La mesure de la littératie dans PISA : la méthodologie est la réponse, mais quelle était la question ?

Ce texte reprend celui d'un article paru dans la Revue française de Pédagogie, *INRP, 2006.*

L'objectif de cet article est de montrer l'interaction entre les aspects méthodologiques et la manière dont est conceptualisée et définie la littératie dans l'enquête PISA. Pour introduire le thème on évoque d'abord les changements conceptuels dont a été l'objet l'évaluation des acquis des élèves dans les enquêtes internationales depuis les premières enquêtes de l'IEA jusqu'à l'enquête PISA. Après avoir rappelé que toute mesure est un construit, on expose les théories psychométriques qui fondent les modèles de mesure utilisés pour l'évaluation des acquis des élèves. La présentation des approches classiques permet d'introduire les concepts de difficulté et de discrimination des items ainsi que le concept de consistance globale d'une épreuve. On insiste sur les conditions de validité des modèles psychométriques qui posent des contraintes fortes sur la façon dont la variable mesurée est construite. On présente ensuite le modèle de mesure utilisé dans PISA qui appartient à la famille des modèles de réponse à l'item (MRI). PISA a privilégié le modèle dit de Rasch, on discute des conséquences de ce choix. On aborde ensuite un point très technique peu souvent abordé : celui de l'algorithme utilisé pour l'estimation des paramètres. La procédure utilisée aboutit à estimer non pas une valeur par sujet mais à construire la distribution des compétences de chaque sujet, on insiste sur les conséquences de cette approche sur la manière de conduire l'analyse des résultats. Ensuite, on discute du caractère réellement unidimensionnel de la variable du point de vue des contenus et du point de vue méthodologique. On s'interroge sur la contradiction apparente qu'il peut y avoir à considérer simultanément des sous-échelles et une échelle globale. En conclusion, on revient sur la manière d'interpréter la littératie telle qu'elle est mesurée dans PISA en la mettant en relation avec d'autres enquêtes visant à évaluer le même type de compétence.

Pierre Vrignaud
vrignaud.pierre@wanadoo.fr
Université Paris X Nanterre
EA 39 84 - Laboratoire « Travail et évolution professionnelle »

Descripteurs (TEE) :
Mots-clés : Littératie, comparaisons internationales, psychométrie, MRI

Les enquêtes internationales sur l'évaluation des acquis des élèves et parmi elles, l'enquête PISA témoignent des changements conceptuels profonds qui ont modifié l'objet de ces enquêtes au cours de la dernière décennie. À la différence des enquêtes internationales conduites pendant un demi-siècle par l'IEA[1], principalement centrées sur les acquis définis à partir des *curricula*, PISA (OECD, 1999) a introduit l'idée qu'il est plus pertinent d'évaluer les compétences pour travailler et vivre dans une société post-industrielle, compétences considérées comme devant être le produit, l'*output* des systèmes éducatifs (Bottani et Vrignaud, 2005). Ce choix s'inscrit tout à fait dans la logique des enquêtes américaines sur ce thème (NAEP, YALS, NALS[2] voir Johnson, 1992) et de la première enquête internationale sur la littératie (IALS[2]; Murray, Kirsch et Jenkins, 1998). Cette approche en termes de compétences plutôt que d'acquis peut apparaître comme l'œuf de Colomb des enquêtes internationales. Elle a permis d'éviter les difficultés que posait la construction d'un « méta-curriculum » - opération indispensable aux premières enquêtes mises en place par l'IEA à partir des années 1950 -, en faisant glisser la mesure du contenu des enseignements vers une compétence suffisamment large pour considérer que tous les systèmes scolaires ont pour objectif de l'enseigner.

Dans cette optique, on argumente l'intérêt des enquêtes internationales en général et de l'enquête PISA en particulier par le fait qu'elles fournissent des informations sur des compétences très générales – transversales – qui interviennent dans la plupart des situations de la vie quotidienne et plus particulièrement de la vie professionnelle. Telles qu'elles sont présentées, ces compétences peuvent être considérées comme des interfaces entre la formation initiale dont elles sont le produit et le monde professionnel où elles sont mises en œuvre. Ces compétences sont porteuses de sens pour

NOTES

1. *International Association for the Evaluation of Educational Achievement*, en français : Association internationale pour l'évaluation du rendement scolaire.

2. *National Assessment of Education in Progress; Young Adult Litteracy Assessment; National Adult Litteracy Assessment; International Adult Litteracy Survey.*

des utilisateurs divers, tant les chercheurs (économistes, psychologues, sociologues, sciences de l'éducation) que les décideurs des politiques éducatives ainsi que les médias. Si l'avantage de cette approche est de présenter une validité écologique importante de par son large spectre d'utilisation, son principal inconvénient est le danger de réification auquel cette compétence peut donner lieu. Dans le cadre de l'évaluation psychométrique qui est celui de ces enquêtes, les compétences sont avant tout des construits et ne sont pas séparables de la manière dont ces construits ont été opérationnalisés. Perdre de vue cette caractéristique laisse la porte ouverte à des généralisations abusives voire à des extrapolations que ne soutiennent pas réellement l'interprétation des résultats.

Ce danger est bien identifié et connu historiquement en psychologie à partir des problèmes liés à la mesure de l'intelligence. On a d'abord critiqué les tests d'aptitudes intellectuelles pour leur caractère réducteur (mesure d'une intelligence plutôt académique) et socialement biaisé (en particulier les débats autour de la possibilité de construire des tests *culture free* ou *culture fair* ; sur ce point voir Vrignaud, 2002). Puis, les apports de la psychologie cognitive ont montré que l'approche unitaire de l'intelligence, véhiculée par l'emploi d'un score unique (par exemple le QI), ne reflétait pas ou, du moins, reflétait imparfaitement le fonctionnement psychologique des sujets (sur ce thème voir Huteau et Lautrey, 1999). La pluralité des processus et des stratégies mis en œuvre par les sujets pour résoudre les problèmes proposés n'était pas prise en compte par le score global. Les variabilités tant inter qu'intra-individuelles obéraient fortement la pertinence de l'interprétation d'un score unique.

Ces critiques et débats font parfois oublier que la construction de tests a permis le développement des méthodes et des concepts qui fondent la mesure en sciences humaines, en particulier, la psychométrie. La pierre de touche de l'évaluation en psychologie et en éducation est de distinguer entre une performance et une compétence. On observe une performance d'un sujet à une épreuve et on infère sur sa compétence (Mislevy, 1994). Loin des représentations naïves que l'idée de calcul d'un score peut véhiculer, on cherche à estimer la compétence ces sujets. Il s'agit d'un processus faisant appel à de nombreux concepts psychométriques et l'estimation de la compétence obtenue est aussi éloignée du simple calcul d'un score par sommation des bonnes réponses que peuvent l'être les premiers travaux sur les localisations cérébrales des recherches actuelles bénéficiant des avancées les plus récentes des techniques d'imagerie cérébrale..

Les enquêtes internationales ont été, depuis leur origine, un laboratoire d'essai et de développement des méthodologies psychométriques les plus sophistiquées (on trouvera une présentation très complète en français de la méthodologie des enquêtes internationales dans Rocher, 2003). Il faut dire qu'en plus du désir de l'ensemble des participants de fournir aux utilisateurs des résultats présentant les meilleures garanties de fiabilité, ces enquêtes présentaient des problèmes méthodologiques redoutables comme celui d'assurer l'équivalence de la mesure dans de multiples contextes linguistiques et nationaux. Les méthodes utilisées comme on peut le constater à la lecture du *technical manual* de PISA (Wu et Adams, 2002) sont extrêmement sophistiquées. Il paraît donc utile de donner aux lecteurs intéressés un aperçu schématique de quelques uns principaux points méthodologiques comme la construction de l'échelle de compétence et l'algorithme d'estimation des scores à ces échelles. Cet exposé est évidemment technique mais c'est justement un des problèmes cruciaux de ces enquêtes que la compréhension des résultats et surtout de leurs limites est liée à des questions méthodologiques complexes.

De fait, l'enquête PISA est un dispositif de mesure de la littératie et l'interprétation de ses résultats doit se faire en gardant présent à l'esprit la manière dont cette compétence a été construite. Il est donc utile pour expliciter ce qu'est et n'est pas la compétence évaluée dans PISA, de donner et de discuter les éléments méthodologiques qui valident ce passage entre la performance à un ensemble de tests et la compétence de populations de nombreux pays. L'objectif de cet article est de montrer les importantes avancées méthodologiques qui ont été intégrées dans PISA pour construire un dispositif de mesure solide et, en même temps, de montrer que l'accent mis sur le dispositif de mesure a peut-être laissé dans l'ombre d'autres interrogations sur la nature et la mesure de la compétence. Cet exposé nécessitera quelques rappels historiques qui montreront que les méthodes utilisées dans PISA sont le produit d'une longue histoire : celle de la psychométrie et des enquêtes sur l'évaluation des acquis des élèves, en particulier aux États-Unis. On présentera d'abord les principaux concepts psychométriques selon l'approche classique. Puis, on présentera le modèle de mesure utilisé dans les enquêtes internationales

en général et dans PISA en particulier. On insistera à la fois sur les avancées réalisées pour la mise au point de ce dispositif et sur les difficultés qui peuvent se rencontrer dans sa mise en œuvre et sa bonne compréhension par les utilisateurs. En conclusion, on reviendra sur les relations entre le dispositif de mesure et la nature de la compétence.

L'approche classique de la mesure psychométrique

La majeure partie des méthodes utilisées pour les enquêtes internationales sur les acquis des élèves ont été élaborées au sein de la psychologie ou plutôt de la psychométrie. On parle aujourd'hui de « l'édumétrie » pour définir un champ équivalent à celui de la psychométrie dans le domaine de l'évaluation en éducation. Cette distinction reste cependant une distinction de surface dans la mesure où les méthodes et les concepts sont largement similaires et où, bien souvent, les chercheurs qui travaillent et publient dans l'un de ces deux champs travaillent et publient également dans l'autre.

Les modèles de mesure

Pour introduire cette présentation des concepts de base de la psychométrie, on peut rappeler que la mesure, c'est à dire l'assignation de grandeurs à des objets en respectant certaines propriétés de ceux-ci, a posé en psychologie des problèmes particuliers qui ont abouti au développement de solutions originales au sein de cette discipline. Ces méthodes se sont trouvées rassemblées dans la psychométrie qui définit les méthodes à mettre en œuvre, depuis les dispositifs de collecte des données jusqu'à la définition de normes de fiabilité (pour une présentation des théories et méthodes psychométriques, on se reportera, en français, à des ouvrages comme ceux de Dickes *et al.*, 1994 ou de Laveault et Grégoire, 2002). La démarche de validation de la mesure en psychométrie repose sur le principe selon lequel toute mesure est un construit. On parlera ici d'un modèle de mesure, et la démarche hypothético-déductive consiste à tester l'adéquation de ce modèle de mesure aux données. Plusieurs approches peuvent être mises en œuvre pour tester cette adéquation (on en trouvera une présentation dans les ouvrages cités plus haut). Les trois modèles de mesure les plus généralement utilisés sont l'approche classique (formalisée par Lord et Novick, 1969), les *modèles de réponse à l'item* (MRI) et les modèles structuraux. Les traitements des données des enquêtes internationales comme PISA utilisent majoritairement les MRI. Cependant, il est commode d'introduire les principaux concepts psychométriques à partir de l'approche classique.

La théorie classique des tests

On peut résumer le principe essentiel de la psychométrie par la formule de l'équation [1] :

Score observé = Score vrai + Erreur de mesure [1].

On cherche à distinguer performance (les résultats observés) et compétence (l'aptitude, le trait qui a produit cette performance et que l'on cherche à évaluer). L'étude de la fidélité interne est de s'assurer que le passage des items à la variable évaluée est fiable. Elle garantit que le score calculé à partir des items, en général en faisant la somme des points accordés pour des réponses correctes a une signification univoque. Ce qui ne serait pas le cas, par exemple, dans le cas où les items mesureraient des compétences différentes. C'est pourquoi on parle ici d'homogénéité ou de consistance interne. L'analyse interne se fait à deux niveaux : local, celui des items et global, celui du score. Au niveau des items, on s'intéresse principalement à deux de leurs caractéristiques : leur difficulté et leur discrimination.

Indice de difficulté de l'item

Dans le cas d'un score dichotomique (bonne ou mauvaise réponse), la difficulté de l'item est souvent estimée par la proportion d'élèves de l'échantillon qui donnent une réponse correcte à cet item. Le score moyen est une variante pour des items polytomiques (réponses multiples ordonnées). L'utilisation de cette proportion observée comme estimation de la difficulté peut être biaisée lorsque la représentativité de l'échantillon n'a pu être démontrée. À l'extrême dans le cas d'échantillons de compétence très faible ou très élevée, l'estimation de la difficulté des items peut conduire à des estimations très différentes. Cette dépendance entre l'estimation de la difficulté des items et l'estimation de la compétence des sujets a été la source de nombreuses réflexions visant à obtenir des estimations indépendantes. Les MRI ont été souvent présentés comme fournissant une solution à ce problème.

La gestion des non-réponses ou plutôt des réponses manquantes est un autre problème pour l'estimation de la difficulté des items et de la compétence des sujets. On peut identifier au moins trois types différents de

réponses manquantes : 1) les omissions intermédiaires ; 2) les omissions finales ; 3) les réponses manquantes structurelles. Les omissions intermédiaires ou finales correspondent à des items présents dans le protocole du sujet mais auxquels il n'a pas répondu. En général, on interprète les omissions intermédiaires, comme une déclaration d'ignorance et/ou une absence de prise de risque ; les omissions terminales comme un manque de temps. La distinction entre ces deux types de non-réponse est importante car elle renvoie à la distinction entre test de puissance ou de vitesse (la rapidité du sujet à accomplir la tache fait partie de la compétence évaluée). Le codage des réponses manquantes comme échecs ou comme items non examinés est donc fondamentale pour l'estimation de la difficulté des items. Le codage des omissions terminales comme des items non examinés ou comme des échecs aboutit à une estimation différente de la difficulté. Si la proportion de réussite est estimée à partir des seuls élèves de l'échantillon qui ont répondu à l'item, cela évite d'interpréter comme absence de maîtrise du domaine ce qui dépend en fait de la vitesse de travail et du temps de passation. Les omissions structurelles proviennent, elles, de l'organisation du plan de collecte des données. L'utilisation de la méthode dite des « cahiers tournants » dans PISA produit des données manquantes structurelles. Pour concilier deux exigences : recueillir de l'information sur de nombreux exercices sans trop augmenter le temps de passation, on va répartir les exercices (items) en plusieurs blocs de longueur (temps de passation) à peu près égale. Chaque sujet ne passera qu'un nombre de blocs correspondant au temps de passation choisi. Pour permettre de traiter les données, il faut que toutes les paires de blocs soit présentes dans le dispositif expérimental. Il s'agit alors de réduire le nombre de combinaisons des paires de cahiers pour maîtriser l'explosion combinatoire que pourrait engendrer la nécessité de construire toutes les combinaisons de paires de blocs. En général, on a choisi de construire des cahiers comprenant trois blocs pour s'appuyer sur une méthode de construction des plans expérimentaux bien connue : celle des triades. Pour neutraliser les effets liés à l'apprentissage et à la fatigabilité, on va contrôler l'ordre de passation des blocs en les contrebalançant. Chaque bloc apparaîtra au moins une fois dans les différentes positions de l'ordre de passation d'où le nom de « cahiers tournants » sous lequel ce dispositif expérimental est souvent désigné en français. Les protocoles contiennent donc des données manquantes structurelles et peuvent contenir des omissions finales et intermédiaires.

La recherche de solutions satisfaisantes pour la gestion de ces trois types réponses manquantes a été un des moteurs qui ont fait évoluer les méthodes employées pour traiter les données. Les réflexions ayant abouti à ces évolutions seront présentées dans le cadre des MRI.

Indice de discrimination de l'item

La discrimination de l'item renseigne sur la qualité et la quantité d'information apportées par l'item pour déterminer la compétence du sujet. Un item au pouvoir discriminant élevé apporte beaucoup d'information sur la compétence du sujet, un item peu discriminant renseigne peu sur la compétence du sujet. Leur pouvoir discriminant est un des principaux critères de sélection des items pour la construction définitive d'une épreuve. L'indice utilisé pour estimer le pouvoir discriminant de l'item se fonde sur la corrélation entre l'item et le critère évalué (en général le score au test). On fait l'hypothèse qu'un item est discriminant si les sujets qui le réussissent ont, en moyenne, un score plus élevé que les sujets qui y échouent. La prise en compte de l'indice de discrimination est importante pour s'assurer de la fiabilité des items de l'épreuve (suppression des items peu discriminants donc peu informatifs). Il faut souligner que le modèle de mesure retenu pour le traitement des données de PISA requiert que tous les items présentent une discrimination égale.

La consistance au niveau global

De la même manière qu'on s'est intéressé à la validité des items, on va étudier la fiabilité de l'épreuve au niveau global. On parle d'homogénéité ou de consistance interne. Dans la théorie classique des tests, celle-ci est estimée par le coefficient α de Cronbach (Cronbach et Meehl, 1955). Cet indicateur répond à la question « l'ensemble des items est-il suffisamment homogène pour que le calcul d'un score soit valide ? » La valeur de l'α dépend à la fois de l'homogénéité des items (appréciée à partir de leurs intercorrélations) et de leur nombre. À homogénéité donnée, on peut augmenter la consistance interne du test en augmentant sa longueur (Cortina, 1993). Ce point est important dans la mesure où les épreuves pour les évaluations internationales sont en général plutôt longues.

Le modèle de mesure classique repose, comme les autres modèles, sur

plusieurs conditions de validité. Les plus connues sont l'unidimensionnalité et l'indépendance conditionnelle des items et des sujets. Ces conditions seront davantage développées dans la présentation des MRI. On peut cependant signaler ici un problème posé par le format des épreuves de littératie par rapport à la condition d'indépendance conditionnelle. L'indépendance conditionnelle se traduit par l'hypothèse selon laquelle la réponse d'un sujet à un item ne dépend pas de ses réponses aux autres items de l'épreuve. La réussite d'un sujet à un item ne dépend que de sa compétence sur le trait latent mesuré par l'item et de rien d'autre (en particulier pas de ses réponses aux items qu'il a examinés avant celui ci). Il est souvent difficile de tester l'hypothèse d'indépendance conditionnelle.

On peut, par contre, identifier de nombreuses situations de *testing* où, par construction, la condition d'indépendance conditionnelle n'est pas respectée (Vrignaud, 2003). Ainsi, dans l'évaluation de la littératie, on demande souvent de répondre à plusieurs questions posées sur le même texte. Cette manière de procéder se justifie par le fait que l'investissement du sujet, tant cognitif que temporel, pour s'approprier des objets complexes, ici un texte, doit être rentabilisé au mieux. On utilise en anglais l'expression de *testlet* pour de tels exercices comprenant plusieurs items. En général, on ne tient pas compte des biais induits par cette dépendance dans le traitement des résultats des enquêtes internationales sur la littératie (Dickes et Vrignaud, 1995). Ces biais ont pourtant des effets non négligeables comme l'ont montré les quelques recherches réalisées sur les *testlets* (par exemple Wainer et Thissen, 1996). Les indicateurs psychométriques classiques tels que l'α de Cronbach sont biaisés dans le sens d'une surestimation.

La référence

On sait qu'un score brut à une épreuve n'est pas interprétable puisqu'il dépend de la difficulté des items intrinsèquement mêlée à la compétence de l'échantillon. En psychologie, on a privilégié l'utilisation d'une population de référence pour situer les performances des sujets. La compétence du sujet va être estimée faible, moyenne ou forte selon que sa performance se situe, respectivement, en dessous de, égale ou supérieure à la moyenne de la distribution de la population de référence. Plusieurs solutions peuvent être adoptées pour situer un score dans une distribution de référence : 1) le calcul d'une note standardisée en utilisant les paramètres (moyenne et écart type) de la distribution de référence – ce calcul s'accompagne souvent d'un changement d'échelle, l'exemple le plus connu est celui du QI, 2) le recours à un étalonnage, 3) la référence à un critère de maîtrise. En éducation, on a plutôt privilégié le recours à un critère traduisant la maîtrise du domaine évalué par l'épreuve. L'approche la plus simple consiste à calculer le pourcentage des items réussis par le sujet et à considérer qu'au-delà d'un seuil donné (en général 75 ou 80 %) le sujet maîtrise le programme évalué par l'épreuve. Cette façon de procéder peut inciter à des interprétations erronées. En effet, le fait que les scores à différents tests se trouvent ainsi standardisés laisse penser qu'ils sont comparables. Or, comme on l'a rappelé plus haut, la difficulté d'un item donc d'un test ne peut être appréciée qu'en relation avec la compétence de l'échantillon sur les résultats desquels cette difficulté a été estimée. Pour placer les résultats obtenus à différentes versions – ici linguistiques et/ou nationales – d'un même test il faut procéder à une opération dite de parallélisation pour placer les résultats sur une même échelle (pour une présentation de ces procédures, voir Kolen et Brennan, 1995). Dans le cadre des enquêtes internationales, la procédure de parallélisation est gérée par le modèle de mesure employé (les MRI).

Les modèles de réponse à l'item (MRI)

Présentation

Ces modèles regroupés sous l'appellation générique de *modèles de réponse à l'item* (MRI) – *Item Response Modeling* (IRM) en anglais[3] – ont été créés il y a une trentaine d'années (voir, pour une présentation, Hambleton et Swaminathan, 1985 ou, en français, Dickes *et al.*, 1994 ; Vrignaud, 1996). Il faut signaler qu'ils ont été « inventés » à peu près simultanément et de manière indépendante au Danemark par le mathématicien Georg Rasch (1960) qui cherchait un modèle permettant de comparer des compétences d'élèves en lecture à plusieurs années d'intervalle et, aux États-Unis, par le statisticien Allan

NOTE

3. En anglais, le terme d'*Item Response Theory* (IRT) est plus largement utilisé. Le terme de modèle paraît plus approprié dans la mesure où il s'agit de rendre compte du comportement du sujet répondant à un item plutôt que de construire une théorie psychologique du comportement du sujet, comme le font remarquer H. Goldstein et R. Wood (1989).

Birnbaum (1959, cité dans Birnbaum, 1968) qui cherchait à améliorer les modèles de mesure en psychométrie. Ces modèles ont profondément renouvelé l'approche psychométrique car d'une part ils offrent un cadre unitaire pour penser l'ensemble des concepts psychométriques (exposés plus haut à propos du modèle classique) et d'autre part, ils offrent un nouveau cadre d'interprétation des résultats aux tests en situant la performance des sujets par rapport à des tâches et non plus par rapport à la performance d'autres sujets. Ces modèles dont le principe est présenté dans l'équation [2] sont probabilistes. On postule que la probabilité qu'un sujet j donne une réponse correcte à un item i est fonction de la compétence (θ_j) du sujet et de la difficulté de l'item (di) :

$$Pr(X=x) = f(d_i, \theta_j) \quad [2]$$

Dans le cas d'items dichotomiques, X prend les valeurs échec [0] ou réussite [1], on obtient donc la probabilité d'un échec ou d'un succès.

Les modèles MRI sont basés sur la recherche d'un modèle mathématique du fonctionnement de l'item permettant de représenter la relation entre difficulté de l'item et compétence du sujet. On utilise en général la fonction logistique. Le modèle le plus général comprend trois paramètres pour modéliser le fonctionnement de l'item : « b_i » la difficulté de l'item « a_i » la pente (discrimination de l'item), « c_i » le paramètre de réponse « au hasard »[4].

On peut les rapprocher des paramètres classiques : « b_i », la difficulté de l'item de la fréquence de réussite ; « a_i », la pente (discrimination de l'item) de la corrélation item/score ; « c_i » de l'étude des distracteurs. Le paramètre de compétence « θ_j » est une estimation de la mesure vraie de la compétence du sujet (c'est-à-dire que les MRI permettent de séparer performance et compétence). L'explication de la compétence et de la difficulté de l'item par une même variable latente justifie explicitement la comparaison entre items et entre sujets. Les paramètres de difficulté vont permettre de comparer les items entre eux. Les paramètres de compétences autorisent la comparaison des sujets et des groupes de sujets. Toutes les opérations de construction de tests et d'interprétation des résultats demandant d'assurer l'équivalence des items et des tests ou la comparaison de différentes populations vont se trouver ainsi facilitées.

Combien de paramètres utiliser pour modéliser la compétence ?

La question du nombre de paramètres du modèle a été souvent discutée. Les options retenues ayant des conséquences sur les conditions de validité des statistiques et la présentation des résultats, ces choix ont un retentissement sur le traitement des enquêtes internationales. Ainsi, pour les traitements de l'enquête PISA, ACER *(Australian Council for Educational Research)*[5] utilise un modèle dérivé du modèle de Rasch implanté dans son logiciel CONQUEST, modèle qui ne comprend, pour expliquer le fonctionnement de l'item, que le paramètre de difficulté alors qu'ETS *(Educational Testing Service)* s'appuie sur un modèle à deux paramètres (difficulté et discrimination) en utilisant des algorithmes d'estimation implantés dans le logiciel BILOG (Zimowski, Muraki, Mislevy, et Bock, 1996) – voir pour un exemple les traitements de l'enquête IALS : Yamamoto, 1998). Cette différence de choix s'explique par au moins quatre raisons. D'abord des raisons historiques, les travaux sur les MRI s'étaient inscrits à ETS dans la suite des travaux de Birnbaum (1968) repris et enrichis par Lord (1980) qui avaient introduit d'emblée un modèle à deux paramètres alors que les travaux d'ACER s'inscrivaient dans le cadre de l'approche de Rasch comme le montrent les logiciels construits par cette organisation (Titan puis Quest : Adams et Khoo, 1994). Ensuite des raisons liées au format des items, PISA comprend des items polytomiques (les réponses peuvent faire l'objet d'un codage ordonné selon des niveaux de réussite). Ce format d'item est facile à traiter par le modèle de Rasch (on sépare le paramètre de difficulté en une partie représentant la difficulté générale de l'item et une autre partie représentant le passage d'un niveau

NOTES

4. L'anglais utilise le terme de *guessing* (traduit parfois en français par « pseudo-chance ») pour désigner, principalement dans les QCM, la probabilité de « deviner » la bonne réponse ou de la donner par hasard. On a jugé utile d'introduire ce paramètre dans les MRI pour rendre compte du fait que la probabilité de bonne réponse d'un sujet ne devient pas infiniment petite au fur et à la mesure que la compétence de ce sujet est estimée faible, mais peut rester dans une zone nettement plus élevée. Par exemple dans le cas d'un QCM comprenant quatre possibilités de réponse, la possibilité de donner la bonne réponse au hasard serait de 25 %. Dans ce cas, le paramètre de *guessing* estimerait la probabilité à ce seuil même pour des sujets de compétence faible.

5. ACER est l'organisation principale en charge du consortium qui a géré PISA, ETS a été l'organisation en charge du traitement des données des enquêtes américaines (NAEP, etc.) ainsi que de plusieurs enquêtes internationales, en particulier, IALS.

de difficulté à un autre) alors que l'estimation des paramètres de difficulté de tels items n'est pas aussi aisément accessible par le modèle à deux paramètres. Une troisième raison peut trouver son origine dans la détermination des niveaux de compétence dont le rationnel sera présenté plus loin. La procédure de classement des items en niveau de difficulté est plus cohérente si la discrimination des items est identique. L'existence de différences de discrimination entre items peut rendre ce classement moins univoque. Enfin, une des phases essentielles de l'étude de l'équivalence en fonction des différentes versions linguistiques et/ou nationales est l'identification des *fonctionnements différentiels des items,* en abrégé FDI (pour une présentation en français voir Vrignaud, 2002, ou Rocher, 2003 dans le cadre des enquêtes internationales). Le FDI est une différence de réussite à un item entre deux groupes de sujets comparables quant au construit mesuré par le test. Le FDI[6] peut porter sur chacune des caractéristiques de l'item : 1) sa difficulté ; 2) sa discrimination. Le recours à un modèle à un seul paramètre simplifie l'approche de cette question. En revanche, l'utilisation du modèle de Rasch nécessite une condition de validité supplémentaire : l'hypothèse d'égale discrimination des items. Cette condition est en général vérifiée *a posteriori* dans la mesure où les tests d'adéquation au modèle de Rasch permettent de retenir l'hypothèse que ce modèle rend bien compte des données sans qu'il soit besoin d'introduire un paramètre supplémentaire pour prendre en compte la discrimination.

Dans le cadre des MRI, l'estimation des valeurs des paramètres de difficulté se fait sous cette hypothèse d'indépendance conditionnelle. Si on ne peut pas retenir l'hypothèse d'indépendance conditionnelle, alors il faudrait introduire un paramètre spécifique représentant la dépendance conditionnelle entre ces deux items comme la probabilité particulière de réussite à ces deux items, leur interaction comme le suggérait le statisticien anglais Harvey Goldstein (Goldstein, 1980). Par exemple E. T. Bradlow, H. Wainer et H. L. Lang (1998) proposent un MRI incluant des paramètres représentant la dépendance locale et élaborent un algorithme permettant l'estimation de ces paramètres.

Évaluer la compétence dans le cadre des MRI

Les modèles MRI ont été présentés par leurs avocats comme renouvelant la théorie de la mesure. G. Rasch argumentait que l'estimation de la difficulté des items et de la compétence des sujets étaient indépendantes, ce qui fondait, selon lui, le concept d'objectivité spécifique (Rasch, 1977). Quels que soient les items passés par un sujet, on obtiendra une même estimation de sa compétence. Quels que soient les groupes de sujets auxquels l'item a été administré, on obtiendra une même estimation de sa difficulté. Cette idée a été souvent considérée comme peu « réaliste » et semble d'ailleurs ne pas avoir donné lieu à de nombreuses études comme on le constate dans un ouvrage de synthèse sur les développements du modèle de Rasch (Fischer et Molenaar, 1995).

Les MRI définissent la compétence du sujet comme sa probabilité de résoudre des items d'une difficulté donnée. La compétence se définit donc par rapport à des tâches et non par rapport à d'autres sujets. Le paramètre de compétence du sujet définit sa zone de compétence qui peut être mise en relation avec les paramètres de difficulté des items. La définition de la zone de compétence nécessite de décider du seuil de probabilité de réussite retenu pour considérer que le sujet maîtrise l'item. Peut-on considérer qu'un seuil supérieur à 50 % est signe que l'item peut être résolu par le sujet ou vaut-il mieux considérer que seul un seuil proche de 100 % peut refléter la réelle maîtrise par le sujet ? Par exemple dans les évaluations éducatives aux États-Unis, le seuil de 80 % est généralement retenu (Kirsch, 1995). Ce seuil a l'avantage de garantir une probabilité quasi certaine de réussite, mais sa sévérité peut être trompeuse quant aux réussites réelles des sujets. En effet, les probabilités sont fortes que les sujets réussissent d'autres items de difficulté plus grande que celle comprise dans leur zone de compétence. Un second problème est celui de la définition de la compétence en fonction du contenu des items. Dire qu'un sujet est capable de résoudre des

NOTE

6. Lorsque la différence de réussite à l'item est de même sens en faveur ou en défaveur du même groupe dans toutes les classes de sujets, le FDI est dit « uniforme ». Le FDI uniforme porte uniquement sur la difficulté de l'item. Il existe un écart en faveur du même groupe à tous les niveaux de compétence. Lorsque la différence de réussite change de sens selon le niveau de performance des sujets (par exemple la différence est en faveur d'un groupe pour les classes de performance faibles et en défaveur du même groupe pour les classes de performance élevée) on parle de FDI « croisé ». Le FDI croisé porte sur la discrimination de l'item si celui-ci est plus discriminant dans un groupe que dans l'autre. Si on se représente aisément la signification psychologique d'un FDI uniforme, celle d'un FDI croisé peut être plus délicate.

items d'une difficulté donnée renvoie à la définition opérationnelle de ces items. Cette définition peut paraître simple quand le contenu des items s'y prête : par exemple la complexité d'opérations arithmétiques, le nombre d'inférences à effectuer pour conduire un raisonnement. Néanmoins, ce type d'analyse apparaît souvent simplificatrice au regard des modèles de résolution proposés par la psychologie cognitive (Rémond, *à paraître*).

La construction de l'échelle de compétence dans les enquêtes utilisant les MRI est essentiellement basée sur les regroupements d'items à partir de leurs indices de difficulté. Ainsi, dans la plupart des enquêtes internationales on définit plusieurs niveaux (en général cinq) de compétences. L'interprétation de chacun de ces niveaux est ensuite enrichie par l'analyse cognitive des items classés dans ce niveau. Ce système de définition d'une compétence est essentiellement psychométrique même s'il reçoit un habillage de psychologie cognitive. Un tel système a été particulièrement développé par Kirsch et collaborateurs dans les enquêtes NAEP puis IALS et PISA (voir par exemple Kirsch, Jungeblut et Mosenthal, 1998). Cette approche présente deux inconvénients majeurs.

Le premier est d'être partiellement tautologique : cet item est facile puisqu'il est réussi par un grand nombre de sujets et qu'il correspond donc à des opérations de niveau faible.

Un second inconvénient est la difficulté de déterminer le niveau auquel appartient un item. En effet, on prend en compte le paramètre de difficulté, non pas en lui-même, mais en recherchant quel niveau de compétence est nécessaire pour maîtriser un item de ce niveau de difficulté. Un item sera donc classé dans la catégorie correspondant au niveau de compétence permettant d'avoir une probabilité (en général 75 ou 80 %) de le réussir. Mais les sujets qui ont un niveau de compétence inférieur ont encore une probabilité élevée de le réussir si leurs compétences sont proches de la coupure séparant les classes de niveau. La qualité de cette séparation peut être appréciée à partir du pouvoir discriminant des items. L'information donnée par ces niveaux apparaît donc relativement floue et imprécise dans la mesure où les coupures sont par nature arbitraires : le fait d'être classé dans un niveau de compétence ne veut en aucun cas dire que le sujet n'est pas capable de fonctionner à des niveaux de compétence plus élevés. L'interprétation des niveaux n'est pas toujours facile car certains niveaux possèdent parfois peu d'items (en général les niveaux supérieurs). Et, surtout, l'interprétation en termes de fonctionnement cognitif n'est pas fondée sur l'analyse des tâches et des processus mais apparaît plutôt comme un produit dérivé du modèle de mesure psychométrique.

Dans PISA, les différents niveaux de compétence ont été définis de telle manière que les sujets dont le paramètre de compétence a une valeur proche de la borne inférieure ont une probabilité de 50 % de réussir les items de ce niveau, et ceux dont le paramètre de compétence a une valeur proche de la borne supérieure, une probabilité de 80 % de réussir ces mêmes items. Par construction, il est donc certain qu'un sujet ne réussit pas uniquement tous les items correspondant à son niveau et a – au moins pour les sujets proches de la borne supérieure – une probabilité non négligeable de réussir ceux du niveau supérieur. Encore une fois, il ne s'agit pas de pointer les insuffisances de la méthode sans en voir les avantages, en premier lieu, ceux de définir la compétence en relation avec des tâches et non plus en relation avec d'autres sujets comme dans l'approche psychométrique classique. Il faut également souligner la prudence avec laquelle ces opérations ont été effectuées et la clarté avec laquelle elles sont exposées dans le *technical manual* (Turner, 2002). Mais, on ne peut passer sous silence le risque d'aboutir à une réification de la notion de niveaux de compétence qui, dans les représentations d'utilisateurs n'ayant pas eu accès à l'ensemble des sources techniques, peuvent paraître plus objectifs qu'ils ne le sont en réalité.

L'estimation des paramètres

La mise en œuvre de l'estimation des paramètres des MRI n'est pas une opération anodine (on trouvera une excellente présentation exhaustive de cette question dans l'ouvrage de Baker, 1992). L'appréciation de l'adéquation des modèles MRI se pose aux différentes étapes de l'estimation des paramètres de difficulté des items et de compétence des sujets. En amont, les modèles MRI reposent sur des conditions de validité nombreuses : unidimensionnalité, indépendance conditionnelle des items, et, pour le modèle de Rasch, égal pouvoir discriminant des items. Ces conditions sont parfois difficiles à tenir et à vérifier. Ainsi R. K. Hambleton, H. Swaminathan et H. J. Rogers (1991) recensent une vingtaine de procédures à mettre en œuvre pour s'assurer de la possibilité d'application du modèle aux données. On peut citer

également l'ensemble de travaux menés par l'équipe de Stout (Bolt et Stout, 1996 ; Shealy et Stout, 1993 ; Nandakumar, 1994) à l'Université de Chicago qui a permis de trouver des cadres conceptuels plus performants pour tester certaines hypothèses (unidimensionalité, indépendance conditionnelle, fonctionnement différentiel des items). On peut regretter que les travaux de cette équipe soient totalement absents des traitements des enquêtes internationales.

L'algorithme d'estimation utilisé dans PISA est issu des travaux du statisticien américain D. Rubin sur l'algorithme dit « EM »[7] *(Expectation-Maximization* ; Dempster, Laird et Rubin, 1977 ; Rubin, 1987 et 1991). Rubin a clarifié le concept de valeur manquante en identifiant trois types de situations. La distribution des valeurs manquantes peut être représentée par une distribution complètement aléatoire (MCAR, *Missing Completely At Random*). Par exemple dans le cas des enquêtes internationales, l'utilisation de la méthode dite des cahiers tournants, les réponses manquantes sont dites MCAR puisque les blocs qui n'ont pas été présentés à l'élève résultent d'une affection au hasard d'un cahier à chaque élève. Le second type de situation est celui où on peut faire l'hypothèse que la distribution des données manquantes peut être représentée par une distribution aléatoire (*Missing At Random*, MAR) mais peuvent dépendre des réponses des sujets à d'autres variables utilisées dans l'enquête. Enfin, le dernier cas dit *Missing Not At Random* ou *not ignorable* est celui où les données manquantes résultent d'un processus dépendant de la variable elle-même par exemple la non-réponse à une question sur le niveau de revenus est plus fréquente dans les classes de revenu élevé.

Cette réflexion sur les données manquantes a conduit Rubin à opérer un renversement de perspective concernant l'estimation de la compétence des sujets. Rubin considère que la valeur manquante fondamentale est la position du sujet sur la variable latente. En effet, la compétence n'est connue que conditionnellement aux réponses du sujet à un nombre réduit de questions : celles qui sont incluses dans le test qu'il a passé y compris dans le cas où il a répondu à toutes les questions du test. Dans le cadre des MRI, cette formulation a conduit à repenser l'algorithme d'estimation des paramètres en utilisant l'algorithme EM (Bock et Aitkin, 1981), procédure implantée dans les logiciels BILOG dédiés à l'estimation des paramètres des MRI (Mislevy et Bock, 1990 ; Zimowski *et al.*, 1996). R. J. Mislevy et ses collaborateurs (Mislevy, 1987 ; Sheehan et Mislevy, 1990 ; Mislevy *et al.*, 1992) ont perfectionné cette approche en introduisant dans l'algorithme d'estimation les données descriptives du contexte du sujet *(background variables)* afin de rendre l'estimation du paramètre de compétence des sujets plus robuste. Il s'agit d'estimer la compétence des sujets conditionnellement aux réponses qu'ils ont données aux items auxquels ils ont répondu (donc sans inclure les items manquant par construction des cahiers tournants et les omissions terminales) et conditionnellement aux variables décrivant le contexte socio-économique des sujets. Il faut préciser que le score de compétence du sujet est conceptuellement une valeur non observée et que son estimation renvoie non pas à un seul paramètre mais à une distribution. Conditionnellement aux réponses et aux caractéristiques de ce sujet, on infère avec une plus ou moins bonne garantie la distribution du paramètre de compétence d'un sujet ayant ces caractéristiques et ce patron de réponses aux items. On ne connaît pas la valeur vraie du paramètre de compétence mais sa distribution. Pour renforcer la robustesse de cette estimation, on va procéder à plusieurs tirages dans cette distribution de valeurs dites plausibles dont la moyenne sera une meilleure estimation de la compétence de ce sujet. On trouvera le détail de cette procédure dans le *technical manual* (Adams, 2002).

On peut faire plusieurs commentaires par rapport à cette approche. En premier lieu, il est certain qu'elle prend au sérieux et qu'elle pousse, de manière particulièrement élégante, à l'extrême les concepts théoriques de la psychométrie. Sur le plan théorique, il est également certain que ces procédures permettent d'assurer une estimation plus rapide

NOTE

7. L'algorithme EM estime selon la méthode du maximum de vraisemblance les paramètres de distributions expliquant un échantillon de données lorsqu'on est en présence de données manquantes, en complétant les données par une variable aléatoire rendant compte de la relation entre les données observées (les réponses aux items) et les données manquantes (ici les paramètres du MRI). Dans une première phase, on va calculer l'espérance de la vraisemblance *(expectation)* et dans une deuxième phase on va opérer une maximisation *(maximisation)* de l'espérance obtenue. Puis, on utilise les valeurs trouvées à l'étape de maximisation pour une nouvelle étape d'espérance. On répétera ce processus de manière itérative dont chaque phase augmente la vraisemblance jusqu'à ce qu'on atteigne un critère d'arrêt (en général un écart faible entre la vraisemblance à deux étapes consécutives).

(convergence accélérée) et plus robuste des paramètres de compétence des sujets. On a pu également montrer qu'elle permet une estimation plus fidèle des moyennes des pays dans le cas des enquêtes internationales. Les points forts de cet algorithme sont la source de ses points faibles : la distribution des paramètres dépendant de plus nombreuses informations, cela introduit de nouvelles sources de biais dans l'estimation (par exemple les caractéristiques des sujets). Il va falloir s'assurer de la fidélité de toutes les informations portant sur les caractéristiques des sujets et de leur équivalence dans les différents contextes nationaux. Elle multiplie également les conditions de validité. Enfin, *last but not least*, cette procédure d'estimation aboutit à un ensemble (cinq dans PISA) de valeurs plausibles pour chaque sujet. D'après les publications sur cette approche, la théorie réalise un apport majeur à la réflexion psychométrique et les procédures semblent donner des résultats robustes pour l'estimation des paramètres des MRI. Il est, d'ailleurs, à noter que cette procédure élaborée par les chercheurs d'ETS pour les enquêtes américaines de type NALS et YALS (en ajoutant des procédures spécifiques au logiciel BILOG MG) puis pour les enquêtes internationales (voir par exemple IALS : Yamamoto, 1998) a été ensuite implantée dans le logiciel Conquest édité par ACER (Wu, Adams et Wilson, 1997) lorsque ce groupe a été chargé du traitement des données PISA. Le recours à la distribution de valeurs plausibles est maintenant généralisé dans les enquêtes internationales (voir par exemple PIRLS : Gonzalez, 2001).

Le fait d'estimer la compétence d'un sujet par cinq valeurs plausibles et non un score unique a des implications importantes sur la manière de conduire les analyses. La dispersion de ces valeurs plausibles est aussi importante que leur moyenne. Toutes les analyses statistiques devraient donc être élaborées à partir des différentes valeurs plausibles et non d'une seule ou d'une agrégation de celles-ci. Par exemple, si on souhaite calculer la corrélation entre une variable de contexte (la PCS de l'élève) et la compétence, il faudra calculer cette corrélation pour chacune des cinq valeurs plausibles fournies pour chaque sujet puis réaliser une agrégation des cinq valeurs obtenues pour la corrélation. La dispersion des valeurs de l'indicateur devra être utilisée pour les tests de signification. On trouvera des descriptions des procédures permettant de réaliser cette agrégation dans les publications traitant des méthodes d'imputations multiples (voir par exemple Schafer et Graham, 2002 pour une revue récente). Il n'est pas certain que les chercheurs réalisant des analyses secondaires à partir des données de PISA aient complètement intégré l'importance d'utiliser ces procédures pour obtenir des estimations sans biais des indicateurs dans le cadre de leurs analyses. Ces éléments sont présentés très explicitement et très clairement dans le *technical manual* (Adams, 2002).

L'unidimensionnalité de la littératie : artefact ou réalité ?

Les MRI ont été l'objet de nombreuses critiques. La plus fondamentale porte sur leur réalisme pour représenter le fonctionnement des sujets répondant à des items. Ainsi, M. Reuchlin (1996) conteste le caractère continu du modèle qui présuppose qu'un sujet peut toujours réussir un item. La réponse à un item a un caractère discret. La réussite à un item difficile n'est pas peu probable pour un sujet peu compétent, elle est tout simplement impossible. Une contestation moins radicale porte sur certaines de leurs propriétés au premier rang desquelles l'unidimensionnalité.

L'unidimensionnalité de la variable latente laisse présupposer que les différences interindividuelles ne sont que des différences de puissance, que les différences de difficulté entre items ne sont que des différences quantitatives. On accrédite ainsi l'idée que quel que soit le niveau de compétence des sujets, ceux-ci mettent en œuvre des processus et des stratégies similaires pour répondre aux items. Cette critique a déjà été souvent portée à l'encontre des scores dont le caractère globalisant n'informe pas sur les processus sous-jacents (Huteau et Lautrey, 1999). Le nombre de variables à introduire dans un modèle pour rendre compte d'un ensemble de comportements est une question classique en psychologie.

La question centrale est la prise en compte de différentes dimensions et par conséquent de plusieurs compétences expliquant la performance des sujets aux items. Si l'on considère par exemple trois échelles, les relations entre leurs scores peuvent se situer entre deux situations extrêmes : 1) il n'existe aucune relation entre elles ; 2) la relation entre les dimensions est tellement élevée qu'il n'y a pas lieu de les distinguer : elles mesurent la même chose. Dans le cas n° 1, les dimensions sont orthogonales (les corrélations sont nulles), il faut présenter et interpréter les résultats de chacune des échelles séparément.

Dans le cas n° 2, les corrélations sont proches de 1, il n'y a pas lieu d'interpréter séparément les dimensions, les compétences mesurées sont complètement redondantes et si l'on devait les distinguer ce serait par un artefact sémantique qui consisterait à les nommer différemment. La plupart du temps, les données se situent entre ces deux pôles. La question est alors de décider à partir de quel seuil la liaison entre les dimensions peut être estimée comme suffisamment faible pour considérer que les dimensions mesurées correspondent à des compétences différentes ? Cette question a été au coeur de la plupart des débats autour des modèles psychologiques des aptitudes. La dimensionnalité des compétences en littératie s'inscrit dans un tel débat. On cherche à savoir si les résultats peuvent être présentés sur une ou plusieurs échelles. Cependant, la pertinence d'une discussion apparaît, dans le cas des enquêtes internationales sur la littératie, comme faussée car pour des raisons de fiabilité de la mesure, on s'attache au fait que les épreuves soient fortement unidimensionnelles. On a montré *supra* que cette condition est requise par le modèle de mesure employé : le MRI. L'unidimensionnalité est à la fois la structure recherchée et la condition de validité (l'hypothèse au sens de l'*assumption*) des MRI. En effet, les modèles de base des MRI nécessitent la condition d'unidimensionnalité : on doit rendre compte des relations entre items (estimés par leurs paramètres) et entre les sujets ainsi qu'entre items et sujets par une seule variable latente.

La solution retenue pour l'interprétation de PISA est de considérer cinq échelles : trois de littératie, une de mathématiques et une de science. On s'intéressera uniquement aux échelles de littératie. Ces trois échelles se distinguent selon les auteurs du dispositif par les opérations auxquelles elles font appel (sur ce point voir Rémond, *à paraître*) : 1) retrouver de l'information ; 2) développer une interprétation ; 3) réfléchir sur le contenu du texte. La distinction entre ces trois échelles et le rattachement des items à chacune d'elle a été fait à partir de jugements d'experts et des résultats de l'analyse des données. Les valeurs des corrélations entre échelles publiées pour les trois échelles de littératie dans PISA 2000 sont très élevées (> .89 ; *cf.* Adams et Caspersen, 2002) et dans bien des cas seraient considérées comme suffisantes pour rassembler les trois échelles en une seule. Ce qui est d'ailleurs le cas puisque certains résultats sont estimés sur une échelle globale qui est, elle-même, considérée par hypothèse comme unidimensionnelle puisqu'elle présente une bonne adéquation à un modèle de Rasch. On peut donc légitimement s'interroger sur le bien-fondé de distinguer trois échelles puisqu'un modèle comprenant une seule échelle rend parfaitement compte des données (selon les décisions prises par les statisticiens quant à l'adéquation du modèle de mesure aux données).

■■■

Ce tour d'horizon du modèle de mesure et de l'estimation des paramètres dans les enquêtes internationales en général et dans PISA en particulier fait ressortir plusieurs points. D'abord la sophistication des méthodes utilisées, le soin apporté à résoudre des problèmes délicats posés par l'évaluation psychométrique. Bien que tous ces éléments soient présentés dans le *technical manual* (Adams et Wu, 2002), on peut s'interroger sur la réalité de leur accessibilité à l'ensemble des utilisateurs potentiels de PISA dans la mesure où la psychométrie, du moins à ce niveau de complexité, ne fait pas forcément partie du socle commun de connaissances de l'ensemble de la communauté scientifique francophone des sciences humaines. Ceci peut conduire certains utilisateurs à des erreurs dans l'utilisation des données comme cela a été souligné à propos de la prise en compte des valeurs plausibles dans les analyses secondaires.

Un second point est que, malgré le soin apporté à ces questions méthodologiques, certaines solutions restent encore insatisfaisantes au regard de la sophistication du reste de l'édifice. On a signalé parmi les aspects les plus techniques la violation de la condition d'indépendance conditionnelle. La question de la dimensionnalité apparaît plus centrale et donc plus gênante dans la mesure où elle est en prise directe avec la présentation et l'interprétation des résultats. Ceci conduit à une interrogation plus générale sur la nature conceptuelle de la compétence évaluée. À ce sujet, il faut signaler que Harvey Goldstein et ses collaborateurs (Goldstein, 2004 ; Goldstein *et al.*, soumis) ont montré, en appliquant les modèles d'équations structurales aux données anglaises et françaises de PISA qu'elles n'étaient pas unidimensionnelles, mais à tout le moins bidimensionnelles. L'écart à l'unidimensionnalité est révélateur de failles dans le dispositif de mesure et ses conséquences sur la définition de la compétence doivent être prises en considération.

Il est certain que cette compétence est bien une compétence

largement transversale dont la plus ou moins grande maîtrise peut être considérée comme le produit des systèmes éducatifs. Mais une telle variable, relativement décontextualisée puisqu'elle ne doit pas être sensible aux différents contextes linguistiques et culturels, n'est-elle pas une sorte de facteur général de réussite protéiforme susceptible de recevoir de multiples dénominations et interprétations ? Le résultat d'une étude conduite sur la comparaison entre une enquête précédente sur la littératie auprès d'adultes IALS et PISA conduit également à s'interroger sur la nature des échelles de PISA. L'enquête IALS comprenait trois échelles définies d'après le contenu du support (prose, document et littératie quantitative). Plusieurs items (15) de l'échelle « Prose » de IALS ont été intégrés dans PISA. Il était donc possible de comparer les deux types d'approches de la littératie celle de IALS et celle de PISA. Cette étude comparative a été conduite par Yamamoto (2002). Malgré les nombreux biais conduisant à rendre difficile la comparaison entre les deux échelles, Kentaro Yamamoto aboutit à la conclusion que la corrélation entre l'échelle de *prose literacy* de IALS et de PISA est de .83. Ce qui correspond à peu près à l'ordre de grandeur des corrélations entre les sous-échelles de IALS. On peut en conclure que ces deux enquêtes bien que constituées de sous échelles interprétées différemment mesurent globalement la même compétence.

On peut également s'interroger sur le fait que ce facteur peut s'apparenter dans une large mesure à des variables du type des aptitudes intellectuelles, en particulier, du raisonnement verbal. Dans une autre enquête menée dans le cadre d'un projet européen (Vrignaud, 2001), on observe une corrélation proche de « .70 » entre un test de vocabulaire (subtest de vocabulaire du WISC III) et des épreuves nationales d'évaluation de la lecture pour deux pays (l'Angleterre et l'Italie). Bien que l'intensité de ces corrélations ne soit pas suffisamment élevée pour assimiler les compétences évaluées par les deux types de tests, elle est néanmoins suffisamment élevée pour faire l'hypothèse qu'une partie relativement importante (près de la moitié de la variance) est expliquée par un test de vocabulaire. Les tests de vocabulaire sont les meilleurs indicateurs du raisonnement verbal et même du raisonnement en général (corrélation élevée avec la mesure globale du QI). Ces tests de lecture mesurent donc également une compétence verbale très générale. On pourrait s'interroger, au moins pour les niveaux supérieurs de PISA qui, selon leur définition, requièrent que les sujets réalisent des opérations d'inférence, sur le fait qu'on mesure autant la capacité au raisonnement verbal que la capacité à tirer de l'information d'un texte écrit.

La seconde question porte sur l'unidimensionnalité du construit mesuré. Le recours à trois dimensions, même s'il est intéressant d'un point de vue conceptuel, n'apparaît pas pleinement convainquant du point de vue psychométrique. L'agrégation de l'ensemble des items dans une seule variable latente est un point qui ne plaide pas particulièrement en faveur de l'utilisation de plusieurs sous échelles. Les contraintes du modèle de mesure sont telles qu'elles conduisent à éliminer toutes les causes éventuelles d'écart à l'unidimensionnalité qui seraient en violation avec l'utilisation des MRI. On peut considérer que cette réduction va s'opérer dès la sélection des items. Par conséquent, l'univers des items risque d'éliminer des informations porteuses de différences qualitatives supportant d'autres dimensions et non plus seulement des différences quantitatives consistant à ordonner les moyennes des pays sur un axe. On peut également s'interroger sur la pertinence d'expliquer les différences entre sujets de manière uniquement quantitative pour les sujets faiblement compétents dont la situation est mieux qualifiée par le terme d'illettrisme que par celui de niveau faible de littératie. Il est plus heuristique de chercher à qualifier ces situations d'illettrisme en identifiant leurs causes plutôt que de les quantifier. L'enquête sur les compétences en littératie des adultes français « Information et Vie Quotidienne » (Murat, 2005) comportait un module particulier pour les sujets identifiés comme étant en situation d'illettrisme visant à diagnostiquer les causes de cet illettrisme.

Le choix fait par des enquêtes de type PISA d'évaluer des compétences n'est pas exempt de tout questionnement scientifique et idéologique. En effet, on se souvient des débats sur la mesure de l'intelligence et de la boutade de Binet. On court ici le risque de déclarer « la compétence ? c'est ce que mesure notre test ! ». Comment être sûr que l'on échantillonne les items (les tâches) de manière à réellement balayer le domaine ? Ne court-on pas le risque comme dans les tests d'intelligence de sur-représenter voire de ne représenter que les tâches en relation avec les apprentissages scolaires et le milieu culturel dominant tels qu'ils sont conçus et valorisés dans certains pays et d'assister aux terribles dérives apparues dans le domaine des aptitudes avec les travaux de Terman comme l'évoquent A. Blum

et F. Guérin-Pace (2000) ? Il y a un risque de dérive idéologique à considérer ces compétences comme dotées d'une réalité autonome et objective alors qu'elles sont étroitement dépendantes d'un modèle de mesure.

Si l'on choisit une approche des compétences, alors, il est nécessaire de définir les compétences en termes de domaines, opération qui seule pourra valider l'interprétation de la mesure psychométrique puisqu'elle permettra de vérifier la couverture du domaine de la compétence par les épreuves construites. Cette approche a été l'objet d'une enquête internationale pilotée par l'OCDE : le programme DESECO (1999). Il s'agissait de demander à différents experts : philosophes (Canto-Sperber et Dupuy, 1999), ethnologue (Goody, 1999), psychologue (Haste, 1999), économistes (Levy et Murnane, 1999), spécialistes des sciences de l'éducation (Perrenoud, 1999) comment on pourrait définir les compétences nécessaires pour vivre et réussir dans le monde moderne. Ce type de travaux pourrait permettre de définir les compétences évaluées sur des bases théoriques et non uniquement psychométriques. La validité du construit et son interprétation s'en trouveraient davantage validées. Il ne semble pas malheureusement que les résultats de DESECO aient été injectés dans les réflexions sur les enquêtes internationales d'évaluation des compétences.

À LIRE

Adams R. J. (2002). « Scaling PISA cognitive data ». *In* M. L. Wu et R. J. Adams (éd.), *PISA 2000 : Technical Report.* Paris : OECD, pp. 99-108.

Adams R. J. et Carstensen C. (2002). « Scaling outcomes ». *In* M. L. Wu et R. J. Adams (éd.), *PISA 2000 : Technical Report.* Paris : OECD, pp. 149-162.

Adams R. J. et Khoo S. J. (1994). *QUEST : The Interactive Test Analysis System Version 2.0.* Hawtorn : ACER.

Adams R. J. et Wu M. L. (2002). *PISA : Technical report.* Paris : OECD.

Baker F. B. (1992). *Item Response Theory : parameter, estimation techniques.* New York : M. Dekker.

Beaton A. E. et Johnson E. G. (1992). « Overview of the scaling methodology used in the national assessment ». *Journal of Educational Measurement*, vol. 29, pp. 163-175.

Blum A. et Guérin-Pace F. (2000). *Des lettres et des chiffres.* Paris : Fayard.

Bock R. D. et Aitkin M. (1994). « Marginal maximum likehood of item parameters : Application of an EM algorithm ». *Psychometrika*, vol. 46, pp. 443-459.

Bolt D. et Stout W. (1996). « Differential item functioning : Its multidimensional model and resulting subtest detection procedure ». *Behaviormetrika*, vol. 23, pp. 67-95.

Bonora D. et Vrignaud P. (1997). *Évolution des connaissances scolaires et modèles de réponse à l'item.* Rapport pour le ministre de l'Éducation nationale. Paris : ministère de l'Éducation nationale.

Bottani N. et Vrignaud P. (2005). *La France et les évaluations internationales.* Rapport établi à la demande du Haut conseil de l'évaluation de l'école. Paris : Haut conseil de l'évaluation de l'école. Disponible sur au format PDF sur Internet à l'adresse : http://cisad.adc.education.fr/hcee/documents/rapport_Bottani_Vrignaud.pdf (consulté le 8 janvier 2007).

Bradlow E. T. ; Wainer H. et Wang H. (1998). « A bayesian random effects model for testlets ». *In ETS Research Report. RR-98-3.* Princeton : Educational Testing Service.

Canto-Sperber M. et Dupuy J.-P. (2001). « Competencies for the good life and the good society ». *In* D. S. Rychen et L. H. Salganik (éd.), *Defining and Selecting Key Competencies.* Göttingen : Hogrefe et Huber, pp. 67-92.

Cortina J. M. (1993). « What is coefficient alpha : An examination of theory and application ». *Journal of Applied Psychology*, vol. 78, pp. 98-104.

Cronbach L. J. et Meehl P. E. (1955). « Construct validity in psychological tests ». *Psychological Bulletin*, vol. 52, pp. 281-302.

Dempster A. P. ; Laird N. M. et Rubin D. B. (1977). « Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion) ». *Journal of the Royal Statistical Society, Series B*, vol. 39, pp. 1-38.

Dickes P. et Flieller A. (1997). *Analyse secondaire des données françaises de la première enquête internationale sur la littératie des adultes (enquête IALS).* Rapport pour le ministère de l'Éducation nationale. Paris : ministère de l'Éducation nationale ; Nancy : université de Nancy II : Laboratoire de psychologie : équipe GRAPCO.

À LIRE

Dickes P. et Vrignaud P. (1995). *Rapport sur les traitements des données françaises de l'enquête internationale sur la littératie*. Rapport pour le ministère de l'Éducation nationale. Paris : ministère de l'Éducation nationale.

Dickes P. ; Tournois J. ; Flieller A. et Kop J.-L. (1994). *Psychométrie*. Paris : PUF.

Fischer G. H. et Molenaar I. W. [éd.] (1995). *Rasch models : Foundations, recent developments, and applications.* New York : Springer.

Gardner H. (1983). *Frames of mind*. New-York : Basic Books; trad. fr. *Les formes de l'intelligence*. Paris : O. Jacob, 1997.

Goldstein H. (1980). « Dimensionality, bias, independance and measurement scale problems in latent trait test score models ». *The British Journal of Mathematical and Statistical Psychology*, vol. 33, pp. 234-246.

Goldstein H. (2004). « International comparisons of student attainment : some issues arising from the PISA study ». *Assessment in Education*, vol. 11, pp. 319-330.

Goldstein H. et Wood R. (1989). « Five decades of item response modelling ». *The British Journal of Mathematical and Statistical Psychology*, vol. 42, pp. 139-167.

Goldstein H. ; Bonnet G. et Rocher T. *(to appear)*. « Multilevel multidimensionnal structural equation models for the analysis of comparative data on Educational performance ». *Journal of Educational and Behavioural Statistics*.

Gonzalez E. J. (2003). « Scaling the PIRLS reading assessment data ». *In* I. V. S. Mullis, M. O. Martin, E. Gonzalez et A. Kennedy, *PIRLS 2001 International Report*. Boston : International Study Center.

Goody J. (2001). « Competencies and Education : Contextual Diversity ». *In* D. S. Rychen et L. H. Salganik (éd.), *Defining and Selecting Key Competencies*. Göttingen : Hogrefe et Huber, pp. 175-190.

Hambleton R. K. et Swaminathan H. (1985). *Item Response Theory. Principles and applications*. Boston : Kluwer-Nijhoff.

Hambleton R. K. ; Swaminathan H. et Rogers H. J. (1991). *Fundamentals of item response theory*. Newbury Park : Sage.

Haste H. (2001). « Ambiguity, Autonomy and Agency ». *In* D. S. Rychen et L. H. Salganik (éd.), *Defining and Selecting Key Competencies*. Göttingen : Hogrefe et Huber, pp. 93-120.

Huteau M. et Lautrey J. (1999). *Évaluer l'intelligence : psychométrie cognitive*. Paris : PUF.

Johnson E. G. (1992). « The design of the National Assessment of Educational Progress ». *Journal of Educational Measurement*, vol. 29, pp. 95-110.

Kirsch I. S. ; Jungeblut A. et Mosenthal P. B. (1998). « The measurement of adult literacy ». In T. S. Murray, I. S. Kirsch et L. B. Jenkins (éd.), *Adult Literacy in OECD countries. Technical report on the first international adult literacy survey.* Washington [D. C.] : US Department of Education : National Center for Education Statistics, pp. 105-134.

Kolen, M.J., et Brennan, R.L. (1995). *Test Equating. Methods and practices*. New York : Springer

Laveault D. et Grégoire J. (2002). *Introduction aux théories des tests en sciences humaines*. Bruxelles : De Boeck.

Levy F. et Murnane R. J. (2001). « Key Competencies Critical to Economic Success ». In D. S. Rychen et L. H. Salganik (éd.), *Defining and Selecting Key Competencies*. Göttingen : Hogrefe et Huber, pp. 151-174.

Lord F. et Novick M. R. [éd.] (1968). *Statistical theories of mental test scores*. Reading : Addison-Wesley.

Mislevy R. J. (1994). « Evidence and inference in educational assessment ». *Psychometrika*, vol. 59, pp. 439-483.

Mislevy, R. J. (1987). « Exploiting auxiliary information about examinees in the estimation of item parameters. Applied Psychological Measurement », vol. 11, pp. 81-91.

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

Mislevy R. J. et Bock R. D. (1990). *BILOG 3 : Item analysis and test scoring with binary logistic models*. Mooresville : Scientific Software [2de éd.]

À LIRE

Murat F. (2005). « Les compétences des adultes à l'écrit, en calcul et en compréhension orale ». *INSEE Première*, n°1044, numéro complet.

Murray T. S. ; Kirsch I. S. et Jenkins L. B. [éd.] (1998). *Adult Literacy in OECD countries. Technical report on the first international adult literacy survey.* Washington [D. C.] : US Department of Education : National Center for Education Statistic, pp. 105-134.

Nandakumar R. (1994). « Assessing dimensionality of a set of item responses-Comparison of different approaches ». *Journal of Educational Measurement*, vol. 31, pp. 17-35.

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

OCDE (2002). *Définitions et sélections des compétences (DESECO) : fondements théoriques et conceptuels : document de stratégie.* Neuchâtel : OFS. Document disponible au format PDF sur Internet à l'adresse : http://www.portal-stat.admin.ch/deseco/deseco_doc_strategique.pdf (consulté le 8 janvier 2007).

Perrenoud P. (2001). « The key to social fields : Competences of an Autonomous Actor ». *In* D. S. Rychen et L. H. Salganik (éd.), *Defining and Selecting Key Competencies.* Göttingen : Hogrefe et Huber, pp. 121-150.

Rasch G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhague : Nielsen et Lydiche.

Rasch G. (1977). « On specific objectivity. An attempt at formalizing the request for generality and validity of scientific statements ». *Danish Yearbook of Philosophy,* vol. 14, pp. 58-94.

Rémond M. *(à paraître).* « Éclairages des évaluations internationales PIRLS et PISA sur les élèves français ». *Revue Française de Pédagogie.*

Reuchlin M. (1996). *Psychologie différentielle.* Paris : PUF [n^lle éd. aug.]

Rocher T. (2003). La méthodologie des évaluations internationales de compétences. *Psychologie et Psychométrie, Numéro spécial : Mesure et Education,* vol. 24, pp. 117-146.

Rubin D. B. (1991). « EM and beyond ». *Psychometrika*, vol. 56, pp. 241-254.

Rubin D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York : Wiley.

Rychen D. et Salganik L. [éd.] (2003). *Key Competencies for a Successful Life and a Well-Functioning Society.* Göttingen : Hogrefe et Huber.

Salganik L. H. ; Rychen D. S. ; Moser U. et Konstant J. W. (1999). *Projects on competencies in the OECD context : Analysis of theoretical and conceptual foundations.* Neuchâtel : Office fédéral des statistiques.

Salganik L. et Rychen D. [éd.] (2001). *Defining and Selecting Key Competencies.* Seattle : Hogrefe et Huber.

Salines M. et Vrignaud P. (2001). *Apprécier et certifier les acquis des élèves en fin de collège : diplôme et évaluations-bilans.* Rapport établi à la demande du Haut conseil pour l'évaluation de l'école. Paris : Haut conseil de l'évaluation de l'école.

Schafer J. L. et Graham J. W. (2002). « Missing data : Our view of the state of the art ». *Psychological Methods*, vol. 7, pp. 147-177.

Schafer J. L. et Olsen M. K. (1998). « Multiple imputation for multivariate missing-data problems : A data analyst's perspective ». *Multivariate Behavioral Research*, vol. 33, pp. 545-571.

Shealy R. T. et Stout W. F. (1993). « A model-based standardization approach that separates true bias/DIF from group ability differences and detect test bias/DTF as well as item bias/DIF ». *Psychometrika*, vol. 58, pp. 159-194.

Sheehan K. et Mislevy R. J. (1990). « Integrating cognitive and psychometric models to measure document literacy ». *Journal of Educational Measurement,* vol. 27, pp. 255-272.

Turner R. (2002). « Constructing the proficiency scales ». *In* M. L. Wu et R. J. Adams (éd.), PISA 2000 : Technical Report. Paris : OECD, pp. 195-216.

À LIRE

Van der Linden W. J. et Hambleton R. K. [éd.] (1997). *Handbook of modern item response theory.* New York : Springer.

Vrignaud P. (1996). « Les tests au XXIe siècle : que peut-on attendre des évolutions méthodologiques et technologiques dans le domaine de l'évaluation psychologique des personnes ? » *Pratiques psychologiques*, vol. 2, pp. 5-28.

Vrignaud P. (2001). « Features of results : comparing data from the different countries ». *In* G. Bonnet, N. Braxmeyer, S. Horner, H. P. Lappalainen, J. Levasseur, E. Nardi, M. Rémond, P. Vrignaud et J. White. *The use of national reading tests for international comparisons : ways of overcoming cultural bias.* A European Project. Socrates contract n° 98-01-3PE-0414-00. Ministère de l'éducation nationale. DPD Edition diffusion. Paris, pp. 81-101.

Vrignaud P. (2002). « Les biais de mesure : savoir les identifier pour y remédier ». *Bulletin de psychologie*, vol. 55, n° 6, pp. 625-634.

Vrignaud P. (2003). « Objectivité et authenticité dans l'évaluation : avantages et inconvénients des questions à choix multiples [QCM] et des questions à réponses complexes [QRC] : importance du format de réponse pour l'évaluation des compétences verbales ». *Psychologie et psychométrie*, vol. 24, n° 2-3, pp. 147-188.

Wainer H. et Thissen D. (1996). « How is reliability related to the quality of test scores ? What is the effect of local dependence on reliability ? », *Educational Measurement : Issues and Practice*, vol. 15, pp. 22-29.

Weinert F. E. (1999). *Concepts of competence.* Neuchâtel : Office fédéral des statistiques.

Wu M. L. ; Adams R. J. et Wilson M. R. (1997). *ConQuest. Generalized item response modelling software.* Hawthorn [Australia] : ACER.

Zimowski M. F. ; Muraki E. ; Mislevy R. J. et Bock R. D. (1996). *BILOG-MG. Multiple-Group IRT analysis and test maintenance for binary items.* Chicago : Scientific Software International.