Broadfoot et alii 2000). Il visait à analyser la façon dont les caractéristiques nationales propres à la France et à l'Angleterre pouvaient avoir un impact sur les apprentissages et la façon dont les élèves se comportaient dans le cadre des évaluations nationales servant à comparer les performances entre pays. Pour cela, furent administrés les mêmes tests dans le domaine de la langue maternelle et des mathématiques aux élèves anglais et français en examinant les liens entre les résultats, les caractéristiques des classes et l'attitude des élèves dans l'apprentissage. Si les conditions d'enseignement se sont rapprochées entre les deux pays, des différences importantes continuent de subsister dans l'enseignement des mathématiques et de la langue maternelle. En France, on légitime une logique de la transmission centrée sur la réalisation de la tâche et l'application précautionneuse des formules et des procédures apprises, alors qu'en Angleterre, on favorise une logique plus inductive centrée sur le développement individuel de l'élève. Ainsi, en mathématiques, les performances des élèves reflètent les contenus des programmes nationaux et les valeurs qui y sont attachées : les élèves anglais sont meilleurs dans les recherches en mathématiques alors que les élèves français le sont davantage pour le calcul et la géométrie. De même, les élèves anglais réussissent mieux aux items leur demandant de se débrouiller et d'expérimenter, alors que les français sont meilleurs dans les items demandant une expertise technique (comme le calcul).

Ces résultats rejoignent les analyses concernant les comparaisons internationales qui montrent que les performances des élèves aux tests de mathématiques peuvent être affectées par les contenus curriculaires et par le degré de familiarité de certains élèves avec les items. L'analyse approfondie des résultats a montré qu'il existait deux mondes opposés dans l'approche des mathématiques de part et d'autre de la Manche. Les élèves anglais mettent en œuvre une démarche expérimentale et individualiste, pensant trouver par eux-mêmes la solution à leurs problèmes. Les élèves français ont une action plus technique et structurée. Les premiers sont plus "explorateurs" et les seconds plus "techniciens". L'ensemble de l'étude confirme que les élèves

anglais et français se conforment à des cadres nationaux dans leur apprentissage des mathématiques et de leur langue maternelle. Les élèves apprennent au travers d'outils culturels créant d'importantes différences en termes de curriculum, de procédures d'évaluation et de tests, de contextes de classe et de stratégies pédagogiques. En conséquence, les conceptions des apprentissages et la participation aux évaluations nationales varient fortement sans que l'on puisse en tirer un quelconque enseignement sur la force ou la faiblesse d'un pays par rapport à l'autre.

Sens de la justice et comparaison par les tests : les limites d'une approche techniciste

Après le rejet des politiques compensatoires et d'égalisation des chances des années 1970-1980, les comparaisons internationales se sont intéressées aux résultats afin d'évaluer l'égalité réelle des chances, c'est-à-dire ce qui était directement disponible et observable. On passa donc d'une conception référée à l'égalité des chances à une autre référée à l'égalité de résultats (Crahay 2000). Cette nouvelle définition alimente aujourd'hui les discussions relatives à la réussite des groupes sociaux aux évaluations nationales et internationales. Ces dernières, qui évaluent la réussite scolaire des élèves, sont présentées comme des mesures "objectives" utilisées dans les études concernant l'égalité des chances et deviennent progressivement un élément important du débat sur l'équité des systèmes éducatifs. L'approche psychométrique traditionnelle, sur laquelle s'appuient ces études, fait l'hypothèse que des solutions techniques peuvent être trouvées pour résoudre ces problèmes d'équité, notamment en élaborant des procédures permettant d'éliminer les biais dans les items. L'usage de procédures statistiques permet de déterminer si les questions du test sont particulièrement difficiles pour certains groupes une fois que la performance d'ensemble est prise en compte. Cette forme d'analyse d'item est appelée "Differential Item Functionning" et permet de distinguer les biais occasionnés par les items et ceux occasionnés par les tests (Goldstein & Lewis 1996). Les premiers concernent les questions qui favorisent tel ou tel groupe de manière disproportionnée, tandis que les autres renvoient à la moyenne des scores pour les différents groupes. Ils montrent par exemple qu'utiliser des mots connus par un seul des groupes contribue à augmenter les biais du test. De même la réduction des biais liés aux items permet d'augmenter la validité du test. Mais cette approche ne s'intéresse pas à la façon dont la matière évaluée est définie (c'est-à-dire l'ensemble du champ disciplinaire à partir duquel les tests sont choisis) pas plus qu'à la sélection des items jugés les plus pertinents sur le plan didactique. De même, alors

qu'elle se limite à une réflexion concernant la manipulation de l'évaluation, elle masque l'importante contribution d'autres facteurs comme la perception par les élèves des matières évaluées, les expériences qu'ils mobilisent pour une matière, et le type de demandes qui leur sont adressées.

De même, on considère généralement que les tests délivrent une évaluation impartiale même s'il n'y a pas d'égalité des chances précédant le test. Selon cette conception, l'utilisation de tests standardisés donne à tous les enfants les mêmes chances de compétition et constitue un cas d'égalité des chances puisqu'il s'agit d'un traitement identique sans considération pour les variations initiales dans la préparation au test. Les opposants à l'usage des tests standardisés pour les enfants des groupes défavorisés rétorquent que ces dispositifs ont des effets sociaux négatifs parce qu'ils ignorent l'importance de la dimension individuelle dans les apprentissages et qu'ils ne contribuent guère à réduire les écarts de performances entre les groupes sociaux (Gillborn & Youdell 2000, Kohn 2000). Pour eux, l'égalité des résultats n'est pas un objectif approprié parce que différents groupes n'ont pas forcément les mêmes qualités, capacités et expériences. Manipuler les items et les procédures de tests de façon à produire une égalité de résultats tend à ignorer la construction des compétences évaluées et à masquer des dissemblances originelles importantes. La première conception, largement dominante, considère qu'il faut détacher les problèmes de validité et de biais des tests du contexte de leur administration, en s'attachant uniquement au respect de normes techniques. La deuxième conception suggère que la validité et les biais des tests doivent être rattachés à des questions de justice, notamment quant aux conséquences sociales résultant de leur usage. Ainsi, un biais peut être défini comme une forme d'invalidité du test se définissant par rapport à certains groupes sociaux (Camilli & Shepard 1994). Le test est considéré comme biaisé à l'encontre d'un groupe particulier s'il est incapable de prédire correctement la performance de ce groupe sur la base d'un critère commun aux autres groupes.

Il faut toutefois distinguer les biais concernant la prédiction (predictive bias) et ceux se rapportant aux critères de validité (criterion bias) (Howe 1997). Le premier type, prédictif, peut-être "externe" ou "interne". Un biais prédictif "externe" se rapporte à la prévision réelle de la performance que le test est censé mesurer. Par exemple, les résultats à un test d'accès à un certain niveau de scolarité doivent correspondre à la performance réelle des élèves une fois qu'ils y ont accès. Un biais prédictif "interne" se rapporte aux caractéristiques internes du test, à savoir les écarts apparaissant parmi les items. Si un test en mathématiques contient un item mobilisant des connaissances sur le football et que les filles possèdent des résultats inférieurs pour cet item à ceux des garçons, mais aussi par rapport à l'ensemble de leur performance au test, on dit que l'item est biaisé. Par contre, si ces filles ont des résultats inférieurs pour l'item mais que cela est valable pour l'ensemble de leurs perfor-

mances au test, on considère que le test n'est pas biaisé. Pour autant, l'absence de biais internes ne garantit pas une meilleure réalisation de la justice s'il subsiste des biais prédictifs externes (par exemple si les filles réussissent moins bien au test que les garçons alors qu'elles obtiennent des performances plus élevées par la suite dans leur scolarité). À la différence des biais prédictifs où l'on part de critères fixés au préalable, les biais relatifs aux critères permettent de voir si le critère de performance lui-même est biaisé vis-à-vis de certains groupes sociaux, indépendamment de la façon dont il est corrélé aux scores de réussite. Ce second type de biais peut prendre deux formes : transversale ou interne aux groupes. Les biais transversaux apparaissent quand les résultats aux tests dépendent fortement d'aptitudes ou de compétences sociales sans lien direct avec l'évaluation, comme on le constate dans des procédures de forte sélection des élèves ou des étudiants. Les biais internes résultent du fait que les critères du test, même parfaitement adaptés et appropriés au domaine évalué, et faisant même apparaître les avantages et les désavantages associés à certains groupes, peuvent quand même pénaliser un groupe social, du fait de configurations institutionnelles ou historiques déterminées. Par exemple, aux États-Unis, comme la pédagogie et le curriculum favorisent les individus de race blanche, les critères sont biaisés en faveur des blancs au détriment des noirs et des minorités ethniques.

On comprend donc que la justice soit mieux assurée si l'on élargit la conception technique des tests en envisageant simultanément, en plus des problèmes posés par leur construction, ceux concernant leur capacité prédictive et les critères utilisés au regard de certains groupes sociaux. En plus de ces considérations épistémologiques, deux postulats sont généralement admis pour défendre les évaluations et les comparaisons par les tests. Dans le premier cas, on considère que l'adaptation de la performance des élèves à celle exigée par les tests est le meilleur moyen de réaliser les programmes et d'instruire les élèves. Toutefois, il est facile de constater, au regard des expériences anglo-saxonnes ou américaines, que l'imposition de standards conduit à des dérives dans les pratiques pédagogiques (teaching to the test) et qu'ils n'améliorent pas la qualité du contenu des apprentissages, les effets positifs de ces évaluations étant en définitive très limités (Kosol 1991, Sacks 1999, Normand 2001). Le second postulat selon lequel les tests favoriseraient l'équité n'est guère plus convaincant. En effet, ils ne permettent pas d'éliminer les biais concernant la validité des critères à l'encontre de certains groupes sociaux. Pire, avec le renforcement des procédures d'évaluation, les différences de performances parmi les groupes auraient tendance à persister, voire à augmenter (Apple 1993, Madaus 1994). Les tests utilisés risquent alors d'exacerber les questions d'injustice en les masquant derrière une défense techniciste. Celle-ci vise à concilier des principes d'efficacité avec un idéal d'égalité de résultats remplaçant l'idéal d'égalité des chances et présenté

comme un passage nécessaire de l'utopie au réalisme (Derouet 2000, 2003). Mais cette nouvelle philosophie politique risque de donner plus d'importance aux exigences de qualité et d'efficacité au moment où les débats sur la justesse technique des instruments d'évaluation l'emportent sur l'examen de la justice des objectifs politiques de l'école.

Conclusion

À la lumière des développements précédents, il demeure une question essentielle : à quoi servent en définitive les comparaisons internationales de résultats ? En dépit de ses limites épistémologiques, il s'agit d'un processus technique qui possède un intérêt pour ses promoteurs : il permet de transformer les qualités et les capacités hétérogènes des élèves en une même mesure, qui est utilisée ensuite pour informer les usagers et les administrateurs de l'éducation, établir des classements entre pays et juger d'une compétition internationale, justifier des décisions politiques concernant l'amélioration de l'efficacité et de la qualité des systèmes éducatifs. La comparaison statistique réduit et simplifie le traitement de l'information grâce à des chiffres que les individus et les institutions parviennent plus facilement à manipuler (Desrosières 2003). Tout en facilitant la prise de décision, elle permet de réduire l'incertitude, d'imposer un contrôle et de fournir une légitimité incontestée grâce à des catégories auxquelles les acteurs de l'éducation n'ont plus qu'à se conformer. Les comparaisons sont essentielles pour les organisations internationales qui utilisent des modèles de décision et souhaitent donner consistance et objectivité à des savoirs et des pratiques éducatives très éloignées sur le plan culturel et géographique. Une fois ces catégories statistiques élaborées, elles sont routinisées dans des instances et des lieux de décision, des lois ou des règlements, des recommandations faites aux États (Desrosières 2000, Thévenot, 1997). Les comparaisons deviennent alors des formes publiques de la connaissance qui tendent à relativiser les particularismes culturels et les savoirs locaux au nom d'une méthode scientifique rigoureuse générant une information distanciée et officielle.

Un autre aspect, tout aussi important, est que ces comparaisons internationales mettent en rapport des statisticiens, des évaluateurs, des chercheurs, des administrateurs ou des représentants des gouvernements et des grandes organisations internationales, dans des lieux ou le long de réseaux permettant la confrontation et l'échange d'expériences, la mise en place de projets communs, la définition de nouvelles procédures et de nouveaux instruments. Ces espaces d'intéressement donnent une place centrale à l'expertise dans la définition de problématiques essentielles au devenir des systèmes éducatifs au niveau européen et international (Dutercq 2001, Derouet &

Normand 2003). S'y trouve, en effet, défini un référentiel qui s'appuie sur une vision commune et partagée : le capital humain est un facteur essentiel à une économie fondée sur la connaissance dont il faut chercher à améliorer l'efficacité et la rentabilité. Cependant, outre son économisme, cette proposition tend à relayer une conception objectiviste de l'éducation dont on commence à évaluer les effets pervers sur le management des écoles et les pratiques pédagogiques (Normand 2003). En fait, cette confiance excessive dans les conclusions d'une expertise tend à confisquer le débat démocratique en empêchant une réflexion collective sur le projet politique de l'école (Charlier 2003). Cette dérive rend de plus en plus nécessaire la création de lieux de concertation publique à l'échelon européen ou international où s'exprimerait la diversité des intérêts dans le respect de principes universels concernant l'éducation et qui fourniraient l'occasion de promouvoir une véritable démocratie technique afin de mettre à l'épreuve les postulats ou les affirmations les plus contestables.

Bibliographie

A Nation at Risk 1983 National Commission on Excellence in Education, Cambridge (Mass.), USA Research

APPLE M. 1989 How equality has been redefined in the conservative restoration, in Secada W. (ed.) *Equity and Education*, New York, Falmer Press

APPLE M. 1993 "The politics of knowledge : Does a national curriculum make sense ?, *Teachers College Record, 95(2)*, 222-241

BOTTANI N. & Tuijnman A. 1994 Évaluer l'enseignement : de l'utilité des indicateurs internationaux, Paris, OCDE

BROADFOOT P. & OSBORN M. F. 1992 "Lessons : comparative perspectives on what it means to be a teacher", Philips D. (ed.) *Oxford Studies in Comparative Education 1*, Wallingford, Triangle

BROADFOOT P., OSBORN M., PANEL C. & SHARPE K. 2000 *Promoting Quality in Learning. Does England Have the Answer ?* Findinds from the Quest Project, London, Cassell

BROWN M. 1998a "FIMS and SIMS : the first two IEA International Mathematics Surveys", *Assessment in Education, 3(2),* 193-212

BROWN M. 1998b The Tyranny of the International Horse Race, in Slee R., Weiner G., Tomlinson S. (eds), *School effectiveness for Whom ? Challenges to the School Effectiveness and School Improvement Movements*, London, Falmer Press

BROWN M. 1996 "International comparisons and mathematics education : a critical review" *Oxford Studies in Comparative Education, 3(2),* 193-212

BURGHES D. 1999 "The kassel project : an international longitudinal comparative project in secondary mathematics", *Oxford studies in comparative education, 9(1)*, 135-155

CAMILLI G. & Shepard L., 1994, *Methods for identifying biased test items*, Thousand Oaks (CA), Sage

CHARLIER J.-É. 2003 Les citoyens et les politiques éducatives européennes, in Derouet J.-L. & Normand R., *L'Europe de l'éducation : entre management et politique*, Paris, PUF, à paraître

CRAHAY M. 2000 *L'école peut-elle être juste et efficace ?*, Bruxelles, De Boeck

DEROUET J.-L. 2000 La sociologie des inégalités d'éducation à l'épreuve de la seconde explosion scolaire, *Éducation et Sociétés, revue internationale de sociologie de l'éducation, 5*, 9-24

DEROUET J.-L. 2003 La sociologie des inégalités d'éducation dans une société critique. Petit guide à l'usage de ceux qui veulent explorer le pays de la pensée gnangnan, in Van Haecht A. (dir.) *Sociologie, politique et critique en éducation*, *Revue de l'Institut de Sociologie*, Bruxelles, ULB, 41-55

DEROUET J.-L. & Normand R. (dir.) 2004 *L'Europe de l'éducation : entre management et politique*, Paris, PUF, à paraître

DESROSIÈRES A. 2000 *La politique des grands nombres. Histoire de la raison statistique*, Paris, La Découverte

DESROSIÈRES A. 2003 Comment fabriquer un espace commun de mesure : harmonisation des statistiques et réalisme de leurs usages, in Lallement M. & Spurk J. (éds.), *Stratégies de la comparaison internationale*, Paris, Éditions CNRS

DUTERCQ Y. (DIR.) 2001 *Comment peut-on administrer l'école ?*, Paris, PUF

DURU-BELLAT M. & Kieffer A. 1999 "La démocratisation de l'enseignement 'revisitée' : une mise en perspective historique et internationale des inégalités de chances scolaires en France", *Les cahiers de l'IREDU, 60*

GILLBORN D. & YOUDELL D. 2000 *Rationing Education. Policy, Practice, Reform and Equity*, Buckingham, Open University Press

GIPPS C. & MURPHY P. 1994 *A fair test ? Assessment, achievement and equity*, London, Open University Press

GOLDSTEIN H. 1996 Statistical and Psychometric Models for Assessment, in Goldstein H. & Lewis T. (eds.), *Assessment : problems, developments and statistical issues*, Chichester, John Wiley & Sons, 41-56

GOLDSTEIN H. & LEWIS T. (eds.) 1996 *Assessment : problems, developments and statistical issues*, Chichester, John Wiley & Sons

GREEN A 1999 "Converging paths or ships passing in the night : an English critique of Japanese school reform", *Comparative Education, 36(3)*

HOLMES B. 1981 *Comparative Education : some considerations of method*, London, Georges Allen and Unwin

HOWE K. R. 1997 *Understanding equal educational opportunity : social justice, democracy, and schooling*, New York, Kenneth R. Howe, Teachers college press

HUGHES M. 1997 "The National Curriculum in England and Wales : a lesson in externally-imposed reform ?", *Educational Administration Quaterly*, *33(2)*

HUSEN T. 1987 "Policy Impact of IEA Research", *Comparative Education Review, 31(2)*, 29-46

INKELES A. 1979 "National Differences in Scholastic Performance", *Comparative Education Review, 23(3)*, 386-407

KOHN A. 2000 *The case against standardized testing : Raising the scores, ruining the schools*, Portsmouth, N. H., Heinemann

KOSOL J. 1991 *Savage inequalities*, New York, Crown

LEES L. H. 1994 "Educational inequality and academic achievement in England and France", *Comparative Education Review, 38(1),* 65-116

MADAUS G. 1994 "A technological and historical consideration of equity issues associated with proposals to change the nation's testing policy", *Harvard Educational Review, 64(1),* 76-95

MC LEAN M. 1992 *The Promise and Perils of Educational Comparison*, London, The Tufnell Press

NEAVES G. 1988 "On the cultivation of quality, efficiency and enterprise : an overview of recent trends in higher education in Western Europe", *European Journal of Education, 23(1-2),* 7-24

NOAH H. J. 1987 "Reflections", *Comparative Education Review, 31(1)*, 137-149

NORMAND R. 2003 "Le mouvement de la 'School effectiveness' et sa critique dans le monde anglo-saxon" in Van Haecht A. (dir.) *Sociologie, politique et critique en éducation*, *Revue de l'Institut de Sociologie*, Bruxelles, ULB, 135-166

ORFIELD G. & KORNHABER M.-L. 2001 *Raising standards or raising barriers ? Inequality and high-stakes testing in public education*, New York, The Century Foundation Press

PAPADOPOULOS G.-S. 1994 *L'OCDE face à l'éducation 1960-1990,* Paris, OCDE

PURVES A. 1987 The evolution of the IEA : a memoir, *Comparative Education Review, 31(1),* 10-28

ROBINSON P. 1999 The tyranny of league tables : international comparisons of educational attainment and economic performance, in Alexander R., Broadfoot P. & Phillips D. (eds) *Learning From Comparing, 1 : Contexts, Classrooms, and Outcomes*, Wallingford, Symposium Books

SACKS P. 1999 *Standardized minds : The high price of America's testing culture and what we can do to change it*, Cambridge, MA, Perseus Books

THEISSEN G. L., ACHOLA P. W. & BOAKARI F. M. 1983 "The Underachievement of Cross-National Studies of Achievement", *Comparative Education Review, 27(1),* 46-68

THÉVENOT L. 1997 Un gouvernement par les normes. Pratiques et politiques des formats d'information, in Conein B. & Thévenot L. (dir.), *Cognition et information en société*, *série "Raisons Pratiques"*, 8, EHESS, 1997, 205-242

WINTER S. 1998 "International comparisons of student achievement and the Asian educational phenomenon : a critical analysis", *Comparative Education, 35(2)*