Comparer les résultats des systèmes éducatifs nationaux : les défis méthodologiques des enquêtes PISA

Noémie Le Donné

Comparer les résultats des systèmes éducatifs nationaux :

les défis méthodologiques des enquêtes PISA

Noémie Le Donné

Notes & Documents

n° 2013-06 Octobre 2013

Résumé :

Les enquêtes PISA (Programme International pour le Suivi des Acquis des Élèves) font actuellement office de référence dans le panorama des évaluations internationales des élèves. Elles ont donné lieu à de nombreux rapports nationaux et travaux comparatifs en éducation. Cette note s'adresse aux utilisateurs des données PISA, ainsi qu'à tous ceux qui souhaitent porter un regard distancié sur ces enquêtes et leurs exploitations. Elle expose de manière réflexive, sous les angles statistique et sociologique, la structure, les atouts et les limites du dispositif PISA depuis sa première édition en 2000. Elle cherche également à examiner la manière dont PISA répond aux défis méthodologiques des comparaisons internationales. Dans un premier temps, nous présentons les objectifs pour lesquels le programme PISA a été initialement conçu, à savoir comparer les résultats des systèmes éducatifs dans leur ensemble. Nous montrons ensuite que ce changement de cap des évaluations internationales est à l'origine d'un certain nombre d'innovations méthodologiques pour assurer la comparabilité des données produites. Ces choix méthodologiques ne sont pas sans conséquence sur la manière de traiter les informations recueillies. Au fil de cet examen critique ainsi qu'en conclusion, nous précisons donc le type d'analyse secondaire à privilégier à partir des enquêtes PISA

Pour citer ce document :

Le Donné, Noémie (2013). « Comparer les résultats des systèmes éducatifs nationaux : les défis méthodologiques des enquêtes PISA », Notes & Documents, 2013-06, Paris, OSC, Sciences Po/CNRS.

Pour une version électronique de ce document de travail et des autres numéros des Notes et Documents de l'OSC, voir le site web de l'OSC : http://www.sciencespo.fr/osc/fr/content/notes-documents-de-l-osc

Abstract:

PISA (Programme for International Student Assessment) is today a reference in the field of international student assessments. It has been used in amounts of national reports and comparative educational studies. This note is intended for PISA data users as well as for those who want to take a distant look at these data and their use in the social sciences. It presents the structure, the strengths and the limitations of PISA surveys since their first edition in 2000, in statistical and sociological perspectives. It also investigates the extent to which PISA design deals with the methodological challenges of international comparisons. We firstly present the new targets that PISA is designed to meet, namely comparing the global performance of educational systems. We then show that this shift in international assessments has triggered a number of methodological innovations to ensure data comparability. These methodological choices are not without consequences on how to handle the information collected. Throughout this critical review and in conclusion, we specify the type of secondary analysis one should conduct on PISA surveys.

Readers wishing to cite this document are asked to use the following form of words:

Le Donné, Noémie (2013). "Comparer les résultats des systèmes éducatifs nationaux : les défis méthodologiques des enquêtes PISA", Notes & Documents, 2013-06, Paris, OSC, Sciences Po/CNRS.

For an on-line version of this working paper and others in the series, please visit the OSC website at: http://www.sciencespo.fr/osc/fr/content/notes-documents-de-l-osc

Le programme PISA (Programme International pour le Suivi des Acquis des Élèves ou *Programme for International Student Assessment* en anglais), piloté par l'OCDE (Organisation de Coopération et de Développement Économiques) a pour vocation de déterminer dans quelle mesure les élèves ont acquis, au terme de la scolarité obligatoire, les savoirs et savoir-faire nécessaires à leur pleine participation à la « société de la connaissance ». Les enquêtes PISA sont réalisées tous les trois ans auprès d'élèves de 15 ans de nombreux pays. Trois disciplines sont évaluées à chaque vague d'enquête et sont tour à tour érigées en domaine majeur : en 2000, la compréhension de l'écrit était à l'honneur, en 2003 la culture mathématique et en 2006 la culture scientifique. Le programme suit des cycles périodiques de neuf ans, composés de trois enquêtes. En 2009, PISA a entamé son deuxième cycle en testant de nouveau en priorité les compétences des élèves en compréhension de l'écrit.

Ces enquêtes font actuellement office de référence dans le panorama des évaluations internationales des élèves[1]. PISA a rencontré un large écho à la fois dans le monde médiatique et dans la communauté scientifique[2], pour des raisons en partie différentes. Une raison du succès de PISA commune aux deux sphères tient à la couverture géographique du programme. Le nombre de participants, déjà substantiel à la première édition en 2000 (32 pays dont 28 pays membres de l'OCDE) n'a cessé de croître, pour atteindre 75 pays et économies partenaires (dont 34 pays membres de l'OCDE) lors de l'édition 2009. L'étendue géographique du programme est appréciée des chercheurs en éducation adeptes des approches comparatives, tout comme des médias qui sont en mesure de diffuser un palmarès international des résultats des systèmes éducatifs.

Le retentissement de PISA s'explique ensuite par la richesse des données collectées par ces enquêtes, et par les objectifs novateurs auxquelles elles répondent. Les bases de données PISA offrent en effet un éventail d'informations sans précédent. Deux séries de facteurs sont pris en considération. Les premiers sont spécifiques aux élèves : leurs caractéristiques sociodémographiques, leur environnement familial, les ressources du domicile, leur expérience de l'école, leurs méthodes de travail et leurs projets. Les deuxièmes relèvent des établissements : leur secteur, leurs ressources financières, matérielles et

[1] Il semble qu'elles soient notamment les premières études statistiques à faire l'objet d'une présentation détaillée dans un petit manuel de la collection « Que sais-je ? » (Félouzis et Charmillot 2012).
[2] La notoriété du programme PISA en France est relativement tardive en comparaison d'autres pays, comme l'Allemagne qui a connu un vrai « PISA-Schock » dès la parution des résultats de la première édition en 2000. Il a fallu attendre la publication de l'ouvrage *L'Élitisme républicain* de C. Baudelot et R. Establet en 2009 pour attirer l'attention de l'opinion sur les résultats de la France à PISA.

pédagogiques, leur fonctionnement et le mode de recrutement des élèves. Les enquêtes PISA ouvrent la voie à de nombreuses analyses comparatives des liens entre les compétences des élèves, leurs caractéristiques et celles de leurs établissements. PISA apporte ainsi de nouveaux éclairages quant à l'efficacité et l'équité des systèmes éducatifs.

Enfin, PISA tient sa force de l'abondante documentation mise en ligne gratuitement sur l'Internet par l'OCDE, qu'il s'agisse des rapports techniques sur le dispositif d'enquêtes, des manuels d'utilisation des données ou des volumes consacrés à la présentation des résultats. Accessible à tous, les enquêtes PISA sont continuellement soumises à la critique et au contrôle scientifique (Félouzis et Charmillot 2012).

C'est ainsi que la capacité des enquêtes PISA à relever les défis méthodologiques des comparaisons internationales a souvent été mise en avant par la communauté scientifique internationale (*cf.* notamment Duru-Bellat, Mons et Suchaut 2004b ; Olsen et Lie 2006 ; Grenet 2008 ; Baudelot et Establet 2009 ; Félouzis et Charmillot 2012), tout comme elle a été parfois décriée par des experts de l'évaluation extérieurs à l'OCDE et par des chercheurs (*cf.* notamment Prais 2003 ; Fertig 2004 ; Goldstein 2004 ; Goldstein, Bonnet et Rocher 2007). Au sein de la communauté française de l'évaluation, deux discours se sont confrontés au sujet des enquêtes PISA (Mons et Pons 2009). Au début des années 2000, le ministère de l'Éducation nationale, par la voix de sa Direction de l'évaluation, de la performance et de la prospective (DEPP), a développé ce que N. Mons et X. Pons appellent « l'argumentaire des biais ». Le but de cet argumentaire est d'expliquer les résultats des élèves français en se centrant sur les biais inhérents à la production des résultats. Ces biais prennent différentes formes : ils peuvent être culturels, politiques ou encore méthodologiques. Nous les décrirons en détail tout le long de ce document. À partir de 2005, « l'argumentaire des biais » décline au profit d'une réflexion constructive sur l'élaboration d'enquêtes comparatives et sur l'exploitation des résultats de PISA (Mons et Pons 2009). N. Mons et X. Pons en veulent pour preuve l'émission, en janvier 2005, d'un rapport favorable aux comparaisons internationales pour le Haut Conseil de l'évaluation de l'école (HCEE). Les auteurs du rapport, N. Bottani (un ancien administrateur de l'OCDE) et P. Vrignaud (un expert français en psychométrie), appellent les chercheurs et les experts à s'investir dans la conception d'enquêtes de grande envergure, malgré les réserves émises sur les biais présents dans PISA. Petit à petit, les discours de l'OCDE se propagent et éclipsent les objections au programme PISA. Sur le terrain de la méthodologie, les experts de l'OCDE, et parmi eux en première ligne R. Adams et M. Wu, cherchent à montrer la robustesse des résultats de PISA en rappelant la rigueur des procédures mises en œuvre lors de l'élaboration de l'enquête. Pour l'analyse des résultats, ils développent « l'argumentaire de la gouvernance idéale » lequel consiste à expliquer les performances des élèves en fonction de l'efficacité et l'équité relatives des systèmes éducatifs et non plus en fonction des imperfections méthodologiques des évaluations (Mons et Pons 2009).

L'analyse des différents argumentaires de défense et de critique du dispositif d'enquêtes PISA révèle la manière dont les résultats produits par PISA peuvent facilement être reformulés et réinterprétés par les différents acteurs de l'évaluation (Mons et Pons 2009). À notre sens, il est donc important que le dispositif soit critiqué du point de vue des théories scientifiques. Avant de prendre au sérieux les données PISA et leurs analyses secondaires, il convient donc de réaliser un examen du dispositif d'enquêtes et de critiques qu'on lui a opposées. Cette démarche vise à porter un regard distancié sur les données de PISA, exploitées dans de nombreuses études et recherches. Comment le dispositif d'enquêtes PISA répond-il aux défis méthodologiques des comparaisons internationales ? Quels sont les objectifs du programme PISA en matière d'évaluation internationale des compétences ? Quels choix méthodologiques ont été effectués pour assurer des comparaisons internationales de qualité ? Qu'impliquent ces choix pour l'analyse des données de PISA ?

Pour apporter des éléments de réponse à ces questions, nous nous appuyons sur les rapports de l'OCDE comme sur les publications d'observateurs extérieurs qui ont pointé les forces et les limites du programme. Nous discutons de manière systématique les objectifs et les spécificités méthodologiques de PISA et la manière dont elles répondent aux défis des comparaisons internationales. Au fil de cet examen critique ainsi qu'en conclusion, nous circonscrivons les ambitions à nourrir vis-à-vis des analyses comparées à partir des données PISA.

1. Les objectifs novateurs de PISA en matière d'évaluation internationale

Nous commençons par pointer les éléments de continuité et de rupture que présentent les évaluations PISA avec les enquêtes internationales passées. Rappelons tout d'abord que les évaluations internationales sont nées de la vieille conception du « monde comme un grand laboratoire d'éducation ». L'idée d'enquêtes comparatives à large échelle a pris forme avec la création de l'IEA (*International Association for the Evaluation of Educational Achievement*, Association internationale pour l'évaluation des résultats éducatifs) en 1961, sous l'égide de l'Institut de l'éducation de l'UNESCO (Organisation des Nations unies pour l'éducation, la science et la culture). Penser le monde comme un laboratoire invitait à élaborer des enquêtes rendant compte de la variété internationale dans les conditions et les activités d'enseignement. Depuis les années 1960, l'IEA a ainsi piloté la plupart des enquêtes internationales (au nombre d'une quinzaine) portant sur les acquis des élèves[3]. Ces évaluations ont cherché à comparer, en

[3] Pilote Twelve-Country Study (1960), FIMS (1964), FISS (1971), Six Subject Survey (1971), SIMS (1982), Classroom Environment Study (1983), SISS (1984), Written Composition Study (1985), Reading Literacy Study

termes de ressources, les établissements scolaires des pays participants. Elles ont mesuré les performances des élèves et se sont attachées à apprécier l'efficacité des pratiques pédagogiques des différents systèmes éducatifs.

À partir des années 1990, l'objectif principal des évaluations internationales s'est vu modifié (Robin et Rocher 2002 ; Olsen et Lie 2006). Les études et les données comparatives sont désormais perçues comme une source d'information pour évaluer l'efficacité de projets politiques plus globaux, prenant en compte le coût et la qualité du fonctionnement du système éducatif. L'objet des évaluations est progressivement passé des élèves, puis des écoles aux systèmes éducatifs dans leur ensemble. La réforme des services publics, souvent nommée « le nouveau management public », mise en œuvre dans de nombreux pays, a contribué au « glissement de la régulation des "inputs" – par exemple, la détermination de l'utilisation des ressources ou du nombre d'élèves par classe – vers un contrôle de l'"output" – résultats, enquêtes auprès des parents ou des élèves » (Olsen et Lie 2006, p. 14). L'enquête PISA constitue un exemple emblématique de ce tournant et c'est pour répondre à la préoccupation des liens entre les politiques éducatives et les finalités (égalité des chances d'apprendre, niveau de compétence des élèves) que l'OCDE l'a mise en place.

De fait, le programme PISA est envisagé par les décideurs politiques et les organisations internationales comme un outil de comparaison des systèmes scolaires, révélateur de leurs atouts et de leurs faiblesses. Cependant, la médiatisation et l'instrumentalisation des résultats ont également contribué à faire de PISA un vecteur de compétition internationale, dont la manifestation la plus symptomatique réside dans le classement des pays suivant leurs performances moyennes. La réification de ce classement international est pourtant regrettée par de nombreux observateurs dans la mesure où elle enfle les écarts de réussite observés entre les pays participants (Grenet 2008) et où elle livre une vision simplifiée du fonctionnement des écoles nationales (Mulford 2002 ; O'Leary 2001). Mulford (2002) s'inquiète de ce que l'utilisation des données de PISA pourrait donner lieu à des solutions politiques simplistes en matière d'éducation tandis que D. Phillips et K. Ochs (2004) soulignent que l'emprunt politique dans le domaine éducatif (*policy borrowing*) demeure un processus complexe (*cf.* encadré 1).

(1991), COMPED (1992), PPP (1995), TIMSS (1995, 1999, 2003, 2007 et 2008), Language Education Study (1995), SITES (1999, 2001 et 2006), CIVED (1999), PIRLS (2001 et 2006).

Encadré 1 : Le processus complexe de l'emprunt d'une politique éducative

Selon D. Phillips et K. Ochs, la transposition d'une politique éducative d'un pays vers un autre s'effectue en quatre étapes théoriques. La première étape réside dans l'attraction exercée sur le pays dit « emprunteur » par une politique éducative d'un autre pays. Cette attraction peut découler de phénomènes divers : l'insatisfaction interne du pays « emprunteur » (manifestée par les parents, les professeurs, les inspecteurs ou d'autres acteurs), l'effondrement de son système scolaire (suite à l'explosion de la dette nationale, suite à une guerre ou à une catastrophe naturelle), une évaluation extérieure négative de son système éducatif (comme les résultats d'une enquête internationale du type de PISA) ou encore des changements politique ou économique dans le pays… Le processus de décision constitue la deuxième étape du modèle. Les gouvernements cherchent alors à justifier leur décision d'emprunt politique en présentant des motivations d'ordre théorique, réaliste ou encore pratique. La mise en œuvre de la politique (qui constitue la troisième phase) dépend, quant à elle, du contexte du pays emprunteur. La vitesse à laquelle s'opère le changement politique dépend notamment des attitudes des acteurs dits « significatifs » – à savoir les individus et les institutions qui ont la capacité de résister ou d'accompagner la transformation politique. Enfin, la dernière phase d'internalisation de la politique est constituée de quatre processus : l'impact de la politique empruntée sur le fonctionnement du système éducatif existant, l'absorption par le système d'éducation des composantes caractéristiques de la politique étrangère, la réinterprétation de la politique adoptée comme faisant partie de la stratégie d'ensemble du système éducatif et enfin l'évaluation de la politique. Ce bref et modeste aperçu du modèle de transposition politique permet néanmoins de rendre compte du large nombre de facteurs impliqués dans le processus d'emprunt politique dans le domaine éducatif. Il dissuade par conséquent d'émettre des interprétations trop hâtives et d'exploiter de manière trop simpliste les résultats de PISA.

Comme mentionné ci-dessus, le programme PISA hérite de plusieurs décennies d'expériences dans le domaine de l'élaboration des enquêtes internationales (qu'il s'agisse de l'organisation, de la méthodologie ou des publications). Cependant, il se distingue des précédentes enquêtes internationales sur trois aspects originaux et fondamentaux, qui ont abondamment été discutés par les spécialistes de l'éducation depuis le lancement de PISA. Ces trois nouveautés tiennent à la nature des compétences que PISA cherche à évaluer, à la population visée et au suivi périodique des acquis des élèves.

1.1. Évaluer les compétences en littératie et en cultures mathématique et scientifique

PISA fonde son évaluation sur les résultats des élèves testés dans trois domaines : la compréhension de l'écrit, la culture mathématique et la culture scientifique (le terme anglais « literacy » a été traduit en français par « compréhension de l'écrit » pour la lecture, « culture » pour les mathématiques et les sciences). Pour chacun des trois domaines, des experts internationaux des pays membres de l'OCDE se sont accordés sur les définitions de cette littératie ou culture et sur un cadre conceptuel pour l'évaluer (OCDE 1999). Dans PISA, ces domaines sont respectivement définis par l'OCDE de la manière suivante :

> « Comprendre l'écrit, c'est non seulement comprendre et utiliser des textes écrits, mais aussi réfléchir à leur propos. Cette capacité devrait permettre à chacun(e) de réaliser ses objectifs, de développer ses connaissances et son potentiel, et de prendre une part active dans la société. » (OECD, 1999, p. 24)

> « La culture mathématique est l'aptitude d'un individu à identifier et à comprendre les divers rôles joués par les mathématiques dans le monde, à porter des jugements fondés à leur propos, et à s'y engager, en fonction des exigences de sa vie présente et future en tant que citoyen constructif, impliqué et réfléchi. » (OECD 1999, p. 49)

> « La culture scientifique est le fait de pouvoir utiliser des connaissances scientifiques, d'identifier les questions et de tirer des conclusions fondées sur des faits, en vue de comprendre le monde naturel et de prendre des décisions à son propos, ainsi que de comprendre les changements qui y sont apportés par l'activité humaine. » (OECD, 1999, p. 68)

Ainsi, la « literacy » est définie comme la capacité des élèves à utiliser leurs connaissances dans des situations de la vie quotidienne, et à analyser, raisonner et communiquer de manière efficace. Le concept de littératie ou de culture est en fait bien plus large que la notion traditionnelle d'aptitude à lire et à écrire. Il ne repose pas en tant que tel sur les programmes scolaires des divers pays participants. Dans PISA, il s'agit des connaissances et savoir-faire dont les élèves ont besoin dans la vie d'adulte.

Jugée équivoque, la définition de la littératie a fait couler beaucoup d'encre au cours des années 2000 (*cf.* notamment Prai, 2003 ; Adams 2003 ; Goldstein 2004 ; Prais 2004 ; Bodin 2005 ; Prenzel et Zimmer 2006 ; Rémond 2006 ; Bulle 2010). D'elle dépend l'interprétation des scores de réussite à PISA et a fortiori du diagnostic formulé sur le système éducatif. En effet, les conclusions à tirer ne seront pas les mêmes si les résultats des élèves reflètent leurs connaissances académiques ou leurs connaissances utiles à la vie de tous les jours. D'un côté, les experts de l'OCDE et du consortium PISA rapportent que le programme PISA cherche à évaluer les produits attendus de l'école (Adams 2003 ; OECD 2002) et non des points spécifiques des programmes scolaires internationaux. De l'autre, les commentateurs discutent de l'ambigüité de cette définition et cherchent à savoir si les épreuves PISA parviennent en effet à tester des compétences utiles pour la vie future ou si elles évaluent des compétences scolaires classiques. D'aucuns ont pointé la proximité de certains items à des situations de la vie quotidienne et d'autres items aux programmes scolaires nationaux (Prais 2003 ; Bodin 2005 ; Prenzel et Zimmer 2006 ; Rémond 2006). S. Prais (2003) donne l'exemple d'items de culture mathématique dont les contextes se rapportent pour certains à la vie quotidienne, pour d'autres à des exercices classiques de mathématiques. A. Bodin (2005) a classé de manière systématique tous les items de mathématiques de PISA 2003 et est parvenu aux conclusions suivantes : les épreuves de mathématiques de PISA ne couvrent que 15 % du programme scolaire français, tandis que les épreuves du Brevet d'études du Premier cycle couvrent quant à elles 35 % du programme et sont moins exigeantes sur le plan des compétences cognitives. En outre, 75 % des questions traitées par PISA correspondent à des enseignements reçus par les élèves au collège. Cela signifie aussi que 25 % des questions ne correspondent pas à ce que les élèves apprennent au collège. Les analyses effectuées sur un échantillon supplémentaire d'élèves allemands en 2003 ont montré que les évaluations de PISA sont en ligne avec la tradition nationale d'enseignement. M. Prenzel et K. Zimmer (2006) affirment que les épreuves PISA répondent globalement aux exigences fondamentales des programmes scolaires allemands. Selon eux, des tests nationaux avec des types d'exercices familiers aux élèves allemands auraient conduit à des résultats similaires à ceux de PISA. La thèse d'une référence

implicite des épreuves PISA aux programmes scolaires anglo-saxons est par ailleurs soutenue par de nombreux observateurs (*cf.* notamment Romainville 2002 ; Rémond 2006). La compréhension de l'écrit s'articule autour de trois dimensions cognitives. Or, l'une d'entre elles – la capacité des élèves à réfléchir de manière critique – est inscrite dans les programmes d'enseignements de tous les pays de langue anglaise (*Reading Curriculum and Standards*) tandis qu'elle demeure la grande absente des programmes français. Au-delà de son inscription dans les programmes, la capacité réflexive est régulièrement évaluée au cours de la scolarité des élèves britanniques. Certains pays valorisent ainsi les activités réflexives dans leur enseignement scolaire, et obtiennent manifestement des scores de performances moyens supérieurs dans ce domaine relativement aux autres domaines de la compréhension de l'écrit. Une manière de concilier ces différents points de vue est de désigner les capacités évaluées par PISA par le concept de « potentiel académique » (Bulle 2010). En effet, PISA mesure les compétences à résoudre des problèmes de type académique indépendamment des savoirs disciplinaires, tandis que ces compétences entretiennent des liens étroits avec les capacités à réussir des études.

Finalement, l'orientation des évaluations PISA ne nous apparaît pas déconnectée des programmes éducatifs mais plutôt cohérente avec les finalités et les objectifs des enseignements scolaires, qui, au-delà de la volonté que les élèves acquièrent des compétences précises, se préoccupent de ce que les élèves peuvent accomplir grâce à leur apprentissage (OECD 1999). Toutes choses considérées, la palette d'items testés auprès des élèves semble entrer en résonnance avec le concept de « literacy » que PISA souhaite évaluer. Il est possible que les définitions de la littératie et de la culture soient plus proches des standards scolaires anglo-saxons que de toute autre aire culturelle ou régionale du monde. Nous ne pensons pas que cette prévalence du modèle anglo-saxon soit en état d'affaiblir les données de PISA. L'important est d'avoir cette donnée en mémoire lors de l'analyse des résultats.

1.2. Cibler les élèves de 15 ans

Le deuxième principe original des enquêtes PISA réside dans le choix de la population visée. Contrairement aux évaluations internationales traditionnelles, la population interrogée dans PISA est définie par rapport à un critère d'âge et non par rapport à un critère de classe (au sens de degré ou niveau scolaire). Précisément, PISA choisit d'évaluer des élèves âgés de 15 ans, c'est-à-dire en fin de scolarité obligatoire. Cette décision comporte deux implications.

Premièrement, le choix de ce groupe d'âge entraîne une comparaison singulièrement différente d'une comparaison des acquis des élèves d'un degré scolaire donné. Pour des pays pratiquant le passage automatique dans la classe supérieure, que l'enquête cible les élèves d'un niveau scolaire donné ou d'un âge donné ne modifie pas beaucoup la nature de la population étudiée puisque presque tous les élèves d'un âge donné sont scolarisés dans un même niveau d'enseignement. En revanche, dans les pays pratiquant le redoublement comme la France, un même niveau d'enseignement rassemble des élèves

d'âges différents et les élèves d'un âge donné sont répartis dans des degrés différents. En juin 2003, S. Prais publie dans *Oxford Review of Education* un ensemble de critiques méthodologiques adressées à PISA, dont l'une porte sur ce choix. S. Prais juge injuste le critère d'âge puisqu'il revient, selon lui, à ignorer la motivation pédagogique des pays qui regroupent les élèves dans des classes suivant leur maturité intellectuelle, scolaire, émotive et physique. Publier les scores des élèves d'une tranche d'âge donnée serait une injustice faite au processus de scolarisation mis en œuvre dans ces pays.

Ce point de vue est loin d'être partagé par tous. R. Adams, au nom du consortium PISA, répond en septembre 2003 dans la même revue que l'on peut de manière équivalente affirmer que les pays pratiquant le redoublement sont eux aussi favorisés par une évaluation portant sur un degré scolaire puisque leurs élèves redoublants bénéficient d'années d'instruction supplémentaires. À l'inverse de certains chercheurs (Paris 2003 ; Goldstein, Bonnet et Rocher 2007), nous considérons que le critère d'âge ne rend pas plus ardue la comparaison internationale qu'un autre critère. Il apparaît au contraire très naturel puisqu'il consiste à retenir un échantillon représentatif d'une cohorte de naissance. Ce qui est crucial, dans tous les cas – qu'il s'agisse d'une évaluation des élèves d'un niveau scolaire donné ou d'un âge donné, c'est d'avoir connaissance de l'organisation des systèmes éducatifs considérés. L'important est que la référence à l'âge plutôt qu'au degré scolaire, qu'elle soit jugée plus naturelle ou non, plus compliquée ou non, soit mobilisée pour l'interprétation rigoureuse des résultats.

La seconde implication du choix de cette population est que les résultats portent sur les compétences d'élèves atteignant la fin de scolarité obligatoire, et ainsi susceptibles d'entrer dans la vie active – l'âge de fin de scolarité obligatoire dans la plupart des pays de l'OCDE étant fixé à 15 ou 16 ans[4]. Cette spécificité constitue un aspect intéressant, assez peu discuté. Toujours dans le même article de 2003, S. Prais affirme que le choix de l'âge de 15 ans peut fragiliser les résultats de l'enquête. Les épreuves cherchent à évaluer si les élèves de 15 ans sont prêts à participer à la société de la connaissance, alors que beaucoup d'entre eux sont encore en cours de formation. En effet, dans certains pays (comme l'Allemagne, l'Autriche, la Suisse, et d'un certain point de vue la France et les Pays-Bas), l'enseignement des mathématiques et de la langue du pays demeure obligatoire jusqu'à l'âge de 18 ans. À l'inverse, en Grande-Bretagne, ces enseignements n'ont plus de caractère obligatoire passé l'âge de 16 ans. Par conséquent, les années finales de l'école obligatoire des pays continentaux évoqués ci-dessus sont généralement consacrées à l'approfondissement des connaissances mathématiques théoriques, réservant les applications pratiques aux années d'études suivantes, tandis qu'en Grande-Bretagne, ces années de fin de scolarité obligatoire sont davantage tournées vers les applications. Pour résumer la

[4] En réalité, l'instruction obligatoire s'achève entre les âges de 14 et 18 ans suivant les pays de l'OCDE (OECD 2000).

position de S. Prais, les résultats des élèves de 15 ans aux épreuves PISA dépendent de la manière dont les systèmes scolaires nationaux conçoivent la transmission des savoirs avant et après l'âge de fin de scolarité obligatoire. Il nous semble qu'une fois encore, son argument ne parvient pas à sa démonstration et qu'il pourrait, inversement, jouer en faveur du choix retenu. Ce qui fait l'intérêt du critère d'âge est justement de pouvoir étudier les variations dans les compétences des élèves liées aux différences organisationnelles des écoles avant la fin de la scolarité obligatoire. Le choix de l'âge de 15 ans permet de dresser le bilan des acquis des élèves avant leur entrée potentielle dans la vie active, quelles que soient les pratiques pédagogiques des pays.

Finalement, l'interrogation des élèves âgés de 15 ans constitue un choix original, qui offre de compléter les résultats des évaluations internationales des élèves de degrés scolaires donnés. Ces discussions montrent que le champ de l'enquête répond à – et a fortiori impose – un questionnement particulier. Les enquêtes PISA ne permettent pas de répondre aux interrogations habituelles des évaluations : il ne s'agit pas d'évaluer les acquis des élèves d'un même niveau, exposés au même programme scolaire. Il s'agit d'évaluer les compétences des élèves d'un certain âge, ayant des trajectoires scolaires variées et se trouvant dans différents niveaux d'éducation.

Ces deux premières innovations de PISA – l'évaluation des compétences « non scolaires » des jeunes d'une même classe d'âge –, semblent plutôt réduire la dépendance des tests aux spécificités scolaires nationales et par conséquent assurer la qualité des comparaisons entre pays. Mais, en contrepartie, elles imposent une conception fortement universaliste de la comparaison internationale en éducation. Notamment, elles érigent l'efficacité et l'équité comme les deux objectifs communs et prioritaires à tous les systèmes éducatifs (Félouzis et Charmillot 2012). Ensuite, les discussions au sujet du dispositif pointent le danger qu'il existe à prendre comme tels les résultats des enquêtes internationales. Comme le rappelle M. O'Leary (2001), il est tentant de ne s'intéresser qu'au classement international puisqu'il résume de manière simple et directe les différences entre les pays. Mais il faut avoir conscience que ce classement n'a qu'une signification spécifique et limitée. Pour étudier les résultats des élèves, il faut ainsi connaître précisément les contextes nationaux dans lesquels s'inscrit l'évaluation.

Cette dernière remarque appelle un commentaire quant au choix de l'échelle spatiale à privilégier pour mener des comparaisons internationales. La comparaison d'un vaste ensemble de pays confère aux résultats établis un certain degré de généralité. Effectuer des comparaisons à grande échelle permet de mettre au jour des lois universelles. En contrepartie, on laisse nécessairement échapper des informations disponibles sur chacun des pays étudiés. Aussi, il apparaît également judicieux d'utiliser des échelles spatiales réduites à un petit ensemble de pays. Dans l'introduction de l'ouvrage qu'il a dirigé avec L. Lesnard, A. Chenu recommande de mener des comparaisons sur un nombre de pays limité « parce qu'elles autorisent un réglage optimal entre proximité et distance, entre familiarité avec les idiosyncrasies

des cultures nationales et maîtrise bien fondée de principes théoriques d'analyse du monde social dont la validité ne pourra être attestée que grâce au jeu sur les contrastes entre les différents terrains étudiés » (p. 11). Une telle démarche nous paraît en effet raisonnable et adaptée à des recherches mobilisant de manière précise les éléments du contexte national pour interpréter les résultats de PISA.

1.3. Suivre les acquis des élèves tous les trois ans

Outre la nature des compétences évaluées et la définition de la population testée, les enquêtes PISA s'illustrent par leur périodicité (Mulford 2002 ; Grenet 2008). Réalisée tous les trois ans, l'enquête permet d'effectuer un suivi régulier des performances des élèves de 15 ans de l'ensemble des pays de l'OCDE et des pays partenaires. L'objectif est d'assurer non seulement des comparaisons des systèmes éducatifs dans l'espace mais également dans le temps. Le cycle d'évaluation est conçu de telle manière que chaque enquête met l'accent sur l'une des trois compétences évoquées précédemment, en y consacrant environ les deux tiers des questions. En effet, la compréhension de l'écrit fut le domaine principal testé en 2000 et en 2009, mais la culture mathématique et la culture scientifique ont également été évaluées à chaque enquête. Chacun des trois domaines évalués comporte des sous-domaines. Par exemple, la compréhension de l'écrit recouvre trois dimensions cognitives. La première, « retrouver l'information », renvoie à la capacité des élèves à localiser des informations dans un texte ; la deuxième, « développer une interprétation », à leur aptitude à dégager du sens et à établir des inférences à partir de l'écrit ; la troisième, « réfléchir sur le contenu du texte », à leur faculté de relier le texte à leurs connaissances, leurs idées et leurs expériences. Pour étudier l'évolution des scores de réussite des élèves dans un domaine, et davantage encore dans un sous-domaine, la comparaison est mieux assurée entre deux dates où le domaine est majeur. Ainsi, sur les 240 items du test de lecture de PISA 2009, 28 ont été administrés à chacun des cycles d'évaluation et 24 autres items ont été repris uniquement du test de PISA 2000 (OECD 2010, p. 37). Les tests de lecture de PISA 2000 et de PISA 2009 ont donc plus de 20 % d'items communs. Les analyses statistiques de l'évolution des performances des élèves entre 2000 et 2009 sont ainsi plus précises quand elles concernent la compréhension de l'écrit que les autres domaines.

Ainsi que le regrettent de nombreux observateurs (*cf.* notamment Fertig 2004 ; Goldstein 2004 ; Goldstein, Bonnet et Rocher 2007), le programme PISA est dépourvu de perspective longitudinale puisqu'il ne suit pas les mêmes élèves au cours de leur scolarité mais évalue tous les trois ans des élèves différents, toujours âgés de 15 ans. H. Goldstein (2004) rappelle que pour comparer les effets des systèmes éducatifs, des données longitudinales sont souhaitables (quoique pas toujours suffisantes). En effet, l'exploitation de données de panel, qui suivent les mêmes individus dans le temps, permet d'aller au-delà d'une analyse en termes de corrélation et de traiter de lien causalité que l'on peut qualifier de « retardée » (Safi 2011). Il est indispensable de disposer de plusieurs mesures de performance des mêmes élèves étalées dans le temps pour conclure sur les effets d'une pratique ou d'une politique éducative et s'assurer que l'évolution

observée est due au « vrai changement » et non à l'évolution de facteurs inobservés, ou même à une erreur de mesure. H. Goldstein (2004) s'offusque de ce que certains rapports de l'OCDE et de nombreux commentaires laissent entendre qu'une comparaison directe des systèmes éducatifs est rendue possible par PISA. Pourtant, comme les enquêtes PISA offrent des données en coupe instantanée, elles ne permettent pas d'analyser les effets causaux de politiques éducatives en tant que tels. Les différences observées entre les pays répercutent indéniablement les écarts entre les systèmes éducatifs mais également les différences culturelles, sociales ou d'autres ordres entre les pays qu'il est difficile d'isoler parfaitement. De fait, les enquêtes PISA permettent seulement de dégager des tendances à partir des performances, mesurées à différents moments du temps, des élèves rendus en fin de scolarité obligatoire.

Les enquêtes PISA et leurs nouveaux objectifs constituent ainsi une forme de tournant dans l'histoire des évaluations internationales : s'intéressant aux résultats globaux des systèmes éducatifs, ces enquêtes ont choisi d'évaluer, à intervalle de temps répété, les élèves d'une classe d'âge (et non d'un niveau scolaire) parvenus en fin de scolarité obligatoire. Comment ces nouveaux principes ont été opérationnalisés lors de la conception du dispositif d'enquêtes ? Notre but, dans la suite, est d'examiner les choix méthodologiques et techniques qui ont été mis en œuvre pour assurer la comparabilité des données entre les pays. L'exposition qui suit repose en bonne partie sur l'analyse critique de la documentation fournie par l'OCDE (OECD 1999 et 2005 ; Adams et Wu 2002). Les spécificités les plus techniques du dispositif sont développées dans des encadrés méthodologiques.

2. PISA face aux défis méthodologiques de la comparaison internationale

2.1. Le pilotage international du programme PISA

L'élaboration et l'implémentation de PISA incombent à un consortium international dirigé par le Conseil Australien pour la recherche en éducation (*Australian Council for Educational Research*, ACER), en partenariat avec divers instituts de recherche nationaux. Le Consortium met en œuvre le programme PISA au sein d'un cadre établi par le comité des pays participants (*Board of Participating Countries*, BPC). Le BPC définit les priorités politiques et les standards pour développer les indicateurs, les instruments d'évaluation et les publications des résultats. Le secrétariat de l'OCDE a la responsabilité complète du programme PISA et encourage le consensus entre le consortium international et le BPC. Des experts des pays participants collaborent au sein de groupes de travail afin de doter le programme PISA d'une validité technique internationale, dans les trois domaines de compétence évalués. Ces experts cherchent à s'assurer que les instruments sont efficaces à l'échelle internationale et qu'ils prennent en compte les contextes culturels et éducatifs des différents pays de l'OCDE. La composition à dominante anglophone du vaste comité PISA est régulièrement commentée par les observateurs : celle-ci ne serait pas sans

conséquence sur l'orientation des enquêtes (Romainville 2002). Nous reviendrons sur ses possibles implications dans la suite.

Pour assurer une forte comparabilité entre les pays, de nombreuses procédures d'assurance-qualité ont été prévues à chaque étape de la conception du dispositif. Ainsi, dans chaque pays participant, le responsable national du projet (*National Project Manager*) veille à ce que PISA soit mis en œuvre dans le respect des procédures techniques et administratives définies et contribue à la vérification des résultats et des rapports. La conduite du programme PISA est en effet soumise à un cahier des charges contraignant. La collecte et le traitement des données doivent répondre à une liste de standards très exigeants concernant entre autres le tirage de l'échantillon, l'élaboration, la traduction et la correction des items, les conditions de passation des tests ou encore la gestion des données. Un des standards les plus importants est notamment le taux de réponse minimal exigé. Le contrôle-qualité est présent à deux niveaux du projet PISA : aux niveaux du centre national et des établissements. Il vise à détecter des erreurs dans la passation des épreuves ou la collecte des données, voire des fraudes. Des sanctions sont prévues dans le cas du non-respect des standards.

2.2. Constituer des échantillons nationaux représentatifs

Une condition nécessaire pour assurer la qualité des comparaisons internationales est de constituer des échantillons représentatifs au niveau national. Le consortium PISA a mis un grand soin dans cette étape. La population ciblée par PISA dans chaque pays est l'ensemble des élèves de 15 ans inscrits dans un établissement scolaire localisé dans le pays. La définition opératoire de l'âge de la population dépend directement des dates de passation du test. Ainsi, selon l'exigence internationale, les épreuves de PISA 2000 devaient être passées dans une période de 42 jours entre le 1er mars et le 31 octobre 2000. La population ciblée correspondait à l'ensemble des élèves âgés de 15 ans et trois mois révolus à 16 ans et deux mois révolus au début de la période d'évaluation (pour un pays où l'évaluation avait lieu en avril 2000, il s'agissait en fait de tous les élèves nés en 1984). Cette fenêtre d'âge a ensuite été conservée pour les évaluations des vagues suivantes.

Le plan de sondage utilisé pour PISA a été conçu de sorte à constituer des échantillons contenant au moins 4 500 élèves par pays. La méthode d'échantillonnage est commune à l'ensemble des pays : il s'agit d'un plan de sondage stratifié à deux niveaux (trois pays ont un échantillon stratifié à trois niveaux). Le premier niveau comporte tous les établissements scolarisant des élèves de 15 ans. Le second niveau comprend tous les élèves âgés de 15 ans scolarisés dans le pays. En amont de l'échantillonnage, tous les établissements éligibles dans le pays ont été répertoriés. Ensuite, au moins 150 établissements parmi eux ont été sélectionnés, avec une probabilité proportionnelle à la taille de l'établissement (mesurée par une estimation du nombre d'élèves de 15 ans scolarisés dans l'établissement). Dans chaque établissement sélectionné, la liste exhaustive des élèves de 15 ans éligibles a été établie. Si la liste contient plus de 35

élèves, 35 élèves sont sélectionnés avec une probabilité égale. Si la liste contient 35 élèves ou moins, tous les élèves sont sélectionnés. Chaque établissement doit finalement compter plus de 20 élèves, afin d'assurer une précision convenable pour l'estimation des composantes de la variance entre et au sein des établissements.

Des standards de qualité ont du être respectés en ce qui concerne la couverture de la population ciblée et les taux de réponse des établissements et des élèves. Concernant la couverture de la population ciblée, le programme PISA a cherché à circonscrire les raisons possibles d'exclusion des établissements et des élèves, afin de limiter le taux d'exclusion. En théorie, un établissement peut être exclu du champ de l'enquête s'il est localisé dans une région reculée difficilement accessible, si la langue d'enseignement de l'établissement diffère de la langue nationale ou s'il s'agit d'une école spécialisée. Dans le cas de la France, par exemple, les territoires d'outre-mer, la Réunion, les établissements régionaux d'enseignement adapté (EREA) et les établissements privés hors contrat ont été exclus de l'échantillon national. Les élèves peuvent être d'office exclus en raison de déficiences intellectuelles ou fonctionnelles ou s'ils ont reçu moins d'un an d'instruction dans la langue du pays. Dans la pratique, des applications variables des standards relatifs aux exclusions ont pu être observées. S. Prais (2003) remarque que les rapports nationaux britanniques et allemands diffèrent dans leur traitement des écoles spécialisées : tandis que les Britanniques reportent leur exclusion, les Allemands indiquent leur inclusion dans l'échantillon national. Selon le standard établi, le taux d'exclusion ne doit cependant pas dépasser 5 % des élèves sélectionnés, ce qui permet de limiter les biais de sélection.

Des standards définissent également les taux de participation minimaux exigés. Un taux de réponse de 85 % est requis pour les établissements visés. Si le taux de réponse initial est compris entre 65 et 85 %, il est encore possible d'atteindre un taux de participation acceptable en recourant à un échantillon d'établissements de remplacement. Un taux de réponse de 80 % des élèves sélectionnés dans les établissements participants est également demandé. En France, le taux de participation des établissements est généralement très élevé : en 2000, le taux final de participation des élèves était de 91,2 %, supérieur au seuil requis (Adams et Wu 2002, chapitre 12). Avec des taux de réponse inférieurs aux standards exigés, les Pays-Bas en 2000 et le Royaume-Uni en 2003 ont donc été exclus des évaluations et leurs résultats ne sont pas publiés. S. Prais (2003) s'inquiète légitimement des biais induits par les faibles taux de réponse des établissements et des élèves britanniques à l'édition de 2000 (le taux de réponse final n'atteignant pas 50 %). Il présage que, dans l'ensemble, comme les établissements et les élèves non répondants correspondent à des populations peu compétentes, les résultats du Royaume-Uni sont probablement biaisés vers le haut. R. Adams (2003) de répondre que les établissements qui ne participent pas aux enquêtes ne sont pas nécessairement ceux dont les scores au GCSE (*General Certificate of Secondary Education*, l'examen de fin du premier cycle du secondaire en Angleterre) sont les plus faibles et où les

élèves sont le plus souvent d'origine sociale défavorisée, et qu'on compterait au contraire, parmi les élèves qui ne participent pas au test PISA, des élèves compétents qui préparent le GCSE.

De manière générale, on peut penser que le respect des standards imposés par l'OCDE assure une représentativité honorable des échantillons nationaux. Ainsi que le rappelle D. Joye en introduction d'un chapitre intitulé « Échantillonnage et pondération » dans l'ouvrage consacré à *La France dans les comparaisons internationales* (Chenu et Lesnard 2011), « la qualité d'une enquête dépend du processus de sélection des répondants. Et si ce moment ne se réalise pas sans défauts, ce qui est finalement le cas le plus fréquent, il est essentiel de le documenter pour qu'une éventuelle correction soit possible ou, à tout le moins, que les limites de l'instrument soient connues » (Joye 2011, p. 143). Les échantillons ainsi constitués permettent d'inférer des résultats pour l'ensemble de la population nationale des élèves de 15 ans. Il est néanmoins nécessaire de tenir compte des erreurs d'échantillonnage pour conclure à la significativité des résultats (l'encadré 2 détaille notamment le calcul de la précision des statistiques élaborées sur les données de PISA). La connaissance de ces erreurs est en particulier précieuse pour l'interprétation des différences observées entre les pays et des tendances temporelles observées à partir des données.

Pour les utilisateurs des données PISA, deux choses sont particulièrement appréciables : d'une part que les rapports de l'OCDE signalent bien les pays dont la qualité des données est jugée douteuse ; d'autre part que les données mises à disposition permettent de tenir compte de l'incertitude liée au tirage de l'échantillon. Ajoutons enfin que le plan de sondage n'est pas neutre pour le choix des méthodes statistiques à mettre en œuvre. En effet, la structure en deux niveaux des données du PISA suggère fortement l'emploi d'une modélisation multi-niveaux de la performance.

Encadré 2 : Plan de sondage, pondération des données et calcul de la précision

Pour assurer la qualité des comparaisons internationales, le consortium de PISA a cherché à constituer des échantillons d'élèves représentatifs au niveau national. Il existe deux moyens d'action pour y parvenir : sélectionner soigneusement les individus de l'échantillon de manière à ce qu'il soit directement représentatif et pondérer les observations recueillies de manière à les rendre représentatives. Ces deux techniques ont été utilisées dans le cas de PISA. S'agissant du tirage de l'échantillon, les établissements ont été sélectionnés avec des probabilités proportionnelles à leur taille. Les plus grands ont alors une probabilité de sélection plus élevée que les petits établissements, mais les élèves de ces derniers ont une probabilité de sélection au sein de leur établissement qui est plus forte que celles des élèves scolarisés dans de grands établissements. Ensuite, un échantillon aléatoire d'élèves (en général 35) est tiré au sein de chacun des établissements sélectionnés.

Ainsi, la probabilité qu'un établissement j soit sélectionné est égale au ratio de la taille de l'établissement Nj multipliée par le nombre total d'établissements échantillonnés n_{sc}, divisé par le nombre total d'élèves *N* dans la population (*cf.* OECD 2005, p. 25):

$$p_j = \frac{N_j . n_{sc}}{N}$$

La probabilité finale qu'un élève *i* de l'établissement *j* soit sélectionné est égale au produit de la probabilité de sélection de l'établissement *j* et de la probabilité de sélection de l'élève *i* au sein de l'établissement *j,* égale au ratio du nombre d'élèves sélectionnés dans l'établissement divisé par la taille de l'établissement (*cf.* OECD 2005, p. 26) :

$$p_{ij} = \frac{n_i}{N_j} \frac{N_j . n_{sc}}{N} = \frac{n_i . n_{sc}}{N}$$

Le plan de sondage garantit alors que tous les élèves ont la même probabilité de sélection et par conséquent le même poids. Pourtant les échantillons n'ont pas été jugés parfaitement représentatifs pour trois raisons : la surreprésentation ou la sous-représentation de certaines strates de la population, les imprécisions concernant le nombre attendu d'élèves de 15 ans scolarisés dans les établissements concernés par l'enquête, la non-réponse aux niveaux élève et établissement. Les données de PISA ont par conséquent été repondérées. Il a été attribué un poids final à chaque élève de la base de données.

Les résultats observés sur les échantillons d'élèves étudiés doivent permettre de déduire des conclusions plus ou moins probables pour l'ensemble de la population d'élèves ciblée. Chaque généralisation ou inférence proposée à partir de l'échantillon, c'est-à-dire chaque estimation d'un paramètre de la population, est de fait associée à une incertitude liée au tirage de l'échantillon dont on souhaiterait rendre compte. Pour calculer cette erreur d'échantillonnage, la méthode retenue par PISA consiste à utiliser plusieurs poids possibles de l'échantillon considéré, que l'on nomme « poids répliqués ». Plusieurs méthodes de réplication d'échantillon à deux degrés existent. Les concepteurs de PISA ont recouru à une variante de la méthode de réplication répétée (Balanced Repeated Replication dite BRR), qui utilise la « modification de Fay » (OECD 2005, p. 50). Cette procédure sélectionne aléatoirement un établissement parmi chaque pseudo-strate de l'échantillon (la pseudo-strate étant définie par rapport à une caractéristique particulière de l'établissement, par exemple la zone – rurale ou urbaine) et fixe son poids égal à $k = 0{,}5$. Les poids des établissements non sélectionnés par la procédure sont alors fixés à $(2 - k) = 1{,}5$. Ainsi les poids des établissements d'un échantillon répliqué sont soit multipliés par 0,5 soit par 1,5. Comme cette méthode donne lieu à un ensemble très large de réplications possibles, un ensemble réduit d'échantillons répliqués est généré afin d'éviter des calculs trop longs. La base de données de PISA comporte ainsi un poids dit « final » (à utiliser de manière prioritaire dans les analyses) et 80 poids répliqués (lesquels permettent de rendre compte de l'erreur d'échantillonnage).

Le calcul d'une statistique d'intérêt $\hat{\theta}$ s'effectue à partir de l'échantillon initial (à l'aide du poids final) et sur les 80 échantillons répliqués (à l'aide des 80 poids répliqués). Pour obtenir l'erreur (ou la variance) d'échantillonnage, les estimateurs obtenus sur les échantillons répliqués sont comparés à l'estimateur de l'échantillon initial (OECD 2005, p. 51):

$$\sigma^2_{(\hat{\theta})} = \frac{1}{80(1-k)^2}\sum_{i=1}^{80}(\hat{\theta}_i - \hat{\theta})^2 = \frac{1}{20}\sum_{i=1}^{80}(\hat{\theta}_i - \hat{\theta})^2$$

Précisions que des macros SAS effectuant le calcul des erreurs d'échantillonnage ont été mises à la disposition des chercheurs sur le site Internet de PISA.

2.3. Les efforts d'harmonisation des évaluations

L'objectif d'homogénéisation des données PISA a non seulement pesé sur la méthode de sondage, mais également sur la conception, la passation et la correction des évaluations. Nous verrons qu'à chacune de ces étapes de nombreuses procédures d'assurance-qualité ont encore été mises en œuvre.

Il s'est tout d'abord agi, pour les experts impliqués dans l'harmonisation des évaluations, de constituer un matériel de test de difficulté équivalente entre les pays. Des évaluations ont été élaborées pour les trois domaines testés. En 2000, la réussite en compréhension de l'écrit est évaluée à partir de 141 items d'une durée approximative de 270 minutes. Les évaluations de cultures mathématique et scientifique consistent respectivement en 32 items et 35 items d'une durée de 60 minutes chacune. Le matériel utilisé dans l'épreuve générale est issu d'un ensemble de près de 600 items expérimentés sur des échantillons réduits (environ 1 200 élèves) dans tous les pays, l'année qui a précédé la véritable enquête. Cet ensemble d'items incluait 16 items repris de IALS (*International Adult Literacy Survey*) et 3 items de TIMSS (*Trends in International Mathematics and Science Study*). Finalement, seuls les items de IALS ont été effectivement inclus dans le test, dans le but d'établir des liens entre les résultats de PISA et de IALS.

À ce stade, nous pouvons déjà donner voix à deux critiques portant sur l'élaboration et le choix des items. La première concerne la reprise des items de lecture de IALS. G. Bonnet (2002) regrette l'inclusion d'unités de lecture des épreuves IALS, alors que ces dernières ont été conçues pour une population d'adultes âgés de 16 à 65 ans dans le contexte d'une enquête auprès des ménages et que la fiabilité des items de IALS a été fortement décriée (*cf.* notamment les critiques de F. Guérin-Pace et d'A. Blum 1999). Quelle que soit la qualité des items de IALS, les conséquences de leur inclusion dans PISA apparaissent somme toute limitées lorsque l'on rappelle qu'ils ne représentent que 16 des 141 items administrés.

La seconde critique porte sur l'origine des items. Bien que des procédures internationales de discussion et d'échange aient été mises en place, l'élaboration des épreuves a incombé à un consortium à

dominante anglophone dirigé par l'ACER, un organisme australien. M. Romainville (2002) souligne que même si les tests PISA ont été en partie conçus sur la base des questions proposées par les pays participants, il n'en demeure pas moins que l'essentiel des épreuves (78 % des items selon M. Romainville (2002) et deux tiers des supports testés selon T. Rocher (2003)) provient des pays de langue anglaise. Un nombre important d'exercices sont inspirés des manuels scolaires des pays anglo-saxons (notamment néozélandais et australiens). L'origine massivement anglo-saxonne des items génère une proximité culturelle entre les épreuves PISA et les élèves de ces pays qui tend à favoriser ces derniers. M. Romainville souligne que « plus fondamentalement, c'est tout un modèle d'école que les tests charrient » (2002, p. 89). Ce jugement rejoint celui de M. Rémond (2006) qui souligne également la dominante anglo-saxonne des évaluations.

À notre sens, ces arguments sont loin d'invalider les résultats des évaluations internationales. Que les épreuves PISA reflètent davantage les attendus des systèmes scolaires anglo-saxons n'est pas à même de révoquer les analyses. L'essentiel est d'avoir à l'esprit la prévalence anglo-saxonne lors de l'analyse des résultats. Elle apporte même un éclairage intéressant sur les compétences comparées des élèves, que des épreuves nationales n'auraient pas décelées (Rocher 2003). Dans son étude des items biaisés en faveur des élèves français et américains, T. Rocher (2003) constate l'influence des documents de support et du mode de questionnement sur les résultats des élèves. Aux États-Unis, les élèves ont l'habitude de travailler sur des supports de la vie professionnelle, aussi réussissent-ils mieux que les élèves français les questions de PISA 2000 portant sur un avis invitant les employés d'une entreprise à se faire vacciner contre la grippe. En revanche, les items portant sur un extrait d'une pièce d'Anouilh sont mieux réussis par les élèves français. Le mode de questionnement des élèves (questions ouvertes appelant une réponse courte ou longue, construite ou non, questions à choix à multiples) peut également constituer une source de biais. Toujours dans sa comparaison France/États-Unis, T. Rocher conclut que les élèves américains sont plus à l'aise pour répondre aux questions à choix multiples et rédiger une réponse longue et construite tandis que les élèves français ont plus de facilités à répondre à des questions ouvertes appelant des réponses courtes et précises. En fin de compte, la proximité culturelle des élèves à tel ou tel type de document ou à tel ou tel mode de questionnement reflète des différences dans les pratiques et les objectifs des deux systèmes éducatifs. T. Rocher (2003) affirme donc l'intérêt d'étudier les fonctionnements différentiels des items suivant les pays plutôt que de considérer ces biais comme des erreurs de mesure. Si la fiabilité du classement international est fragilisée par l'orientation anglo-saxonne des épreuves, les résultats à ces tests n'en sont pas moins éclairants sur les compétences et les connaissances comparées des populations d'élèves.

Les concepteurs de PISA ont ensuite pris grand soin de l'étape de traduction et d'adaptation des items aux contextes culturels nationaux. En effet, les erreurs de traduction sont connues pour être la première cause du mauvais fonctionnement des tests internationaux. L'objectif principal est naturellement

d'assurer un matériel de test et un niveau de difficulté équivalents dans tous les pays. G. Bonnet (2002) rappelle qu'un tel but n'est jamais atteignable et que l'erreur est donc de considérer qu'une fois traduits les items ont les mêmes niveaux de difficulté. Cependant, un certain nombre de procédures d'assurance-qualité ont été mises en œuvre dans les enquêtes PISA, de manière bien plus stricte que dans les évaluations internationales passées. Ces procédures comprennent la traduction parallèle des items dans deux versions sources (l'une en anglais, l'autre en français) et recommandent que chaque pays propose une double traduction suivie d'une conciliation par un tiers. De plus, un groupe de vérification internationale a été formé et entraîné pour assurer la cohérence de la version nationale finale avec les versions sources. La correction des cahiers est également soumise à de rigoureuses procédures d'assurance-qualité. Les concepteurs de l'enquête ont veillé à donner des consignes précises de correction des épreuves. Ces consignes ont également fait l'objet d'un processus de traduction. Les correcteurs ont suivi une formation afin de les familiariser avec les exigences de notation. Pour s'assurer de la cohérence des corrections, des sous-échantillons de questions ont été soumis à de quadruples corrections.

Pourtant, des fonctionnements différentiels des items subsistent suivant la langue (*cf.* Bonnet 2002 ; Robin 2002 ; Romainville 2002 ; Rémond 2006). Il arrive, par exemple, que lorsqu'un item A est plus difficile qu'un item B en France, cela soit aussi le cas dans les pays francophones mais pas dans les autres pays. L'existence de biais culturels pourrait donc remettre en cause le classement international établi par PISA. Force est de constater avec T. Rocher (2003) que « la question de la comparabilité des résultats est ainsi réduite à celle de la robustesse des palmarès ». Une fois encore, l'important est d'avoir connaissance de la subsistance de ces quelques biais culturels, malgré les efforts entrepris pour rendre les instruments d'évaluation universels. Si des biais culturels (notamment en faveur des pays anglo-saxons) existent bel et bien, on peut penser qu'ils affecteront le niveau des performances des élèves, mais pas nécessairement les associations statistiques entre le niveau de la performance et les autres variables qui intéresse le chercheur. Or, ce sont précisément les liens entre la réussite des épreuves, les caractéristiques personnelles et familiales de l'élève, les ressources des établissements que les études analysent.

2.4. Mesurer les compétences de groupes d'élèves

La constitution d'échantillons représentatifs au niveau national et l'harmonisation des évaluations sont deux démarches contribuant à la comparabilité des données. Les méthodes retenues par le consortium pour mesurer les compétences en littératie n'échappent pas non plus à cet objectif.

Le but du programme PISA est de produire des données capables de décrire des *groupes* d'élèves au sein de la population d'intérêt, et non de cerner individuellement chaque élève. Les évaluations de grande échelle comme PISA répondent ainsi à des questionnements spécifiques. Elles cherchent à apprécier le plus fidèlement possible les compétences de la totalité des élèves d'un pays. On reconnaît ici

le glissement des objectifs des évaluations internationales vers les résultats des systèmes éducatifs, décrit plus haut.

Des mesures de réussite adaptées à ces objectifs sont donc employées : le programme PISA – tout comme les récentes évaluations internationales de compétences telles que PIRLS[5] et PIAAC[6] – utilise des « valeurs plausibles » du score et non des scores totaux classiques. Ces mesures sont dérivées de modèles psychométriques sophistiqués : les modèles de réponse à l'item et notamment le modèle de Rasch[7]. Ces modèles psychométriques trouvent leur principe essentiel dans la distinction entre performance (les résultats observés de l'élève) et compétence (la variable ou le trait latent qui a produit ces résultats). On peut résumer ce principe à l'aide de la formule suivante : la performance observée pour l'élève est égale à la « vraie compétence » de l'élève plus une erreur de mesure (Vrignaud 2008). L'objectif des modèles psychométriques utilisés dans PISA est donc d'inférer la compétence de l'élève à partir de ses résultats au test. Il est ainsi crucial de ne pas perdre de vue la manière dont les compétences ont été opérationnalisées à partir des évaluations s'il on souhaite interpréter correctement les résultats des enquêtes PISA. Cette section vise à présenter succinctement ces instruments de mesure et les principales critiques qui leur ont été adressées[8]. Surtout, nous préciserons comment ces instruments répondent à l'objectif de comparaison internationale.

En pratique, le modèle de Rasch employé dans PISA consiste à ordonner, dans un premier temps, les items d'un même domaine testé selon leur difficulté. Dans un second temps, il échelonne sur ce même continuum les compétences des élèves. Dans le cas d'un score dichotomique (réussite ou échec), la difficulté d'un item correspond à la proportion de réponses incorrectes données à l'item. Une échelle relative peut ainsi être établie : la difficulté relative d'un item résulte de la comparaison de cet item aux autres items. Il convient donc de définir un point d'ancrage arbitraire (comparable au degré Celsius 0 sur l'échelle des températures) et une unité de mesure. Dans le cas du modèle de Rasch, l'unité de mesure est définie par la fonction logistique prenant en arguments la difficulté de l'item et la capacité de l'élève. Ce modèle postule que la probabilité qu'un élève *i* donne une réponse correcte à un item *j* est une fonction logistique de la compétence de l'élève (β_i) et de la difficulté de l'item (δ_j) :

$$P(Y_{ij} = 1 | \beta_i, \delta_j) = \frac{\exp(\beta_i - \delta_j)}{1 + \exp(\beta_i - \delta_j)}$$

[5] Progress in International Reading and Literacy Study.
[6] Programme for the International Assessment of Adult Competences.
[7] PISA utilise le modèle de Rasch de régression latente ("the latent regression Rasch model", Wilson et De Boeck 2004).
[8] Sur la méthodologie de la mesure de la littératie dans PISA et ses implications, *cf.* les articles de Vrignaud (2006 et 2008).

Un seul point d'ancrage est défini : dans le cadre de PISA, les difficultés des items ont été centrées en zéro, et cette valeur nulle de la capacité qui constitue le point de référence de l'échelle. Lorsque la capacité de l'élève égale la difficulté de l'item, la probabilité de bonne réponse est toujours égale à 0.5, quelle que soit la position de la capacité de l'élève et de la difficulté de l'item sur le continuum. En fait, le seul facteur qui influe sur la probabilité de réussite est la distance sur l'échelle de Rasch entre la capacité de l'élève et la difficulté de l'item[9].

Une fois que les difficultés des items sont placées sur le continuum de Rasch, les capacités des élèves peuvent alors être estimées. Il s'agit d'abord de calculer la probabilité d'apparition de l'ensemble des réponses données par un élève au test. Pour ce calcul, le modèle de Rasch suppose l'indépendance entre les items, c'est-à-dire que la probabilité de réussite à un item ne dépend pas des réponses données aux autres items. Ainsi, la probabilité de réussite à l'ensemble des items est égale au produit des probabilités de réussite à chacun des items. La procédure d'estimation est alors la suivante. Rappelons tout d'abord que les difficultés des items sont fixées et ordonnées. Considérons un élève en particulier : il a donné un ensemble de réponses au test qu'il a effectué. Pour une valeur β de la capacité de l'élève et étant donné les difficultés δ des items qu'il a passés, on calcule la probabilité que l'élève a de donner les réponses qu'il a fournies. Ce calcul est réitéré pour toutes les valeurs possibles β de la capacité de l'élève. Le modèle de Rasch retient alors la valeur de la capacité la plus probablement associée à la configuration particulière des réponses aux items fournies par l'élève. L'estimateur de la capacité de Rasch s'apparente donc à un estimateur du maximum de vraisemblance[10]. Le paramètre de compétence β_i est une estimation de la vraie valeur de la compétence de l'élève *i*.

Le modèle de Rasch place donc sur la même échelle – autrement dit sur la même variable latente (Rocher 2003) – les difficultés des items et les compétences des élèves, permettant ainsi des comparaisons entre les items et des comparaisons entre les élèves. Les élèves peu compétents et les

[9] Si la difficulté de l'item excède la capacité de l'élève d'une unité de Rasch, correspondant à un logit, alors la probabilité de réussite de l'élève à l'item sera toujours égale à 0,27. Si la capacité de l'élève excède la difficulté de l'item, alors la probabilité de réussite sera toujours égale à 0,73. On peut également interpréter le modèle de Rasch dans les termes suivants. Si 100 élèves de capacité 0 répondent à un item de difficulté 0, alors le modèle prédit 50 bonnes réponses et 50 mauvaises réponses. Si un élève de capacité 0 répond à 100 items de difficulté 0, alors le modèle prédit 50 bonnes réponses et 50 mauvaises réponses.

[10] Warm (1989) a montré que cet estimateur du maximum de vraisemblance est biaisé et a proposé de pondérer la contribution de chaque item par l'information qu'il fournit. Par exemple, un item difficile apporte peu d'information sur un élève avec une faible capacité. En revanche, il renseigne bien sur un élève avec une grande capacité. Par conséquent, pour les élèves les moins performants, les items faciles contribuent davantage à la vraisemblance que les items difficiles et inversement, pour les élèves les plus performants, les items difficiles contribuent plus que les items faciles. Les estimateurs de Warm (EW) et du maximum de vraisemblance (EMV) constituent donc deux types d'estimateurs de la capacité. Dans le PISA, les estimateurs de vraisemblance pondérés (EW) sont calculés en appliquant des poids aux EMV pour tenir compte du biais inhérent aux EMV, comme Warm le propose.

items faciles seront placés au bas de l'échelle, tandis que les élèves très compétents et les items difficiles seront placés en haut de l'échelle. Ce procédé permet de déduire aisément la probabilité qu'un élève d'un niveau de compétence donné a de réussir un item de tel ou tel niveau de difficulté. On voit ainsi apparaître une des hypothèses fondamentales du modèle de Rasch – l'unidimensionnalité de l'échelle – sur laquelle nous reviendrons ci-après.

Pour que l'enquête soit considérée valide, un nombre important d'items doivent être évalués auprès des élèves lors des tests finaux. Les items doivent être d'autant plus nombreux que la discipline évaluée est un domaine étendu recouvrant plusieurs dimensions cognitives. Cependant, il n'est pas raisonnable d'évaluer chaque élève de l'échantillon sur l'ensemble des items construits. Des tests trop longs s'accompagnent de nets changements dans les attitudes et les comportements des élèves, ce qui pourrait fortement biaiser les résultats. En effet, les élèves peuvent ressentir de la fatigue ou une baisse de motivation, ou encore bénéficier d'effets d'apprentissage au cours de l'épreuve. En outre, les chefs d'établissement refuseraient probablement que leurs élèves participent à des enquêtes chronophages. Pour pallier ce problème d'arbitrage entre taille réduite du questionnaire et large couverture de la discipline évaluée, le programme PISA fait donc passer à chaque élève un sous-ensemble d'items (OECD 2005).

Le dispositif d'évaluation de PISA repose sur la méthode dite « des cahiers tournants ». Cette méthode consiste à morceler le matériel de test — l'ensemble des items conçus pour l'évaluation — en différents blocs d'items et à les regrouper dans des cahiers. Dans un souci de clarté, prenons l'exemple du test de PISA 2003. Les items conçus pour PISA 2003 ont été regroupés en 13 blocs d'items. Puis 13 cahiers ont été constitués à partir de ces blocs d'items, de sorte que chacun des 13 cahiers contienne au total 4 blocs d'items, et que chaque bloc d'items apparaisse dans 4 cahiers différents et à une place différente dans le déroulement du test (une fois en première position, une fois en seconde, une fois en troisième et une fois en dernière position) : les blocs d'items « tournent » – d'où l'appellation de la méthode des cahiers tournants). Enfin, chaque élève passe les épreuves d'un des 13 cahiers, qui lui est attribué de manière aléatoire.

La méthode des « cahiers tournants » est satisfaisante au regard de l'objectif de comparaison internationale, puisque cette méthode permet d'évaluer précisément le niveau national des compétences. En pratique, les élèves interrogés sont soumis à des épreuves pendant deux heures, durant lesquelles ils répondent à l'un des cahiers d'exercices formés d'une sous-partie des items du test global. Afin d'assurer la comparaison des résultats des élèves, les cahiers ont été conçus de façon à comporter des degrés de difficulté voisins. Cependant deux cahiers ne seront jamais exactement de même difficulté. La distribution des difficultés des items affecte donc la distribution des performances des élèves, reflétées par leurs scores de réussite. L'attribution aléatoire des différents cahiers aux élèves permet de supposer l'égalité de la

moyenne et de la variance des performances des élèves entre les différents tests (la standardisation des scores obtenus par les élèves gomme ensuite les écarts éventuellement existants).

Comme les élèves ne passent pas tous les mêmes épreuves, on ne peut pas observer leur score total à l'ensemble des items du test. Les scores totaux ne sont donc pas directement observables, et les scores obtenus ne sont pas directement comparables. Cependant, comme les items se recoupent entre les différents cahiers, les scores de compétence à l'ensemble du test peuvent être inférés, avec plus ou moins de fiabilité. En effet, ces estimations de la compétence s'accompagnent nécessairement d'une erreur de mesure substantielle, dont on souhaiterait rendre compte. En outre, comme nous l'avons déjà dit, l'objectif n'est pas d'estimer précisément les scores de réussite obtenus par chaque élève à l'évaluation, mais les scores de réussite obtenus par des groupes d'élèves dotés des mêmes caractéristiques (et notamment des populations nationales d'élèves). Pour ce faire, la procédure d'estimation des compétences doit prendre en compte les structures de ces groupes. Une manière de notifier l'incertitude associée aux estimations de la compétence et simultanément d'obtenir des estimations non biaisées des compétences de groupes d'élèves est d'utiliser plusieurs valeurs de score représentant la compétence probable de l'élève (Wu, 2005 ; von Davier, Gonzales et Mislevy 2009). Ces multiples valeurs tirées de la distribution a posteriori de la compétence d'un élève sont appelées des « valeurs plausibles ». Dans le cadre de PISA, ces valeurs plausibles sont inférées à partir des réponses au sous-ensemble d'items passé par l'élève et des informations contextuelles disponibles et pertinentes sur l'élève (Mislevy 1991 cité dans von Davier, Gonzales et Mislevy 2009). Il est nécessaire d'inclure des variables de contexte dans le modèle générant les valeurs plausibles si l'on souhaite ensuite estimer correctement les associations statistiques entre ces variables de contexte et les valeurs plausibles de compétence (Wu 2005 ; Monseur et Adams 2009). M. Wu (2005) explique que l'inclusion des variables de contexte dans le modèle qui produit les valeurs plausibles est dans l'intérêt des analyses secondaires de PISA qui examineraient les relations entre les compétences et les caractéristiques des élèves. Si ces études cherchent à régresser la compétence de l'élève sur un ensemble de variables contextuelles, alors le véritable coefficient de régression sera « retrouvé », seulement si le modèle qui produit ces valeurs plausibles de compétence incluait les variables de contexte. Ce modèle d'estimation de la compétence constitue une forme étendue du modèle de Rasch, appelé « modèle de Rasch de régression latente » ("the latent regression Rasch model", Wilson et De Boeck 2004). Ces procédures d'inférence des compétences sont détaillées dans le chapitre 9 du *PISA 2000 Technical Report* (Adams et Wu 2002).

Les valeurs plausibles représentent ainsi un ensemble de compétences que l'élève pourrait raisonnablement avoir, étant donné ses réponses aux items et ses caractéristiques (Wu 2005). Comme les valeurs plausibles ne sont pas des scores individuels au sens habituel, elles ne doivent donc pas être analysées de manière traditionnelle. Leur usage est délicat : toute analyse statistique faisant intervenir les compétences des élèves doit être mise en œuvre avec chacune des valeurs plausibles (*cf.* encadré 3).

Encadré 3 : De l'usage des valeurs plausibles

Les analyses statistiques faisant intervenir des variables de compétences des élèves doivent être conduites à partir de chacune des cinq valeurs plausibles. La procédure à mettre en œuvre est la suivante. Chaque statistique de population est estimée pour chacune des valeurs plausibles prises séparément. Toute statistique portant sur la performance de population finale correspond à la moyenne des cinq statistiques calculées à partir des cinq valeurs plausibles. Par exemple, si l'on souhaite calculer le coefficient de corrélation entre l'indice socioéconomique et la compétence en littératie des élèves, on doit d'abord calculer les cinq coefficients de corrélation entre l'indice socioéconomique et chacune des valeurs plausibles, puis en prendre la moyenne arithmétique. En termes statistiques, si θ est le paramètre d'intérêt de la population, et θ_i la statistique d'intérêt calculée pour la valeur plausible *i,* alors : $\theta = \frac{1}{5}\sum_{i=1}^{5}\theta_i$. On peut alors également calculer l'incertitude liée à la précision de la mesure. L'erreur de mesure (ou la variance associée à la mesure) est égale à : $B_M = \frac{1}{4}\sum_{i=1}^{5}(\theta - \theta_i)^2$. Enfin, on peut calculer la variance totale (V) à partir de l'erreur d'échantillonnage (notée U) et de l'erreur de mesure : $V = U + \left(1 + \frac{1}{4}\right)B_M$.

En raison de leur sophistication, ces méthodes ont été peu débattues par la communauté scientifique (notamment française) des sciences sociales. Les développements suivants se concentrent sur une série de critiques adressées seulement à l'un des principes du modèle de Rasch : le principe d'unidimensionnalité.

L'hypothèse d'unidimensionnalité signifie que le trait latent mesuré – la compétence – est unidimensionnel : chaque élève peut être situé sur le continuum des performances en fonction de ses réponses au test. Cela implique également que la dimension évaluée est invariante quel que soit le pays, c'est-à-dire qu'elle ne dépend pas des caractéristiques des élèves du pays. Dans le cadre de PISA, plusieurs sociologues, statisticiens et psychométriciens ont critiqué cette hypothèse tant du point de vue des résultats statistiques que des tâches évaluées (Goldstein 2004 ; Vrignaud 2006 ; Bautier *et al.* 2006 ; Rochex 2006 ; Goldstein, Bonnet et Rocher 2007).

Goldstein (2004) et Goldstein, Bonnet et Rocher (2007) suspectent tout d'abord le consortium de PISA d'avoir à tort favorisé l'invariabilité du test en n'impliquant que les pays participant effectivement au projet (évitant ainsi l'introduction de différences culturelles additionnelles entre les pays) et en retirant les items jugés louches (*dodgy*) lorsqu'ils étaient suspectés de biais culturels. L'objectif d'unidimensionnalité du test implique que, même en présence d'une seconde dimension qui s'exprimerait à travers plusieurs items, ces derniers seraient probablement jugés louches, puisqu'ils violeraient l'hypothèse en question. Par conséquent, les échelles de compétences construites ne comportent généralement qu'une seule dimension, mais souvent au prix de l'exclusion de certaines informations potentiellement importantes. Cependant, en appliquant de nouvelles analyses aux réponses des élèves anglais et français au test de mathématiques de PISA 2000, H. Goldstein est parvenu à la conclusion que les compétences des élèves seraient davantage bidimensionnelles qu'unidimensionnelles. Le modèle de réponse à l'item utilisé dans PISA serait donc trop restrictif car trop simpliste (il fait notamment l'hypothèse d'égale discrimination des

items[11]). Plusieurs chercheurs ont également pointé la contradiction apparente entre l'existence d'une échelle unidimensionnelle des compétences pour un domaine et la construction de plusieurs sous-échelles de compétences dans les sous-domaines (Goldstein 2004 ; Vrignaud 2006). En effet, la coexistence de ces échelles est satisfaisante du point de vue conceptuel mais pas nécessairement du point de vue psychométrique. P. Vrignaud (2006) suggère que les hypothèses de l'approche psychométrique retenue et la sophistication méthodologique qui en découle ont pu prendre le pas sur l'approche conceptuelle de la littératie que tente d'évaluer PISA.

Du point de vue de la conception de l'évaluation, l'hypothèse d'unidimensionnalité comporte deux implications. Premièrement, les différences interindividuelles résultent de différences de puissance entre les élèves ; deuxièmement, tous les élèves ont la même stratégie de résolution des items (Vrignaud 2006). À nouveau, la validité de cette hypothèse, du point de vue conceptuel, a été questionnée. É. Bautier et ses collègues (2006), et J.-Y. Rochex (2006) ont conduit de nouvelles analyses qualitatives et quantitatives à partir de deux cahiers tirés des évaluations PISA et administrés à une quarantaine d'élèves de 15 ans. Leur objectif était de mettre en évidence la variété des stratégies de résolution des élèves. Les réponses ont donc été recodées en fonction d'une grille élaborée pour tenter d'appréhender la diversité des registres mobilisés par les élèves. Les analyses confirment le caractère hétérogène des modes de travail et des univers de référence mobilisés par les élèves évalués par PISA. D'après les auteurs, l'hétérogénéité et la diversité des modes de résolution sont repérables non seulement dans les comparaisons que les auteurs ont pu faire entre les élèves ou entre diverses catégories d'élèves créées par le traitement statistique, mais aussi chez un même élève, d'une épreuve à l'autre, en fonction des thèmes abordés et des tâches attendues.

En conclusion, les discussions précédentes invitent à prendre avec précaution le principe d'unidimensionnalité des compétences évaluées par PISA. Ils ne disqualifient pas les résultats des enquêtes internationales mais cherchent à minorer les effets de palmarès et appellent une analyse plus détaillée des similitudes et des différences entre pays.

[11] La discrimination de l'item renvoie à la capacité de l'item à bien distinguer entre les élèves compétents et les élèves moins compétents. Un item très discriminant renseigne bien sur la compétence de l'élève tandis qu'un item peu discriminant apporte peu d'information sur la compétence de l'élève. L'indice de discrimination de l'item correspond généralement à la corrélation entre l'item et le trait latent évalué. Un item sera dit discriminant si les élèves qui le réussissent, réussissent en moyenne mieux le reste du test que les élèves qui ne répondent pas correctement à cet item (Vrignaud 2008). L'hypothèse d'égale discrimination des items suppose que les items apportent tous la même qualité et la même quantité d'information sur les compétences des élèves. Selon P. Virgnaud (2008), cette condition est en général vérifiée *a posteriori* puisque les tests de validation du modèle de Rasch confortent bien l'adéquation du modèle aux données sans que l'introduction du paramètre de discrimination soit nécessaire.

2.5. L'harmonisation internationale des données contextuelles

Outre les épreuves d'évaluation des élèves, le programme PISA comprend deux questionnaires « contextuels », l'un rempli par l'élève et l'autre par le chef d'établissement, son adjoint ou un conseiller principal d'éducation. Un temps de réponse de 20 à 30 minutes est prévu pour ces questionnaires. Le questionnaire « établissement » porte sur les ressources humaines et matérielles, le secteur, le mode de gestion et de financement, les processus de décision et le climat général de l'établissement. Les informations collectées sont cruciales car elles permettent de faire le lien entre les résultats et les caractéristiques des élèves et des établissements. Les questions aux élèves portent sur leur environnement familial, y compris leurs ressources sociales, économiques et culturelles, ainsi que sur les attitudes des élèves vis-à-vis de l'apprentissage à l'école et dans leur famille. L'harmonisation du recueil et du codage de ces informations contextuelles est nécessaire pour assurer la comparabilité des analyses entre pays. Or, trouver la meilleure façon d'harmoniser les variables sociodémographiques habituellement présentes dans les enquêtes nationales, constitue également un défi (Joye et Mochmann 2011). Ce problème concerne dans leur globalité tout le processus d'interrogation des individus et toute la mise en forme des données collectées.

L'interrogation des acteurs du système éducatif

Le dispositif PISA s'illustre par l'enregistrement de la voix des élèves et, dans l'immense majorité des pays, des chefs d'établissement.[12] D'aucuns regrettent même que les professeurs ne soient pas interrogés (Mulford 2002). L'interrogation des acteurs du système scolaire constitue indéniablement un point positif de l'enquête. Cependant, la fiabilité et la qualité des informations recueillies restent discutables. Les déclarations d'élèves de 15 ans relatives à leur environnement familial peuvent présenter des discordances avec la réalité. En effet, il est fort probable que les enfants ne soient pas toujours en mesure de fournir des informations exactes concernant leurs parents, notamment leur profession ou encore leur niveau d'études (Bonnet 2002 ; Adams et Wu 2002). En outre, certaines réponses au questionnaire « contextuel » requièrent d'avoir une idée relativement précise du volume de ressources au domicile, par exemple du nombre de livres, ce qui n'est pas toujours une tâche aisée. Ainsi, des dissonances entre les réponses formulées par l'élève et la situation effective peuvent apparaître et affecter les résultats. Cependant, même si les déclarations des élèves peuvent s'éloigner de la réalité, elles reflètent néanmoins leur vision de leur environnement familial. Et sauf à penser que les adolescents de 15 ans se méprennent complètement sur

[12] Nous signalons que la France n'a pas autorisé, après PISA 2000, que ce questionnaire soit rempli par les chefs d'établissement. En 2003, 2006 et 2009, aucune information n'a donc été recueillie en France auprès des établissements. Les raisons du refus (regrettable) d'administrer ce questionnaire demeurent relativement opaques.

leur situation, on peut considérer que leurs déclarations correspondent à des effets « objectifs », éventuellement atténués ou amplifiés, mais bien réels[13].

L'exploitation des données collectées auprès des chefs d'établissement pose également d'éventuels problèmes de décalage avec la réalité pour deux raisons principales. D'abord, le questionnaire adressé aux établissements requiert une bonne connaissance des ressources et des actions conduites par les acteurs de ces institutions. Ce questionnement exhaustif pourrait générer des biais dans les réponses des chefs d'établissement, si ces derniers ne sont pas suffisamment informés des pratiques internes. Ensuite, les interrogés pourraient consciemment infléchir leurs réponses en fonction de l'exploitation qu'ils anticipent de l'enquête. Par exemple, ils pourraient être incités à déclarer un manque de ressources éducatives quelle que soit la situation.

Le recours à des classifications internationales de référence et à des indices

L'harmonisation des données contextuelles présentent des difficultés certaines : elles tiennent au fait que l'origine sociale des élèves, appréhendée via les niveaux d'éducation et les professions occupées par les parents, présente de fortes spécificités nationales. Dans un chapitre de l'ouvrage *La France dans les comparaisons internationales* (Chenu et Lesnard 2011), L.-A. Vallet rappelle que ces variables n'ont fait que récemment l'objet d'un codage via des nomenclatures internationales dans les enquêtes à large échelle. Les enquêtes PISA utilise d'une part la Classification Internationale Type des Professions (CITP) ou *International Standard Classification of Occupations* (ISCO) établie par le Bureau International du Travail pour le codage des professions, d'autre part la Classification Internationale Type de l'Éducation (CITE) ou *International Standard Classification of Education* (ISCED), élaborée sous l'égide de l'UNESCO, pour la transcription des niveaux d'éducation. Le recours à ces nomenclatures de référence pour le codage des caractéristiques individuelles et familiales des élèves favorise toute entreprise comparative de qualité à partir des données PISA.[14] Dans de nombreux cas, les réponses des élèves et des chefs d'établissements ont été recodées et échelonnées afin d'obtenir de nouvelles variables synthétiques, à savoir des indices (Adams et Wu 2002).

C'est à partir de la nomenclature ISCO qu'a pu être construit l'indice socioéconomique international (ISEI), dû à H. Ganzeboom, P. de Graaf et D. Treiman (1992). Les ISEI de la mère et du père sont les seuls indicateurs synthétiques de l'environnement socioéconomique de l'élève à figurer dans chacune des vagues d'enquêtes PISA. L'ISEI est utilisé dans de nombreuses analyses secondaires de PISA pour

[13] L'âge de 15 ans est le seuil souvent retenu pour interroger des jeunes indépendamment de leurs parents (Joye, 2011). C'est notamment la limite inférieure choisie pour définir la population sondée dans le cadre de l'enquête European Social Survey (ESS).
[14] Ces deux nomenclatures internationales ne sont cependant pas sans poser de question (*cf.* les chapitres 8 et 10 de l'ouvrage *Advances in Cross-national Comparison* (Hoffmeyer-Zlotnik et Wolf, 2003) qui présentent respectivement le rôle et les limites des classifications ISCO et ISCED).

caractériser l'origine sociale de l'élève. Ces éléments justifient que l'on s'attarde un peu sur sa construction et sa signification.

Dans le questionnaire contextuel de PISA, il est demandé aux élèves de préciser la profession de chacun de leurs parents. Deux questions leur sont posées à ce sujet : "What is your mother/father currently doing?" et "What does your mother/father do in her/his main job? (e.g., school teacher, nurse, sales manager). *If she/he is not working now, please tell us her/his last main job*". Les réponses ouvertes données par les élèves permettent alors de coder la profession de chacun des parents suivant la CITP. Ce sont ensuite les valeurs des ISEI, codés à partir des variables de profession, qui ont pu être ajoutées à la base de données PISA. Cet indice associe à chaque profession de la CITP[15] une valeur égale à la somme pondérée du niveau d'éducation moyen et du revenu moyen du groupe socioprofessionnel.

L'ISEI est en filiation directe avec la vision anglo-saxonne de la stratification sociale. Selon H. Ganzeboom, P. De Graaf et H. Treiman (1992), les positions des systèmes de stratification sociale se laissent appréhender de trois façons. Par des catégories sociales – par exemple, la nomenclature française des catégories socioprofessionnelles de l'Insee (Institut national de la statistique et des études économiques) ou la CITP, ou encore la nomenclature d'Erikson, Goldthorpe et Portocarero (EGP) –, par des échelles de prestige ou ratings (par exemple la *Standard International Occupational Prestige Scale*) et enfin par des scores mesurant le statut socio-économique (*International Socio-Economic Index of Occupational Status*). À chaque instrument de mesure de la position sociale correspond une conception de la stratification sociale. La classification des catégories sociales repose sur l'idée, relativement partagée par les sociologues français, d'un espace social à la fois multidimensionnel (conformément à la vision bourdieusienne) et discontinu (conformément à la conception boudonienne). Dans le monde anglo-saxon, les positions sociales sont généralement appréhendées par une échelle unidimensionnelle construite à partir d'un critère parfois subjectif (le prestige) ou d'un indice composite de critères. C'est le cas pour l'indice socioéconomique international (ISEI) de H. Ganzeboom, P. De Graaf et D. Treiman (1992).

Les approches « continues » de la stratification sociale diffèrent des approches catégorielles de deux points de vue. Premièrement, elles autorisent un nombre illimité de distinctions entre les groupes professionnels. Deuxièmement, elles supposent que des différences substantielles entre les groupes professionnels peuvent être reflétées par une seule dimension et par conséquent être représentées dans des modèles statistiques par un unique paramètre. Les approches continues sont puissantes puisqu'elles résument de nombreuses distinctions à l'aide d'un seul chiffre. En dépit du caractère multidimensionnel de la stratification sociale, de bonnes raisons justifient le recours à l'indice socio-économique international. En

[15] C'est la CITP de 1988 qui est utilisée dans PISA.

effet, une variable quantitative continue est plus aisément manipulable qu'une ou plusieurs variables catégorielles. L'interprétation des paramètres devient plus facile et plus réaliste et la perte d'information liée à la compression des données lors du passage au continu est compensée par la possibilité de mener des analyses multivariées.

La construction de l'ISEI découle de l'interprétation donnée par O. Duncan (1961) du rôle joué par la profession dans les liens entre niveau d'éducation et niveau de revenus : la profession est la composante qui permet la conversion de l'éducation en revenus. L'indice associé à chaque catégorie professionnelle est conçu comme la variable intermédiaire qui maximise l'effet indirect de l'éducation sur le revenu et qui minimise l'effet direct de l'éducation sur le revenu. L'ISEI a donc pour but de capter le plus possible des différences de niveaux d'éducation et de revenu entre les catégories socioprofessionnelles de la CITP. La méthode de construction suivie par les auteurs a permis d'établir une échelle des professions qui explique de façon optimale la relation entre niveau d'éducation et niveau de revenu et qui satisfait la définition d'O. Duncan. Techniquement, l'ISEI est calculé comme une somme pondérée des niveaux standardisés d'éducation et de revenus, contrôlés des effets d'âge[16].

Quatre pays participant à PISA 2000 – le Canada, la République Tchèque, la France et le Royaume-Uni – ont évalué la validité des réponses des élèves concernant les professions de leurs parents et des indices socioéconomiques correspondant. Ces études ont en général conclu que les informations récoltées auprès d'élèves de 15 ans sont fiables et ne sont pas moins précises que les réponses obtenues auprès d'adultes interrogés sur la profession de leur conjoint (Adams et Wu 2002). L'essentiel, ici, est que l'ISEI dérivé des réponses de l'élève soit très corrélé à l'ISEI qu'on aurait déduit de la déclaration du parent. L'important est bien que ces deux valeurs de l'ISEI soient très proches. Les quatre études menées dans les quatre pays mentionnés ci-dessus ont conclu que la corrélation entre ces deux scores d'ISEI est élevée (le coefficient de corrélation est compris entre 0,70 et 0,86 suivant le pays).

L'exemple de l'ISEI est un exemple parmi d'autres. Dans de nombreux cas, les réponses des élèves et des chefs d'établissements ont été recodées et échelonnées afin d'obtenir des indices synthétiques plutôt que de nombreuses variables catégorielles (Adams et Wu 2002).

[16] Les données utilisées pour l'estimation de l'échelle concernaient un échantillon de plus de 70 000 hommes actifs âgés de 21 à 64 ans, issu de 16 pays développés. Les valeurs de l'indice ISEI sont comprises entre 15 et 90. En utilisant cette échelle, les analyses secondaires de PISA font donc l'hypothèse que la distance entre deux professions est conservée dans l'ensemble des pays participants à PISA.

Chaque indice complexe[17] synthétise une série de réponses d'élèves ou de chefs d'établissements à un ensemble de questions portant sur un aspect, conformément aux considérations théoriques et aux résultats de recherches empiriques retenus. Par exemple, l'indice de richesse de la famille ("family wealth") est construit à partir des déclarations des élèves concernant la disposition au domicile familial d'une machine à laver la vaisselle, d'une chambre à coucher pour l'élève seul, de logiciels éducatifs et d'une connexion à Internet, ainsi que le nombre de téléphones, de télévisions, d'ordinateurs, d'automobiles et de salles de bain du foyer. L'indice d'autonomie de l'établissement est, quant à lui, construit à partir des déclarations du chef d'établissement. Des techniques de psychométrie ont été utilisées pour échelonner les indices. Puis, des modèles d'équation structurelle ont été appliqués pour confirmer la validité des indices et leur comparabilité entre les pays. Les concepteurs de PISA ont évalué la fiabilité de ces indices : leur validité interne au sein de chacun des pays est modérée mais cependant satisfaisante, compte tenu du faible nombre d'items pour chaque échelle (Adams et Wu 2002).

On peut cependant interroger la pertinence de l'objectivation statistique que ces indices incarnent et douter de leur comparabilité entre les pays (Bonnet 2002). Par exemple, l'indice de richesse de la famille tente de synthétiser des items hétéroclites qui ne sont pas dotés des mêmes significations économique et sociale et peuvent avoir des influences variables sur les résultats scolaires des élèves suivant les pays. Néanmoins, les différences internationales dans la signification des ressources familiales sont essentiellement constatées entre des pays en voie de développement participant à PISA (*cf.* Kespaik et Rocher 2011). En travaillant seulement sur les pays d'une même aire régionale (par exemple les pays de l'Union européenne), c'est-à-dire sur des pays bénéficiant d'un niveau de développement socio-économique comparable et partageant des traits culturels communs, on ménage un minimum de comparabilité entre les indices caractéristiques des élèves et des établissements.

Conclusion

Le consortium PISA, mu par une forte volonté de comparabilité des résultats, a porté un très grand soin à la conception et à la mise en œuvre des enquêtes. Le programme PISA est apparu innovant du point de vue des objectifs poursuivis et de leur opérationnalisation. Il évalue périodiquement les compétences en littératie des élèves de 15 ans. Il cible ainsi une classe d'âge et non un niveau scolaire. Il porte sur les connaissances et savoir-faire utiles à la vie en société et non sur des points précis des programmes scolaires. Évaluant tous les trois ans les élèves de dizaines de pays occidentaux et non occidentaux, il offre

[17] La construction de l'ensemble des indices est présentée dans le chapitre 17 du *PISA Technical Report* (Adams et Wu 2002).

de comparer les aptitudes des élèves à la fois dans l'espace et dans le temps. Les enquêtes fournissent un volume conséquent d'information contextuelle sur les élèves et leurs établissements. Le souci de comparabilité s'est traduit par l'imposition de nombreux standards, de procédures de contrôle et par d'importantes sophistications méthodologiques.

Ces choix méthodologiques ne sont pas sans conséquence pour le type d'études à mener sur les données PISA. Nous conclurons cette note en insistant sur trois implications du dispositif pour les analyses secondaires. Premièrement, l'instruction des choix méthodologiques opérés invite à porter l'attention sur les contextes éducatifs et socioculturels nationaux dans lesquels s'inscrivent les évaluations PISA. Il nous est apparu judicieux de combiner des analyses effectuées à partir de plusieurs échelles spatiales. D'abord une comparaison sur un nombre important de pays permettrait de dégager les facteurs explicatifs principaux des phénomènes étudiés. Ensuite, à partir de ces conclusions, une autre comparaison sur un petit ensemble de pays favoriserait la mobilisation d'éléments contextuels précis et l'analyse approfondie des mécanismes en jeu. Deuxièmement, le consortium PISA a veillé à ce que l'on puisse juger de la significativité des écarts constatés entre pays. Cette opportunité est essentielle à toute entreprise comparative. Ainsi, les données disponibles dans les bases PISA permettent de rendre compte de l'incertitude contenue dans les mesures. Il est donc recommandé de considérer les erreurs-types (ou les intervalles de confiance) associés aux paramètres statistiques calculés, afin d'apprécier la significativité des différences observées entre les pays. Troisièmement, la structure hiérarchique des données PISA, organisés en deux niveaux au sein de chaque pays (l'élève et l'établissement), commande la mise en œuvre de modèles de régression multi-niveaux (encore appelés modèles à effets mixtes ou modèles linéaires hiérarchiques). Contrairement aux modèles linéaires, ces modèles tiennent compte des effets générés par le regroupement des élèves au sein des établissements, et permettent d'évaluer les variations relatives dans la variable d'intérêt entre les élèves au sein des établissements et entre les établissements.

Enfin, fort de la connaissance des atouts et des limites des données PISA, le chercheur doit désormais prendre au sérieux la construction des outils de mesure et de comparaison du programme PISA. Dans ses analyses, il s'attachera à expliciter systématiquement les principes de construction des données et des variables et la méthodologie statistique employée pour procéder aux analyses (Desrosières 1989 ; Gollac 1997). La restitution de ces processus autorisera alors des interprétations éveillées et éclairées des résultats.

BIBLIOGRAPHIE

Adams R. J., 2003. “Response to ‘cautions on OECD’s recent educational survey (PISA)’”. *Oxford Review of Education* 29(3): 377-389.

Adams R. J., Wu M. L. (eds), 2002. *PISA 2000 Technical Report*. Paris: OECD Publications.

Baudelot C., Establet R., 2009. *L’élitisme républicain. L’école française à l’épreuve des comparaisons internationales*. Paris: Seuil, La République des Idées.

Bautier E., Crinon J., Rayou P., Rochex J.-Y., 2006. « Performances en littéracie, modes de faire et univers mobilisés par les élèves : analyses secondaires de PISA 2000 ». *Revue française de pédagogie* (157): 85-101.

Bodin A., 2005. “What does PISA assess? What it doesn’t? A French view”. Communication présentée à la conférence franco-finnoise *Enseigner les mathématiques : au-delà de l’enquête PISA*. Paris: 6-8 octobre. http://smf4.emath.fr/VieSociete/Rencontres/France-Finlande-2005/BodinGB.pdf

Bonnet G., 2002. “Reflections in a critical eye: On the pitfalls of international assessment in education: Principles, policy and practice”. *Assessment in Education* 9(3): 387-399.

Bottani N., Vrignaud P., 2005. *La France et les évaluations internationales*. Rapport national au HCEE. Paris: HCEE.

Bulle N., 2010. « L’imaginaire réformateur – PISA et les politiques de l’école », *Revue skhole.fr-penser et repenser l’école*. 18 janvier. http://skhole.fr/l-imaginaire-r%C3%A9formateur-pisa-et-les-politiques-de-l-%C3%A9cole

Chenu A., Lesnard L. (dir.), 2011. *La France dans les comparaisons internationales. Guide d’accès aux grandes enquêtes statistiques en sciences sociales*. Paris: Presses de la Fondation nationale des sciences politiques.

von Davier M., Gonzales E., Mislevy R. J., 2009. “What are plausible values and why are they useful?”. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments* 2 (chapitre 1): 9-36.

Desrosières A., 1989. « Comment faire des choses qui tiennent : histoire sociale et statistique ». *Histoire et Mesure* 4(3-4): 225-242.

Duru-Bellat M., Mons N., Suchaut B., 2004. « Caractéristiques des systèmes éducatifs et compétences des jeunes de 15 ans. L’éclairage des comparaisons entre pays ». *Cahiers de l’Iredu* (66).

Félouzis G., Charmillot, S., 2012. *Les enquêtes PISA*. Paris: PUF.

Fertig M., 2004. “What can we learn from international student performance studies. Some methodological remarks”. *RWI Working Paper* (23).

Ganzeboom H. B. G., De Graaf P. M., Treiman D. J., 1992. “A standard international socio-economic index of occupational status”. *Social Science Research* (21): 1-56.

Goldstein H., 2004. “International comparisons of student attainment: Some issues arising from the PISA study”. *Assessment in Education* 11(3): 319-330.

Goldstein H., Bonnet G., Rocher T., 2007. “Multilevel structural equation models for the analysis of comparative data on educational performance”. *Journal of Educational and Behavioral Statistics* 32(3): 252-286.

Gollac M., 1997. « Des chiffres insensés ? Pourquoi et comment on donne un sens aux données statistiques ». *Revue française de sociologie* 38(1): 5-36.

Grenet J., 2008. « PISA : une enquête bancale ? ». *La vie des idées*. 8 février. http://www.laviedesidees.fr/PISA-une-enquete-bancale.html

Guérin-Pace F., Blum, A., 1999. « L'illusion comparative: les logiques d'élaboration et d'utilisation d'une enquête internationale sur l'illettrisme ». *Population* 54(2): 271–302.

Hoffmeyer J. H. P., Wolf C. (eds), 2003. *Advances in Cross-National Comparison: A European Working Book for Demographic and Socio-Economic Variables*. New York: Kluwer and Plenum.

Joye D., 2011. « Échantillonnage et pondération », p. 143-152 *in* A. Chenu, L. Lesnard (dir.). *La France dans les comparaisons internationales. Guide d'accès aux grandes enquêtes statistiques en sciences sociales*. Paris: Presses de la Fondation nationale des sciences politiques.

Joye D., Mochmann E., 2011. « Comparaisons interculturelles et recherches comparatives internationales », p. 153-160 *in* A. Chenu, L. Lesnard (dir.). *La France dans les comparaisons internationales. Guide d'accès aux grandes enquêtes statistiques en sciences sociales*. Paris: Presses de la Fondation nationale des sciences politiques.

Mons N., Pons X., 2009. "The reception of PISA in France: A cognitive approach of institutional debate (2001-2008)". *Sisifo. Educational Sciences Journal* (10): 27-40.

Monseur C., Adams R., 2009. "Plausible values: How to deal with their limitations". *Journal of Applied Measurement* 10(3): 320-334.

Mulford B., 2002. "Sorting the wheat from the chaff – knowledge and skills for life: First results from OECD's PISA 2000". *European Journal of Education* 37(2): 211-221.

OECD, 1999. M*esurer les compétences et connaissances des élèves. Un nouveau cadre d'évaluation*. Paris: OCDE.

—. 2005. *PISA 2003 Data Analysis Manual*. Paris: OCDE.

—. 2010. *PISA 2009 Results: What Students Know and Can Do – Student Performance in Reading, Mathematics and Science (volume 1).* Paris: OCDE.

Olsen R. V., Lie S., 2006. « Les évaluations internationales et la recherche en éducation : principaux objectifs et perspectives ». *Revue française de pédagogie* (157): 11-26.

O'Leary M., 2001. "The effects of age-based and grade-based sampling on the relative standing of countries in international comparative studies of student achievement". *British Educational Research Journal* 27(2): 187-200.

Phillips D., Ochs K., 2004. "Researching policy borrowing: Some methodological challenges in comparative". *British Educational Research Journal* 30(6): 773-784.

Prais S. J., 2003. "Cautions on OECD's recent educational survey (PISA)". *Oxford Review of Education* 29(2): 139-163.

—. 2004. "Cautions on OECD's recent educational survey (PISA): Rejoinder to OECD's response". *Oxford Review of Education* 30(4): 569-573.

Prenzel M., Zimmer K., 2006. « Études complémentaires de PISA 2003 en Allemagne : principaux résultats et enseignements ». *Revue française de pédagogie* (157): 55-70.

Rémond M., 2006. « Éclairages des évaluations internationales PIRLS et PISA sur les élèves français ». *Revue française de pédagogie* (157): 71-84.

Robin I., 2002. « L'enquête PISA sur les compétences de lecture des élèves de 15 ans : trois biais culturels en question ». *VEI Enjeux* (129): 65-91.

Robin I., Rocher T., 2002. « La compétence en lecture des jeunes de 15 ans. Une comparaison internationale », p. 93-101 *in* Insee *Données sociales : La société française*, Paris: Insee.

Rocher T., 2003. « La méthodologie des évaluations internationales de compétences ». *Psychologie et Psychométrie* 24(2-3): 117-147.

Rochex J.-Y., 2006. “Social, methodological, and theoretical issues regarding assessment: Lessons from a secondary analysis of PISA 2000 literacy”. *Review of Research in Education*, 30 (Special Issue on Rethinking Learning: What Counts as Learning and What Learning Counts): 163-212.

Romainville M., 2002. « Du bon usage de PISA ». *La Revue nouvelle* (3-4): 86-99.

Safi M., 2011. « L’analyse longitudinale. Données et méthodes », p.161-172 *in* A. Chenu, L. Lesnard (dir.). *La France dans les comparaisons internationales. Guide d’accès aux grandes enquêtes statistiques en sciences sociales*. Paris: Presses de la Fondation nationale des sciences politiques.

Vallet L.-A., 2011. « Construction, inégalité et mobilité des statuts professionnels et sociaux », p.103-120 *in* A. Chenu, L. Lesnard (dir.). *La France dans les comparaisons internationales. Guide d’accès aux grandes enquêtes statistiques en sciences sociales*. Paris: Presses de la Fondation nationale des sciences politiques.

Vrignaud P., 2006. « La mesure de la littératie dans PISA : la méthodologie est la réponse, mais quelle était la question ? ». *Revue française de pédagogie* (157): 27-41.

—. 2008. « La mesure de la littératie dans PISA : la méthodologie est la réponse, mais quelle était la question ? ». *Éducation et formations* (78): 69-84.

Warm T. A., 1989. “Wheighted likelihood estimation of ability in item response theory”. *Psychometrika* 54(3): 427-450.

Wilson M., De Boeck P., 2004. “Descriptive and explanatory item response models”, p. 43-74 *in* P. De Boeck, M. Wilson (eds). *Explanatory Item Response Models. A Generalized Linear and Nonlinear Approach.* New York: Springer.

Wu M., 2005. “The Role of plausible values in large-scale surveys”. *Studies in Educational Evaluation* (31): 114-128.