

# Le caractère relatif des indicateurs de tendance

**Christian Monseur et Dominique Lafontaine**

*Depuis le début des années nonante, en réponse à une demande de plus en plus forte des responsables des politiques éducatives, les surveys internationaux en sciences de l'éducation ont adjoint aux objectifs poursuivis traditionnellement la publication d'indicateurs de tendance, indicateurs destinés à suivre l'évolution des performances de populations scolaires. En témoigne le nom donné aux études contemporaines: Programme International de Suivi des Acquis des élèves (PISA), Progress International Reading Literacy Study (PIRLS), Trends International Mathematics and Science Study (TIMSS).*

*Alors qu'une différence de 10 points sur une échelle dont l'écart-type est généralement fixé par convention à 100 ne change pas l'interprétation pédagogique des résultats pour une collecte de données, une telle différence prend une toute autre dimension pédagogique et politique si elle représente le changement de performance entre deux collectes de données.*

*Pour être valable, cette comparaison diachronique suppose que la méthodologie, au sens large du terme, soit parfaitement identique pour les deux collectes de données, ou que les éventuels changements méthodologiques n'affectent en rien la comparabilité temporelle.*

*A travers quelques exemples issus des données PISA pour l'essentiel, ce présupposé sera confronté à la réalité empirique; les résultats montrent la nécessité d'une grande prudence à l'égard de ces indicateurs de tendance.*

## Introduction

Sous l'impulsion de l'Association Internationale pour l'Évaluation du rendement scolaire (I.E.A.), la recherche comparative en sciences de l'éducation s'est dotée au début des années soixante d'une nouvelle méthodologie: les enquêtes à large échelle. «Learning from the experience of other nations was seen as a way of developing the necessary understanding of education and its socio-political implications» (Purves, 1989). Selon Husen (1979), «education may be perceived as a

global laboratory where similar human and societal goals are set in terms of different environments and pursued through different means and educational solutions».

A cette époque, l'I.E.A s'est institutionnalisée sous forme d' «une association internationale à but scientifique ayant principalement pour but:

- (i) d'entreprendre des recherches pédagogiques à l'échelle internationale;
- (ii) de promouvoir la recherche relative aux problèmes pédagogiques communs à plusieurs pays dans le but de recueillir des informations destinées à améliorer ultérieurement les systèmes éducatifs;
- (iii) de fournir, dans le cadre de l'association, les moyens permettant aux centres de recherche membres de l'association, de collaborer à des tâches communes ».

A ce jour, l'I.E.A. a conduit plus d'une vingtaine d'enquêtes internationales principalement en langue maternelle, en mathématiques et en sciences.

A la fin des années nonante, l'arrivée de l'Organisation de Coopération et de Développement Economiques (OECD) sur le marché des surveys internationaux en éducation a bouleversé le paysage des enquêtes comparatives.

Comme l'indique le tableau 1, cette nouvelle décennie se caractérise par une explosion de l'offre, explosion par ailleurs amorcée lors des années nonante. Plus d'une dizaine d'enquêtes auront lieu durant cette décennie.

Tableau 1: Liste des surveys internationaux par discipline

	60	70	80	90	2000
Langue maternelle		Reading, Literature	Composition	RLS	RLS_R PIRLS 2001, PIRLS 2006 PISA2000 PISA2000 PISA2003 PISA2006 PISA2009
Math	FIMS		SIMS IAEP I	IEAP II TIMSS, TIMSS_R	PISA2000 PISA2003 PISA2006 PISA2009 TIMSS 2003, TIMSS 2007 TIMSS Ad. 08
Science		FISS	SISS IEAP I	IAEP II	TIMSS, TIMSS_R PISA2000 PISA2003 PISA2006 PISA2009 TIMSS 2003, TIMSS 2007 TIMSS Ad. 08
Autres		Anglais, Français comme langue étrangère Education civique	Comped	Civic PPP Sites	Sites Teds

Cet accroissement de l'offre résulte de la participation de plus en plus importante des décideurs politiques dans la planification, la gestion et la publication des résultats de ces enquêtes. Comme l'indique Bottani (2004), l'influence du programme OECD/PISA sur les recherches internationales en éducation présente trois composantes:

- (i) reconnaissance de la présence des gouvernements, qui sont à présent de réels partenaires dans la mise en œuvre de ces programmes;
- (ii) renforcement de la façon dont ces programmes sont subventionnés;
- (iii) collecte de données récurrentes avec l'introduction d'une perspective diachronique dans la planification de ces études.

Depuis son avènement, le programme OECD/PISA est conçu pour produire trois types d'indicateurs (OECD, 1999):

- (i) «basis indicators providing a baseline profile of knowledge and skills of students;
- (ii) contextual indicators, showing how much skills relate to important demographic, social, economic and educational variables;
- (iii) indicators of trends that emerge from the on-going, cyclical nature of the data collected and that will show changes in outcome levels, changes in outcome distributions and changes in relationship between student-level and school-level background variables and outcomes over time».

Ce troisième type d'indicateurs, dénommé indicateurs de tendance, s'accorde pleinement avec les deux objectifs majoritairement partagés dans la plupart des pays industrialisés, à savoir l'élévation de la performance moyenne des populations scolaires et la réduction des inéquités, voire des inégalités.

Pour assurer la validité de ces indicateurs de tendance, l'OECD, en concertation avec les décideurs politiques et les responsables scientifiques, a conçu une méthodologie qui minimise les changements d'une collecte de données à l'autre: continuité dans (i) la définition de la population cible et des procédures d'échantillonnage, (ii) le(s) cadre(s) de l'évaluation et (iii) les conditions d'administration des épreuves pour ne citer que les principales composantes méthodologiques.

Néanmoins, entre deux collectes de données, des changements méthodologiques considérés comme mineurs et sans conséquences sur la validité des indicateurs de tendance, sont introduits. Certains de ces changements étaient planifiés dans la conception initiale du projet: à cet égard, citons l'alternance entre domaine majeur et domaines mineurs. D'autres résultent soit d'une volonté politique d'élargir les finalités assignées – adjonction d'un domaine mineur en 2003, insertion d'items d'attitudes dans les épreuves de rendement en 2006 –, soit d'un besoin méthodologique pour accroître la validité de l'étude – changement du test design en 2003 pour mieux contrôler notamment les effets de fatigue observés en 2000 (Adams & Wu, 2002).

Quels sont les effets de ces changements sur la validité des indicateurs de tendance et plus globalement quels facteurs peuvent les altérer, voire les biaiser ? Différents exemples permettront de démontrer le caractère relatif des indicateurs. En soi, ce caractère relatif ne présente pas de dangers particuliers. Cependant, la perspective temporelle ou de suivi assignée depuis peu aux enquêtes internationales fige ces indicateurs, leur attribuant en quelque sorte une valeur absolue, échappant à l'influence du contexte méthodologique de la collecte de données qui a permis leur calcul. Il peut dès lors s'ensuire des surinterprétations qui pourraient aboutir à des réformes pédagogiques que les faits empiriques ne suffisent pas à fonder.

Les exemples qui suivent sont issus d'analyses secondaires des bases de données PISA et TIMSS essentiellement. Les données chiffrées seront fournies pour la Suisse et quelques pays limitrophes, l'Autriche, l'Allemagne, la France, la Communauté française de Belgique, lorsque ceux-ci sont disponibles.

Ces exemples peuvent s'articuler selon deux axes:

- (i) les indicateurs de rendement ou les indicateurs d'équité,
- (ii) les problèmes d'échantillonnage ou problèmes de modèles de mesure.

*Tableau 2: Structuration des exemples selon deux axes*

	Indicateurs de rendement	Indicateurs d'équité
Echantillonnage	<ul style="list-style-type: none"> <li>• Participation différentielle des élèves</li> </ul>	<ul style="list-style-type: none"> <li>• Définition de l'école comme unité d'échantillonnage</li> </ul>
Modèles de mesure	<ul style="list-style-type: none"> <li>• Sélection des items d'ancrage</li> </ul>	<ul style="list-style-type: none"> <li>• Différence de rendement entre les filles et les garçons</li> <li>• Coefficient de corrélation intraclasse</li> </ul>

Dans le cadre de cet article, le premier axe, à savoir la distinction entre indicateurs de rendement et indicateurs d'équité, a été privilégié. Par ailleurs, les exemples ont été retenus en fonction de leur importance en matière de politiques éducatives.

## Le caractère relatif des indicateurs de performance Participation différentielle des élèves

Les enquêtes à large échelle sont souvent confrontées à des problèmes de non-réponse et malheureusement, les surveys internationaux en sciences de l'éducation n'échappent à la règle (Winglee, Kalton, Rust & Kasprzyk, 2001). Dans le cadre des évaluations de la performance scolaire des élèves, la non-réponse peut se produire à trois niveaux distincts:

- (i) une école échantillonnée refuse de participer;
- (ii) un élève échantillonné au sein d'une école participante ne peut compléter les

- tests de rendement et/ou les questionnaires contextuels pour des raisons de santé, d'absentéisme, de connaissance de la langue des outils d'évaluation, voire tout simplement parce qu'il refuse de participer;
- (iii) un élève participant refuse de répondre à une ou plusieurs questions.

Toute forme de non-réponse est susceptible d'introduire un biais dans les résultats selon (i) l'importance de cette non-réponse, (ii) le caractère aléatoire ou non de cette non-réponse, c'est-à-dire selon le lien qui existe entre la non-réponse et les variables mesurées par l'enquête (Groves, Dilman, Eltinge & Little. 2002). Intéressons-nous à la non-réponse des élèves pour concrétiser ces composantes. On peut supposer que l'absentéisme des élèves pour des raisons de santé touche indifféremment l'ensemble de la population scolaire. En effet, un garçon n'a pas plus de chances de souffrir d'un état grippal qu'une fille, un bon élève pas plus qu'un mauvais élève. Dans ce cas particulier, l'absentéisme pour des raisons de santé peut être considéré comme une variable aléatoire non corrélée avec les mesures de rendement scolaire. En d'autres termes, cette non-réponse ne devrait pas introduire de biais dans les résultats. Par contre, plusieurs études ont démontré le lien entre absentéisme non justifié et performance scolaire: les élèves les plus faibles ou les élèves dits en décrochage scolaire sèchent plus souvent les cours que leurs condisciples. Cette forme d'absentéisme corrèle donc avec les mesures de l'enquête et selon l'importance de cette non-réponse, engendrera un biais plus ou moins important.

Pour contrecarrer les effets de ces différents types de non-réponses, les surveys internationaux ont recours à trois grands types de stratégies:

- (i) imposer des minima de participation des écoles et des élèves;
- (ii) les écoles qui refusent de participer peuvent être remplacées par d'autres écoles selon des règles très strictes;
- (iii) compenser la non-réponse des écoles et des élèves par des mesures d'ajustement pondéral.

PISA exige ainsi dans les grandes lignes un taux de participation des écoles de 85% et un taux de participation des élèves de 80% (Adams & Wu, 2002; OECD, 2005). Les pays qui ne satisfont pas à ces exigences encourent le risque de ne pas figurer dans les rapports internationaux. Les données des pays qui satisfont aux exigences relatives à la participation des écoles et des élèves sont généralement considérées comme valides. En fait, il serait scientifiquement plus correct d'affirmer que ces données ne présentent pas de biais considérés comme suffisamment larges pour invalider la signification des résultats.

Cependant, ces différentes stratégies ont été élaborées à une époque où les indicateurs de tendance ne figuraient pas parmi les objectifs assignés aux surveys, et une différence peut à la fois revêtir peu ou prou d'intérêt sur le plan synchronique mais une importance pédagogique et/ou politique considérable sur le plan diachronique. L'exemple du Portugal dans PISA 2000 illustre parfaitement ces

propos. En 2000, ce pays obtient un taux de participation des écoles de 95.27% et un taux de participation des élèves de 86.28%, résultats largement supérieurs aux exigences minimales. Cependant, cette participation des élèves masque des taux variables de participation selon l'année scolaire fréquentée et selon le sexe de l'étudiant. Ainsi, les filles sont plus de 87% à avoir répondu aux tests de rendement alors que les garçons ne sont que quelque 82%. Quant aux taux de participation par année scolaire, ils sont respectivement de 76%, 80%, 87% et 88% pour les 7e, 8e, 9e et 10e années d'enseignement obligatoire. Monseur (2005) a généré de nouveaux scores en tenant compte de ces taux de participation différentiels. Au lieu d'obtenir 470 sur l'échelle combinée de compréhension de lecture, le Portugal obtient 463. Ce changement ne modifie nullement le classement du pays. Cependant, une différence de 7 points entre les collectes 2000 et 2003 pourrait s'avérer statistiquement significative.

En d'autres termes, les exigences minimales relatives à la qualité de l'échantillon ont été établies à une époque où les objectifs poursuivis ne comportaient pas d'indicateurs de tendance. Cet exemple montre qu'une différence de rendement entre deux collectes de données pourrait s'avérer significative alors qu'elle résulte essentiellement des caractéristiques des deux échantillons. Ne conviendrait-il pas de revoir ces exigences à la hausse pour éviter de tels écueils ?

### Sélection des items d'ancrage

Pour établir un indicateur de tendance entre deux ou plusieurs collectes de données, il importe de rapporter les performances des élèves sur une seule et même échelle. Pour cela, il suffit d'insérer dans toute nouvelle épreuve de rendement, des questions (classiquement appelées items en psychométrie) d'épreuves précédentes, selon certaines règles et pour autant que ces items n'aient pas été diffusés. Ainsi, l'ensemble des épreuves de lecture dans PISA2000 comportait quelque 130 items répartis en 38 unités. En règle générale, une unité de lecture débute par un texte suivi de 3 ou 4 questions. En 2003, le nombre de questions de lecture s'élevait à 28, réparties en 8 unités.

En 2000, la Suisse obtenait un score de 494 sur l'échelle combinée de compréhension de lecture et un score de 499 sur la même échelle en 2003. Cette différence n'est pas statistiquement significative. En appliquant une méthodologie strictement identique à PISA2003 (OECD, 2005), Monseur (2006) a recalculé la valeur que prendrait l'indicateur de tendance sur chaque sous-ensemble de 7 unités parmi les 8 unités. Le tableau 3 présente les variations dans la différence de rendement en compréhension de lecture entre PISA2000 et PISA2003 pour les différents sous-ensembles d'unités.

Tableau 3: Variation dans l'indicateur de tendance

Pays	Changement de l'indicateur de tendance en compréhension de lecture							
	Sans l'unité 1	Sans l'unité 2	Sans l'unité 3	Sans l'unité 4	Sans l'unité 5	Sans l'unité 6	Sans l'unité 7	Sans l'unité 8
Allemagne	5.73	-0.79	1.07	4.39	1.77	-2.25	-6.24	-1.70
Autriche	3.34	-2.66	2.48	1.58	0.40	-4.17	-4.79	2.13
CF Belgique	6.79	-1.03	0.07	1.87	1.67	-0.53	-5.54	0.24
France	3.06	3.01	2.30	2.71	-0.36	-6.11	-6.97	1.53
Suisse	5.60	-3.40	3.68	3.21	1.68	-2.21	-6.66	-3.05

La suppression de l'unité 1 aurait amélioré la performance des élèves suisses de 5.60 points sur l'échelle combinée de compréhension de lecture et se serait également traduite par une amélioration entre 2000 et 2003 de 11 points au lieu de 5, différence qui aurait été rapportée dans le rapport initial comme statistiquement significative à 0.10. Par contre, la suppression de l'unité 7 aurait causé une perte de plus ou moins 7 points sur l'échelle, amenant la performance moyenne de 2003 à un niveau légèrement inférieur à celle de 2000. Le choix des unités communes aux deux collectes de données génère donc une variabilité de l'indicateur de tendance, variabilité d'autant plus grande que les épreuves de rendement comportent peu d'items d'ancrage. Augmenter le nombre d'items d'ancrage minimise le problème mais ne le supprime en aucune circonstance.

L'augmentation de 5 points entre 2000 et 2003 reste parfaitement valide, pour autant qu'on limite l'interprétation de ce changement aux 8 unités d'ancrage. Cependant, les décideurs pédagogiques et/ou politiques n'ont pas pour habitude de s'encombrer de telles restrictions ou précautions méthodologiques et généralisent rapidement ces résultats à l'ensemble du domaine évalué, généralisation abusive comme l'indiquent les données présentées au tableau 3. Dans ce contexte, il serait judicieux de privilégier l'interprétation relative de ces indicateurs de tendance au détriment de l'interprétation absolue. En d'autres termes, il importe de comparer l'évolution en Suisse comparativement à l'évolution observée dans les autres pays plutôt que de se concentrer sur la seule évolution suisse.

### **Le caractère relatif des indicateurs d'équité Définition de l'école comme unité d'échantillonnage**

Les rapports initiaux de l'OECD (2001, 2004) ainsi que les rapports thématiques publiés ultérieurement accordent une place de plus en plus importante à l'effet des variables scolaires et aux indicateurs de disparité entre établissements. L'étude PISA offre-t-elle les garanties méthodologiques suffisantes pour assurer la

comparabilité internationale des analyses relatives à l'effet des variables scolaires ? Différents éléments invitent là aussi à la prudence dans l'interprétation des résultats.

Les analyses statistiques qui intègrent le niveau école nécessitent l'adoption d'une définition précise et standardisée du concept d'école. Sans cette définition, les effets entre pays ne peuvent être comparés et recouvrent des réalités potentiellement différentes.

Au sein d'un pays, et de surcroît, à travers le monde, le concept d'école peut recouvrir différentes réalités. Ainsi, la législation belge francophone différencie l'unité administrative (sous l'appellation «école») du lieu géographique circonscrit (bâtiment ou ensemble de bâtiments sous l'appellation «implantation»). Une école peut dès lors comporter une ou plusieurs implantations alors qu'une implantation n'appartient qu'à une et une seule école. Cependant, certains complexes de bâtiments scolaires qui accueillent un très grand nombre d'élèves sont scindés en deux unités administratives (donc, constituent deux écoles). Cet exemple pris en Communauté française de Belgique illustre la complexité qu'il y a à standardiser le concept d'école au niveau international. A ce jour, il n'existe toujours pas une définition internationale du concept école, malgré les nombreuses tentatives des responsables scientifiques d'enquêtes à large échelle.

Malheureusement, les manuels d'échantillonnage des études internationales du type PISA ou IEA ne précisent pas le concept «école» et s'en remettent dès lors aux traditions et usages nationaux.

Ainsi, lors de PISA2000, la Communauté française de Belgique avait sélectionné des «implantations» alors que la Communauté flamande de Belgique privilégiait l'unité administrative comme unité d'échantillonnage. Quelques pays, comme l'Autriche, limitaient l'unité d'échantillonnage à une filière d'enseignement au sein d'une unité administrative donnée. Précisons que ce choix est parfaitement justifié sur le plan de l'échantillonnage puisqu'il permet de réduire considérablement la variance d'échantillonnage pour toute estimation de paramètres de population. Or, comme le dit le sens commun, il y a plus de diversité dans les grands groupes que dans les petits groupes. En scindant un établissement scolaire en différentes unités d'échantillonnage selon la filière d'enseignement, l'Autriche a donc réduit considérablement la diversité académique au sein de ces unités mais, en même temps, a artificiellement augmenté les différences entre «écoles». Il n'est dès lors pas étonnant d'observer que l'Autriche figure parmi les pays qui présentent un coefficient de corrélation intraclasse très élevé (OECD, 2001, 2004) et que les analyses multi-niveaux isolent l'absence d'un effet des variables socio-culturelles au sein de ces «établissements».

Cette difficulté à définir internationalement le concept «école» limite la validité et la pertinence des analyses sur les effets établissement. Par ailleurs, d'un cycle à l'autre, un pays peut également changer la définition du concept école. Ainsi, la Communauté flamande de Belgique et les Pays-Bas sont passés de l'unité administrative dans PISA2000 à l'implantation dans PISA2003 pour mi-

nimiser les perturbations scolaires au sein des établissements testés et de la sorte augmenter leur taux de participation. Pour des raisons administratives, la Communauté française de Belgique a suivi le chemin inverse. Ces changements altèrent évidemment la validité des indicateurs de tendance relatifs aux effets établissement.

### Différences de rendement entre les filles et les garçons

Comparativement à l'étude I.E.A. Reading Literacy (Elley, 1994), l'étude PISA 2000 (OECD, 2001) a mis en exergue des différences non négligeables entre les filles et les garçons en compréhension de l'écrit. Pour rappel, dans l'étude de l'I.E.A. de 1991, l'ampleur de l'effet moyen entre les filles et les garçons s'élevait à 0.07 écart-type et cette différence était non significative dans la plupart des pays. Par contre, en 2001, l'OECD rapportait une ampleur de l'effet moyen de 0.32 sur l'échelle combinée de compréhension de l'écrit.

Ces études diffèrent à plus d'un égard. Ainsi, les études de l'IEA ont pour habitude de définir comme population cible l'ensemble des élèves qui fréquentent une année scolaire alors que PISA a opté pour une population d'âge, à savoir l'ensemble des élèves de 15 ans, indépendamment de l'année scolaire fréquentée. L'impact de telle ou telle population cible dépend des pratiques de redoublement en vigueur dans les différents pays. Au sein des pays de l'Europe du Nord et du Japon, pays qui pratiquent la promotion automatique, les élèves de 15 ans fréquentent pour ainsi dire un seul grade. Par contre, en France et en Belgique, pays qui recourent abondamment au redoublement, le choix de la population cible influencera les résultats. A titre d'exemple, quelque 45% d'élèves de 15 ans présentent un retard d'au moins un an en Communauté française de Belgique. Par ailleurs, ce redoublement affecte différemment les filles et les garçons. Alors que seulement 36% des élèves de sexe féminin présentent un retard, les garçons sont 50% à avoir redoublé au moins une année d'étude. Ces différences sont donc susceptibles d'influencer l'ampleur de la différence de rendement entre sexes.

Un second facteur souvent négligé dans l'interprétation des différences entre filles et garçons réside dans le cadre de l'évaluation. Ce cadre de référence précise entre autres les caractéristiques des stimuli et des items susceptibles d'influencer leurs difficultés. Parmi ces caractéristiques, citons:

- (i) pour la nature du stimulus, la distinction entre textes continus et textes non continus et au sein des textes continus:
  - a. les textes narratifs
  - b. les textes descriptifs
  - c. les textes argumentatifs
- (ii) pour la nature du processus cognitif, la distinction entre retrouver des informations, interpréter, évaluer la forme ou le contenu du texte;

- (iii) pour la forme de la question, la distinction entre question à choix multiple, question ouverte à réponse courte, réponse ouverte construite.

Les différences relatives au type de texte entre IEA Reading Literacy et PISA2000 restent ténues. Par contre, l'étude de l'IEA ne comporte aucune question qui sollicite la réflexion ou l'évaluation alors qu'elles représentent 20% dans PISA. Enfin, la nature des questions varie considérablement selon l'étude: 75% de questions à choix multiple dans l'étude de l'IEA contre seulement 45% dans PISA, pour ainsi dire aucune question ouverte à réponse construite dans l'étude de l'IEA contre 45% dans PISA (Lafontaine & Monseur, 2003).

Il importe donc de connaître l'influence respective de ces différentes dimensions sur l'ampleur de l'effet entre les filles et les garçons. Les rapports internationaux de PISA2000 (OECD, 2001; Kirsch *et al.*, 2003) fournissent des réponses partielles:

Tableau 4: Ampleur de l'effet en compréhension de lecture

Echelles	Ampleur de l'effet
Echelle combinée	0.32
Retrouver des informations	0.23
Interpréter	0.28
Evaluer	0.41
Textes continus	0.39
Textes non continus	0.17

Comme l'indique ce tableau, la différence entre les filles et les garçons semble varier selon le processus cognitif et selon le type de texte. Cependant, ces échelles ne tiennent pas compte de la nature de la question et la littérature scientifique a à plusieurs reprises démontré la différence de comportement de l'élève selon son sexe et selon la nature de la question (Benett, 1993; Lafontaine & Monseur, 2003). Par ailleurs, le type de question ne se répartit pas équitablement selon le processus cognitif. Le tableau 5 présente le nombre de questions dans les tests PISA2000 selon le processus cognitif et selon le format de la question. Les chiffres entre parenthèses indiquent le nombre de questions attendu si les dimensions «type de questions» et «processus cognitifs» étaient indépendantes.

Tableau 5: Distribution des questions de compréhension de l'écrit PISA2000 selon le processus cognitif et selon le format des questions

	Questions à choix multiple	Questions ouvertes	Total
Retrouver des informations	12 (16.7)	24 (19.3)	36
Interpréter	43 (29.8)	21 (34.2)	64
Réfléchir et évaluer	5 (13.5)	24 (15.5)	(15.5)
Total	60	69	129

Ainsi, le processus cognitif «interpréter» est proportionnellement plus souvent évalué par des questions à choix multiple que par des questions ouvertes. A l'inverse, le processus «retrouver des informations» et surtout le processus «Réfléchir et évaluer» sont plus souvent évalués par des questions ouvertes.

Pour différencier ces effets confondus, Lafontaine et Monseur (2006) ont créé dix nouvelles échelles à partir des données cognitives de PISA2000:

1. compréhension à la lecture – retrouver des informations- questions à choix multiple
2. compréhension à la lecture – retrouver des informations – questions ouvertes
3. compréhension à la lecture – interpréter – questions à choix multiple
4. compréhension à la lecture – interpréter – questions ouvertes
5. compréhension à la lecture – réfléchir et évaluer – questions à choix multiple
6. compréhension à la lecture – réfléchir et évaluer – questions ouvertes
7. compréhension à la lecture – textes continus- questions à choix multiple
8. compréhension à la lecture – textes continus – questions ouvertes
9. compréhension à la lecture – textes non continus – questions à choix multiple
10. compréhension à la lecture – textes non continus – questions ouvertes.

Le tableau 6 présente la différence entre les filles et les garçons pour ces dix nouvelles échelles. Les valeurs positives indiquent une performance supérieure des filles et les valeurs négatives, une performance supérieure des garçons.

Tableau 6: Différences entre filles et garçons en fonction du processus, du type de texte et du format de la question

		ALL	AUT	BEL	FRA	SUI
Retrouver des informations	QCM	12	3	19	15	18
	QO	38	22	29	28	26
Interpréter	QCM	33	22	29	27	25
	QO	40	25	36	32	30
Evaluer	QCM	34	33	34	31	26
	QO	58	41	47	40	48
Textes continus	QCM	40	27	34	31	32
	QO	56	39	45	41	49
Textes non continus	QCM	4	-6	15	6	5
	QO	23	12	23	17	11

Comme le suggèrent les résultats du tableau 6, les différences entre les filles et les garçons dépendent de la nature du texte, du format de la question et des processus cognitifs. En d'autres termes, l'ampleur de la différence entre les filles et les garçons peut être modulée selon l'importance accordée à tel type de textes, à tel processus et selon la proportion de questions ouvertes.

D'un point de vue diachronique, il importe de garantir la continuité du test dans ces différentes dimensions. Tout changement dans la répartition des questions selon leur format, le type de textes auxquelles elles se réfèrent ou selon le ou les processus cognitifs sollicités risque de modifier la différence de performances entre sexes. Ainsi, l'épreuve TIMSS 1995 (Beaton et al, 1996) comportait 75 pourcent de questions à choix multiple. En 2003, l'épreuve TIMSS n'en comportait plus que 50 pourcent (Mullis, Martin, Gonzalez & Chrostowski, 2004). Entre ces deux périodes, on observe une réduction de la différence de performance entre les filles et les garçons, réduction qui pourrait être considérée comme une amélioration de l'équité, attribuée à des réformes pédagogiques ou à d'autres facteurs, alors qu'elle pourrait simplement résulter d'une modification de la composition des épreuves.

Assurer la continuité dans la composition d'une épreuve est d'autant plus difficile que les items d'ancrage sont peu nombreux. Dans ce contexte, avec ses 28 items portant sur 8 textes, l'épreuve de lecture de PISA2003 n'offre pas les meilleures garanties de validité diachronique de la différence de performance entre sexes.

## Le coefficient de corrélation intraclasse

Parmi les indicateurs d'équité les plus souvent publiés figure le coefficient de corrélation intraclasse. Cet indice correspond globalement au pourcentage de diffé-

rences associées aux écoles. Comme indiqué ci-dessus, ce coefficient dépend largement du concept d'école adopté dans l'étude.

Monseur et Adams (2002) ont démontré la sensibilité de ce coefficient aux estimateurs de performance retenus et à la fidélité de la mesure. Ainsi, les estimateurs classiques sous-estiment le coefficient de corrélation intraclasse, sous-estimation d'autant plus grande que la fidélité de la mesure diminue.

Depuis la troisième étude internationale en mathématiques et en sciences de l'IEA (Martin *et al.*, 1997) et sous l'influence de la méthodologie adoptée dans le cadre des National Assessment of Educational Progress (Beaton, 1987), la performance des étudiants aux tests de rendement est rapportée sous la forme de Valeurs Plausibles. L'idée sous-jacente peut être résumée comme suit: toute mesure, qu'elle relève des sciences exactes ou des sciences humaines, comporte une composante d'erreur. La performance d'un étudiant à un test peut dès lors varier plus ou moins, non seulement en fonction de ses dispositions physiques et mentales, mais aussi en fonction des conditions d'administration de l'épreuve et des caractéristiques psychométriques du test. Depuis TIMSS 1995, les bases de données comportent pour chaque étudiant testé un ensemble de «valeurs plausibles» qui peuvent dès lors être considérées comme un ensemble de performances possibles de celui-ci.

Le succès incontesté des Valeurs Plausibles réside dans leur capacité à fournir des paramètres de population non biaisés. En effet, les estimateurs classiques sur-estiment la variance de la performance des étudiants et peuvent renvoyer des performances moyennes biaisées si la difficulté du test n'est pas adaptée à la population testée (Wu & Adams, 2002).

Cependant, à l'inverse des estimateurs classiques, les Valeurs Plausibles nécessitent des présupposés sur la distribution de la performance des étudiants, présupposés qui doivent être cohérents avec les différentes analyses secondaires qui seront menées au départ de ces données. Ainsi, la différence de rendement entre sexes sera sous-estimée si le modèle psychométrique ne précise pas la présence de deux sous-populations: la sous-population de filles et la sous-population de garçons. De même, les différences de rendement moyen par établissement seront sous-estimées si le modèle psychométrique n'intègre pas l'appartenance des élèves testés à leur établissement scolaire. C'est précisément l'erreur commise par les responsables de TIMSS 1995. La performance des élèves qui ont participé à cette étude a été estimée par l'intermédiaire de valeurs plausibles générées selon le modèle de réponse à l'item à un paramètre ou modèle de Rasch. Cependant, l'appartenance de l'élève à telle ou telle école n'a pas été précisée dans le modèle psychométrique.

En 1999, l'IEA a généré de nouvelles valeurs plausibles pour les données de 1995 en utilisant un modèle de réponse à l'item à 3 paramètres et en intégrant la variable appartenance à l'école dans le modèle psychométrique. Le tableau 7 présente les coefficients de corrélation intraclasse calculés au départ des deux séries de valeurs plausibles.

Tableau 7: Coefficients de corrélation intraclasse (Monseur, C. & Lafontaine, D., à paraître)

Pays	95 PVs	99 PVs
Allemagne	0.46	0.59
Autriche	0.33	0.45
Communauté française de Belgique	0.39	0.51
France	0.26	0.35
Suisse	0.43	0.56

Entre 1995 et 1999, le coefficient de corrélation intraclasse pour la Suisse passe de 0.43 à 0.56. Ce changement, essentiellement dû à l'insertion de la variable appartenance à l'école dans le modèle psychométrique, démontre l'influence du modèle de mesure adopté sur certains indicateurs d'équité et rappelle l'importance du contexte méthodologique dans l'interprétation d'un indicateur de tendance.

## Conclusion

En quelque quatre décennies, les surveys en sciences de l'éducation ont largement bénéficié des avancées scientifiques considérables qu'ont connu le domaine de la mesure et les modèles statistiques. De plus, l'intérêt croissant des décideurs politiques pour ces études a permis de dégager des moyens financiers importants. Cet ensemble de facteurs a sans nul doute contribué à la valeur des données recueillies qui atteignent aujourd'hui une qualité et une richesse inégalées. Ces enquêtes ont par ailleurs permis d'améliorer notre compréhension des systèmes éducatifs et nul ne peut nier l'impact incontestable qu'elles ont eu sur les politiques éducatives et les réformes pédagogiques de ces trente dernières années.

La volonté politique d'accroître la visibilité de ces surveys s'est entre autres traduite par la mise à la disponibilité gracieuse des données aux chercheurs dans le cadre du programme PISA, qui peuvent dès lors s'adonner à des analyses secondaires pour mettre à l'épreuve des hypothèses aussi multiples que variées. Ces analyses négligent parfois le contexte méthodologique dans lequel les données ont été recueillies ou tout simplement en oublient ou en ignorent les limites. L'avènement récent des indicateurs de tendance accentue ce risque. En effet, une différence qui revêt peu d'intérêt pédagogique sur le plan synchronique peut se voir attribuer une importance considérable si on l'envisage dans une perspective diachronique. Dès lors que de tels indicateurs sont publiés, il n'est pas besoin d'une longue démonstration pour comprendre qu'ils focalisent toute l'attention des décideurs, soucieux avant tout d'obtenir une réponse scientifiquement fondée à une question cruciale s'il en est, dès lors qu'il s'agit de performances sco-

laire: le niveau des élèves a-t-il ou non progressé ? Les analyses menées ci-dessus témoignent de la sensibilité des indicateurs de tendance au contexte méthodologique de la recherche et soulignent combien il convient, en la matière, d'être circonspect dans le maniement et l'interprétation de tels indicateurs, dont «l'objectivité» est toute relative.

## Références

- Adams, R. J. & Wu, M. (Éds.). (2002). *PISA 2000 Technical Report*. Paris: OECD.
- Beaton, A. E. (1987). *Implementing the new design: The NAEP 1983-1984 technical report*. (Report No 15-TR-20). Princeton, NJ: Educational Testing Service.
- Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Kelly, D. L. & Smith, T. A. (1996). *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: Boston College.
- Benett, R. E. (1993). On the meaning of constructed response. In R. E. Benett & W.C. Ward (Éds). *Construction versus choice in cognitive measurement. Issues in constructed response, performance testing, and portfolio assessment*. (pp. 1-29) Hillsdale: Lawrence Erlbaum Associates.
- Bottani, N. (2004). Les évaluations internationales des acquis des élèves et leur impact sur les politiques d'éducation. *Politiques d'Education et de Formation*, 2 (11), 9-20.
- Elley, W. B. (1994). *The IEA Study of Reading Literacy: Achievement and instruction in thirty-two school systems*. Oxford: Pergamon Press.
- Groves, R. M., Dilman, D. A., Eltinge, J. L. & Little, R. J. A. (Éds.). (2002). *Survey nonresponse*. New York: Wiley.
- Husen, T. (1979). An international research venture in retrospect: the IEA surveys. *Comparative Education Review*, 23 (3), 386-407.
- Kirsch, I., de Jong, J., Lafontaine, D., McQueen, J., Mendelovits, J. & Monseur, C. (2003). *La lecture, moteur de changement. Performances et engagement*. Résultats de PISA 2000. Paris: OCDE.
- Lafontaine, D. & Monseur, C. (2003). *Influence du contenu et des modalités d'évaluation sur les indicateurs d'équité*. Actes du 16e Colloque international de l'Admée. (Liège, 4-6 septembre 2003). (pp. 335-347).
- Lafontaine, D. & Monseur, C. (2006, April). *Impact of Test Characteristics on Gender Equity Indicators in the Assessment of Reading Literacy*. Paper presented at the AERA conference, San Francisco.
- Martin, M. O. & Kelly, D. L. (1996). *Third International Mathematics and Science Study: Technical Report, (Vol. II): Implementation and Analysis: Primary and Middle School Year*. Chestnut Hill, MA: Boston College.
- Martin, M. O., Mullis, I., Beaton, A., Gonzales, E.J., Smith, T.A. & Kelly, D.L. (1997). *Science achievement in the primary school years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chesnut Hill, M.A.: Boston College.
- Monseur, C. (2005). An exploratory alternative approach for student non response weight adjustment. *Studies in Educational Evaluation*, 31 (2/3), 129-144.
- Monseur, C. (2006, April). *The Computation of Equating Errors in International Surveys in Education*. Paper presented at the AERA conference, San Francisco.
- Monseur, C. & Adams, R. (2002, April). *The limitation of the plausible values*. Paper presented at the International Objective Measurement Workshop, New Orleans, L.A.
- Monseur, C. & Lafontaine, D. (sous-presses). *Methodological Issues Raised by Equity Indicators Derived from Multilevel Analyses*.
- Mullis, I. V. S., Martin, M. O., Gonzalez, E. J. & Chrostowski, S. J. (2004). *TIMSS 2003. International Mathematics Report: Findings from the IEAs Trends in International Mathematics and Science Study at the fourth and eighth grades*. Chestnut Hill, MA: Boston College.

- OECD (1999). *Measuring Students Knowledge and Skills: A New Framework for Assessment*. Paris: OECD.
- OECD (2001). *Knowledge and Skills for Life. First Results from PISA 2000*. Paris: OECD.
- OECD (2004). *Learning for Tomorrow's World. First Results from PISA 2003*. Paris: OECD.
- OECD (2005). *PISA 2003 Technical Report*. Paris: OECD.
- Purves, A. (Éd.). (1989). *International Comparisons and Educational Reform*. Virginia: Association for Supervision and Curriculum Development.
- Winglee, M., Kalton, G., Rust, K. & Kasprzyk, D. (2001). Handling item nonresponse in the U.S. Component of the IEA Reading Literacy Study. *Journal of Educational and Behavioral Statistics*, 26, 343-359.

**Mots clés:** études comparatives, indicateurs, méthodologie, taux de participation, ancrage, valeurs plausibles, différences garçons-filles, corrélations intra-classe

## Der relative Charakter der Tendenz-Indikatoren

### Zusammenfassung

Seit Beginn der neunziger Jahre und wegen der immer größeren Anfrage seitens der verantwortlichen Politiker des Bildungswesens verfolgen die internationalen Survey-Studien in den Erziehungswissenschaften ein zusätzliches Ziel zu dem der traditionellen Veröffentlichung von Tendenz – Indikatoren: das Erheben von Indikatoren der Leistungsentwicklung bei Schülern.

Dies zeigt sich deutlich in folgenden Studien: Programme International de Suivi des Acquis des élèves (PISA), Progress International Reading Literacy Study (PIRLS), Trends International Mathematics and Science Study (TIMSS).

Obschon der Unterschied von 10 Punkten auf einer Skala mit einer Standardabweichung, die generell per Konvention auf 100 festgelegt wird, nichts an der pädagogischen Interpretation der Resultate für eine einzelne Datenerhebung ändert, erhält ein solcher Unterschied eine ganz andere pädagogische und politische Dimension, wenn sie den Leistungsunterschied zwischen zwei Datenerhebungen darstellt.

Um gültig zu sein, setzt dieser diachronische Vergleich voraus, dass die Methodologie, im weiten Sinne, perfekt identisch für beide Datenerhebungen ist oder dass eventuelle methodologische Veränderungen den zeitlichen Vergleich nicht beeinflussen.

Diese Annahme wird im Folgenden anhand einiger Beispiele größtenteils aus der Pisa-Studie mit der empirischen Wirklichkeit konfrontiert und belegt wie wichtig es ist, Tendenz-Indikatoren mit großer Vorsicht zu interpretieren.

**Schlagworte:** internationale vergleichende Studien, Leistungsindikatoren, Partizipationsrate, Geschlechterunterschiede

## Il carattere relativo degli indicatori di tendenza

### Riassunto

In risposta a richieste sempre maggiori dei responsabili delle politiche educative, a partire dai primi anni novanta le indagini internazionali nel campo delle scienze dell'educazione hanno affiancato, agli obiettivi tradizionalmente perseguiti, la pubblicazione di indicatori di tendenza finalizzati al monitoraggio delle prestazioni delle popolazioni scolastiche. Testimone ne è il nome degli studi più recenti: *Programme International de Suivi des Acquis des élèves (PISA)*, *Progress International Reading Literacy Study (PIRLS)*, *Trends International Mathematics and Science Study (TIMSS)*.

Sebbene una differenza di 10 punti su di una scala la cui deviazione standard è generalmente fissata per convenzione a 100, non modifichi l'interpretazione pedagogica dei risultati all'interno di una medesima raccolta di dati, una tale differenza assume tutt'altra dimensione pedagogica e politica se rappresenta l'evoluzione di una prestazione fra due raccolte di dati. Perché sia considerato valido, tale raffronto diacronico presuppone però che la metodologia, intesa in senso largo del termine, sia perfettamente identica nelle due raccolte di dati, o che le eventuali modifiche metodologiche non intacchino assolutamente il confronto temporale. Utilizzando degli esempi estrapolati dai dati PISA, questo presupposto è confrontato alla realtà empirica. I risultati dimostrano la necessità di una grande prudenza nell'interpretazione degli indicatori di tendenza.

**Parole chiave:** indagini comparative, indicatori, metodologia, tasso di partecipazione, plausible values, differenze di genere, correlazione intraclasse

## The Relativity of Trend Indicators

### Summary

Starting in the nineties, and under the increasing demands of policy makers, international surveys in education pursue not only the goal of revealing trend indicators but also to predict the future development of students' achievement. The titles of current studies point to this: *Programme International de Suivi des Acquis des élèves (PISA)*, *Progress International Reading Literacy Study (PIRLS)*, *Trends International Mathematics and Science Study (TIMSS)*. Whereas a 10 point difference on a scale of which the standard deviation via convention is fixated at 100 does not change the pedagogical interpretation of results for a data collection, such a difference takes a very different meaning at the pedagogical and political levels if it corresponds to the difference of performance between two data collections. In order to be valid, this diachronic comparison premises that on the one hand, the methodology – broadly speaking – is perfectly identical for both data collection, and on the other hand that the possible

methodological changes do not have any impact on the time comparison. This statement is submitted to empirical evidence through a few examples mainly drawn from PISA; the results implicate that careful interpretation of such trend indicators is required.

**Keywords:** comparative surveys, indicators, methodology, participation rate, anchoring, plausible values, gender differences, intraclass correlation