Éducation comparée / nouvelle série, n°7, pp 17-37.

Les évaluations internationales menées dans les pays pauvres sont-elles fiables ?

Gérard Lassibille
Institut de Recherche sur l'Economie de l'Education, Centre National de la Recherche Scientifique,
Dijon,
France

RÉSUMÉ

Cet article passe en revue les principales études internationales d'évaluation qui sont menées dans le contexte des pays pauvres, et souligne l'absence d'éléments empiriques permettant d'apprécier le degré d'appariement des tests aux programmes scolaires. A titre d'illustration, nous évaluons la sensibilité d'un test international en comparant les réponses apportées à un même instrument par des élèves qui fréquentent deux années d'études consécutives. En utilisant la méthode du score de propension, qui permet de contrôler les caractéristiques des élèves et des enseignants, nous montrons que le test est imparfaitement calibré, ce qui remet probablement en cause les conclusions de nombreux travaux empiriques qui empruntent des instruments de même nature.

MOTS CLÉS

Evaluations internationales ; validité de contenu ; pays en développement.

ABSTRACT

This paper reviews the main international assessment studies that are conducted in the context of poor countries and denounces the lack of empirical evidence showing curriculum match between country curricula and the tests content. To illustrate, we evaluate the sensitivity of an international testing instrument by comparing the responses of students in two consecutive grades on the same battery of tests. Using propensity score matching to control for student and teacher characteristics, the results show that the tests are not grade sensitive, which raises the question of the validity of many empirical works that are based on similar instruments.

KEYWORDS

International achievement tests; content validity; developing countries.

Introduction

Une littérature empirique considérable a été consacrée jusqu'à ce jour aux performances des systèmes d'enseignement et à la qualité de l'offre des services éducatifs (voir par exemple, Hanushek 1986, 1995 ; Pritchett et Filmer 1999). Les études internationales d'évaluation des acquis des élèves sont la pierre angulaire de la plupart de ces travaux, qui ont à la fois pour but de fournir des diagnostics objectifs et d'alimenter le débat de politique éducative (Bourdon 2006). Ces évaluations posent un grand nombre de questions d'ordre méthodologique et pratique, largement commentées dans la littérature (voir par exemple, Monseur et Demeuse 2004 ; Greaney et Kellaghan 2008 ; Winter 1998). Un aspect auquel les utilisateurs des études internationales n'accordent probablement pas assez d'importance concerne la qualité des instruments mis en œuvre, et leur aptitude à réellement rendre compte des acquis scolaires. En effet, alors que les spécialistes de l'éducation, et notamment les économistes de l'éducation, ont porté leur attention sur la spécification des facteurs qui entrent dans la détermination des processus pédagogiques, et ont raffiné plus récemment leurs méthodes empiriques d'estimation, ils ont par contre fait peu de cas de la qualité des tests qu'ils utilisent pour évaluer l'efficacité de l'école. Or il va sans dire, que la faculté de ces instruments à rendre compte des apprentissages est tout à fait cruciale pour qui veut mesurer les performances des secteurs d'enseignement, et formuler à l'appui de ses analyses des recommandations de politiques éducatives réellement fondées.

Cet article s'interroge sur la capacité des évaluations internationales menées dans le contexte des pays pauvres à mesurer sans biais les performances des élèves. Une revue des principales d'entre elles montre que ces programmes sont relativement peu diserts en la matière, et que l'on dispose de bien peu d'éléments factuels permettant d'apprécier l'aptitude des tests à appréhender ce pour quoi ils sont conçus (section 1). Profitant de données d'une grande originalité, collectées récemment à Madagascar, nous testons a posteriori, et à titre d'exemple, l'aptitude des instruments conçus pour ce pays dans le cadre du Programme d'Analyse des Systèmes Educatifs de la CONFEMEN (PASEC) à recouvrir les contenus des

programmes scolaires. Nos résultats portent sur une cohorte de quelques 16.000 élèves répartis dans environ 1.000 écoles primaires publiques. Ces élèves, qui fréquentent les classes de troisième et de quatrième, ont tous passé au même moment un test théoriquement conçu pour évaluer les acquis en classe de quatrième. En confrontant les résultats obtenus par l'une et l'autre catégorie d'élèves, cet article apporte un certain éclairage sur le degré d'alignement du test. Notre analyse utilise la méthode du score de propension afin d'éliminer d'éventuels effets aléatoires d'élèves et d'enseignants (section 2). Faute de pouvoir confronter de façon précise les différents items au contenu des programmes scolaires, nos résultats montrent qu'il y a tout lieu de penser que les instruments du PASEC pour le pays en question sont imparfaitement calibrés, et ne permettent sans doute pas d'évaluer sans risque les performances du secteur d'enseignement (section 3).

Bien que nos résultats portent sur un seul pays, ils posent d'importantes questions sur la validité de contenu des batteries de tests mis en œuvre dans les pays pauvres, et par conséquent sur les recommandations de politique éducative qui peuvent découler de leur utilisation. Nos conclusions peuvent être utiles aux chercheurs qui utilisent ce type d'évaluation dans leurs travaux ; ils peuvent l'être aussi pour les spécialistes de l'éducation qui conçoivent ces instruments. Au-delà de ceux-ci, nos résultats peuvent également intéresser tous ceux qui conduisent des expérimentations et des évaluations d'impact dans le domaine de l'éducation, en faisant appel à des instruments similaires (Kremer et Holla 2009). Cet article contribue à la littérature empirique sur la qualité des tests d'évaluation. Alors que des analyses d'appariement aux curricula ont été réalisées pour apprécier la validité de certains tests menés dans le cadre des pays les plus avancés, comme cela a été le cas pour la Troisième étude internationale sur les mathématiques et les sciences – TIMSS – réalisée par l'Association internationale pour l'évaluation du rendement scolaire – IEA - (Beaton et al. 1996 ; Demeuse 2002), cet article est à notre connaissance le premier à apporter un certain éclairage sur la qualité d'une batterie d'instruments conçus dans le contexte des pays pauvres.

Un aperçu des évaluations internationales dans les pays moins avancés

Les pays les moins riches sont pour la très grande majorité d'entre eux exclus du champ des évaluations PIRLS, PISA et TIMSS. Dans ces pays, des systèmes d'évaluation spécifiques ont été mis en place dès le début des années 1990, afin de nourrir les comparaisons internationales et le débat sur les politiques éducatives dans les différentes parties du monde. Nous passons les principaux systèmes en revue, en mettant l'accent sur la difficulté qu'il y a à juger de la qualité de ces instruments d'évaluation.

LLECE

Le *Bureau régional d'éducation de l'Unesco pour l'Amérique Latine et les Caraïbes* (OREALC[1]), à travers le *Laboratoire Latino-américain d'Evaluation de la Qualité de l'Education* (LLECE[2]), réalise depuis 1998 des évaluations comparatives qui rétro-alimentent le débat sur la politique éducative de la région. Le second - et à ce jour, le dernier programme - du LLECE a été mis en place en 2002 (OREALC 2008). Il inclut 16 pays - Argentine, Brésil, Chili, Colombie, Costa Rica, Cuba, Ecuador, El Salvador, Guatemala, Mexique, Nicaragua, Panamá, Paraguay, Pérou, République Dominicaine et Uruguay – et concernent quelques 200.000 élèves du primaire répartis dans environ 3.000 écoles de la région. Les tests réalisés dans le cadre de ce programme visent à évaluer les performances des élèves de 3ème et de 6ème années du primaire en lecture/écriture et en mathématiques dans les 16 pays participants, ainsi que les acquis en sciences en 6ème année pour un sous-groupe de 10 pays[3]. Selon les concepteurs du projet, les tests sont construits à partir d'une analyse préalable des curricula officiels en vigueur dans les pays concernés et portent sur les éléments qui sont communs à ceux-ci. Toutefois, rien dans les documents publiés ou mis

1 OREALC est l'acronyme de *Oficina Regional de Educación para América Latina y el Caribe*.

2 LLECE est l'acronyme de *Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación.*

3 Ces pays sont les suivants : Argentine, Colombie, Cuba, El Salvador, Panama, Paraguay, Pérou, République Dominicaine, Uruguay et Mexique (Nuevo León).

en ligne sur le site web de OREALC/UNESCO ne permet de vérifier empiriquement qu'il en est bien ainsi.

PASEC

Le *Programme d'Analyse des Systèmes Educatifs des Pays de la CONFEMEN* (PASEC) a vu le jour lors de la *43ème session ministérielle de la Conférence des Ministres de l'Education* des pays *ayant le Français en partage* (CONFEMEN) qui s'est tenue à Djibouti en 1991. Les premières évaluations ont été réalisées en 1993. Les acquis des élèves sont mesurés en français (et/ou en langue nationale si celle-ci est la langue d'enseignement) et en mathématiques. Les tests s'adressent à des élèves de 2ème et de 5ème années du primaire qui fréquentent l'enseignement public et l'enseignement privé. Contrairement à la majorité des programmes internationaux, les acquis des élèves sont testés à deux reprises, en début et en fin d'année scolaire. A ce jour, 18 pays d'Afrique Francophone et de l'Océan Indien ont participé au programme : Bénin, Burkina Faso, Burundi, Cameroun, Congo, Côte d'Ivoire, Djibouti, Gabon, Guinée, Madagascar, Mali, Maurice, Mauritanie, Niger, République Centrafricaine, Sénégal, Tchad et Togo[4]. Les tests sont dits être élaborés en référence aux programmes scolaires en cours dans les pays participants, et sont sensés porter sur les aspects communs à ceux-ci, sans que cela ne soit toutefois confirmé par une quelconque analyse empirique. Sur la base de ces données, le PASEC publie de nombreux rapports et de nombreuses études nationales qui contribuent au dialogue politique sur les questions de la qualité de l'éducation en Afrique francophone.

MLA

Le *Monitoring Learning Achievement* (MLA), administré de façon conjointe par l'UNESCO et l'UNICEF, a vu le jour au début des années 1990 suite à la Conférence Mondiale sur l'Education pour Tous qui s'est tenue à Jomtiem. Le programme vise à créer un système de suivi permanent axé sur les objectifs de l'éducation pour tous qui s'attache à la qualité de l'éducation et aux acquis des élèves (Chinapah 1995). Un de ses objectifs principaux est de renforcer la capacité des pays à

[4] Il faut noter que tous les pays participants ne sont pas testés au même moment.

conduire des évaluations périodiques de leurs systèmes d'enseignement. Les tests sont élaborés en référence à des compétences de base définies au niveau de chaque pays, et incluent également, mais en nombre variables selon les contextes nationaux, des items communs à plusieurs pays. Depuis 1992, 72 pays, la très grande majorité à faibles revenus, ont participé au programme, dont 39 pays d'Afrique Sub-saharienne, 13 d'Afrique du Nord et du Moyen Orient, 10 d'Europe et d'Asie Centrale, 5 d'Asie et du Pacifique, et 3 d'Amérique Latine et des Caraïbes[5]. Le MLA est conçu pour évaluer la qualité de l'école à différents niveaux d'enseignement. Ainsi le programme dit MLA I mesure les acquis des élèves de 4ème et 5ème années d'études en calcul et en lecture, de même que leurs compétences relatives à la vie courante ; le MLA II évalue quant à lui les performances en mathématiques, en science et en vie courante des élèves qui fréquentent la 8ème année d'études. A l'instar des autres programmes internationaux, des questionnaires sont administrés simultanément aux élèves, aux enseignants et aux directeurs d'établissement, de façon à identifier les principaux facteurs qui déterminent les performances scolaires des élèves. La construction des tests est, à notre connaissance, peu documentée et apparemment rien ne permet d'apprécier la qualité des instruments développés dans le cadre du programme.

SACMEQ

Le *Southern and Eastern Africa Consortium for Monitoring Educational Quality* (SACMEQ) regroupe plusieurs ministères de l'éducation qui, avec l'assistance de *l'Institut International de Planification de l'Education* (IIPE) de l'UNESCO, ont décidé de partager leur expertise dans le domaine de l'évaluation de la qualité de l'éducation (IIEP 2006). A ce

[5] Ces pays sont les suivants : Afrique du Sud, Botswana, Burkina Faso, Burundi, Cameroun, Comores, Congo, Côte d'Ivoire, Erythrée, Ethiopie, Gabon, Gambie, Ghana, Guinée, Guinée Bissau, Guinée Equatoriale, Kenya, Lesotho, Liberia, Madagascar, Malawi, Mali, Maurice, Mozambique, Namibie, Niger, Ouganda, Rwanda, Sénégal, Seychelles, Sierra Leone, Somalie, Soudan, Swaziland, Tanzanie, Togo, Zambie, Zanzibar, Sao Tome ; Algérie, Maroc, Tunisie, Iraq, Liban, Syrie, Yémen, Jordanie, Bahreïn, Koweït, Qatar, Palestine, Oman ; Arménie, Azerbaïdjan, Croatie, Kazakhstan, Kirghizistan, Tadjikistan, Turkménistan, Ouzbékistan, Yougoslavie, Slovaquie ; Chine, Mongolie, Népal, Pakistan, Sri Lanka ; Equateur, Haïti, Maldives.

jour, le SACMEQ est composé des 15 « pays » suivants: Afrique du Sud, Botswana, Kenya, Lesotho, Malawi, Maurice, Mozambique, Namibie, Seychelles, Swaziland, Tanzanie (Continentale), Tanzanie (Zanzibar), Ouganda, Zambie et Zimbabwe. Trois évaluations, menées à 5/6 ans d'intervalle afin d'apprécier d'éventuels changements dans le temps, ont été réalisées jusqu'à présent. Chacune d'elles s'adresse à des élèves de 6ème année. La première évaluation (SACMEQ I) a été conduite entre 1995 et 1998 dans sept pays, avec pour objectif de mesurer les acquis des élèves en lecture. La seconde évaluation (SACMEQ II) a été réalisée au cours de la période 1999-2004 dans 14 pays. Le SACMEQ II a été théoriquement conçu pour analyser l'évolution des performances scolaires depuis la réalisation du SACMEQ I, en 1995; à la différence de ce dernier, le SACMEQ II cherche à évaluer à la fois les compétences en lecture et en mathématiques[6]. Le dernier projet du SACMEQ (SACMEQ III) est plus ambitieux que ses prédécesseurs, dans la mesure où il vise à évaluer les niveaux atteints par les élèves et leurs enseignants en lecture et en mathématiques, et les connaissances qu'ont les uns et les autres en matière de VIH et de SIDA. De nombreux rapports d'études et d'analyses sont régulièrement publiés par les membres du projet; ils présentent pour chaque pays les principaux résultats des évaluations et font des recommandations en matière de développement de la politique éducative. Les méthodes d'échantillonnage, les procédures d'enquêtes et de saisie des données sont largement documentées et facilement accessibles aux usagers. Mais comme le note le rapport d'audit qui a été conduit récemment sur le SACMEQ (Ercikan et alli 2008), rien ne permet d'évaluer le degré d'alignement des tests aux programmes scolaires. Pour cette raison, il a été fortement recommandé aux responsables du programme de conduire des tests de validité de contenu et d'en publier les résultats afin de pouvoir juger de la pertinence des outils qui sont élaborés.

6 Certains items ont été conçus spécifiquement pour le SACMEQ ; d'autres ont été empruntés à TIMSS.

Les données

Dans le cadre d'un vaste programme d'analyse, le ministère de l'éducation de Madagascar a administré en juin 2007 un test de quatrième année d'études primaires à une cohorte d'élèves qui étaient inscrits en troisième année de ce cycle au cours de l'année scolaire 2005-2006, et ceci indépendamment de la classe dans laquelle ils étaient en 2006-2007. En comparant les résultats obtenus par des élèves de troisième et de quatrième à un test qui est théoriquement conçu pour évaluer les acquis en classe de quatrième, les données ainsi constituées permettent d'apprécier la sensibilité du test et renseignent ex-post sur la façon dont certains tests internationaux sont calibrés.

Le test emprunte les items du pré-test qui a été réalisé en septembre 2004 à Madagascar par le PASEC auprès d'élèves de cinquième année. Parmi les 115 items qui constituent le test originel du PASEC, un ensemble de 80 items a été retenus : 27 en français, 25 en malgache et 28 en mathématiques (Tableau 1). Cet ensemble d'items a été administré en juin 2007 à 15.990 élèves répartis dans environ 1000 écoles primaires publiques[7]. Parmi ces élèves, 11.611 étaient en quatrième année au moment où le test a été réalisé, et 4.379 étaient en classe de troisième. Vingt-cinq élèves au maximum ont été sélectionnés au hasard dans chaque école, puis testés dans les trois matières. Des élèves de troisième et de quatrième étaient présents dans chaque école. Le test a été supervisé par *l'Institut National de la Formation Pédagogique* et l'équipe malgache du PASEC. Il s'est déroulé selon les pratiques courantes en la matière. Tous les administrateurs du test ont suivi les mêmes consignes et ont donné les mêmes instructions aux examinateurs ; aucune irrégularité n'a été signalée tant au cours de la passation que de la correction du test. Comme beaucoup d'autres tests internationaux, le test du PASEC est un test à choix multiples. Chaque item comporte quatre réponses au choix; autrement dit, la probabilité pour un élève d'obtenir un bon score par

[7] La base de données inclut environ 300 écoles dans lesquelles de nouvelles approches pédagogiques étaient implantées. Pour les besoins de notre analyse, et afin de constituer un ensemble homogène, ces écoles ont été supprimées de l'échantillon.

le simple jeu du hasard est relativement faible. Comme cela a été indiqué plus haut, le test est théoriquement conçu pour couvrir le contenu du programme scolaire dans les différentes matières, et donc pour mesurer ce qu'un élève a appris au cours de l'année.

Tableau 1: Nombre d'items administrés aux élèves de troisième et de quatrième

	Test originel du PASEC (2004–05)	Items testés du PASEC (2006–07)
Mathématiques	34	28
Français	43	27
Malgache	38	25
Total	115	80

Les résultats

Les élèves qui ont redoublé la classe de troisième en 2006-2007 n'ayant pas forcément les mêmes caractéristiques que ceux qui ont été promus en classe de quatrième, nous utilisons la technique du score de propension pour comparer les résultats obtenus par les uns et les autres au même test de quatrième année. Par cette méthode, nous nous assurons que les élèves de quatrième (groupe de traitement) et ceux de troisième (groupe de contrôle) diffèrent seulement par le fait que les premiers ont été exposés au programme de la quatrième année d'études, alors que les seconds ne l'ont pas été. Le score de propension (Rosembaum et Rubin 1985) est défini comme la probabilité conditionnelle d'avoir été exposé au programme de quatrième étant donné un ensemble de covariables ou de caractéristiques de pré-traitement. Suivant une pratique courante, le score de propension est estimé à l'aide d'un modèle probit dont les résultats sont présentés en annexe. La probabilité d'appartenance à un groupe plutôt qu'à un autre est prédite sur la base des variables individuelles et familiales suivantes : l'âge et le sexe de l'élève, le niveau d'éducation du père et de la mère et le niveau de richesse du ménage (approximé par le fait que le ménage dispose ou non d'électricité). Dans la mesure où les progrès des élèves peuvent aussi être influencés par les caractéristiques des enseignants, le statut du maître est inclus parmi

l'ensemble des variables susceptibles de déterminer l'appartenance à un groupe plutôt qu'à un autre[8]. Cette variable oppose les enseignants fonctionnaires aux enseignants contractuels. A Madagascar, comme dans de nombreux autres pays pauvres, le statut dans l'emploi est généralement un bon proxy du niveau de qualification de l'enseignant. En effet, alors qu'à Madagascar les enseignants contractuels ont en moyenne un niveau d'éducation générale supérieur aux autres, leur niveau de qualification professionnelle est nettement plus faible que celui des enseignants fonctionnaires. Les deux échantillons d'élèves ont été constitués en appariant individuellement chaque élève traité à un élève non traité ayant le score de propension le plus proche[9]. Lorsqu'un élève traité ne pouvait être apparié à un élève non traité, il était exclu de l'analyse. Comme le montre le Tableau 2, en opérant de la sorte, il est alors possible de constituer un groupe de contrôle qui a des caractéristiques tout à fait similaires à celle du groupe de traitement.

8 Les données disponibles ne permettent pas de contrôler le processus d'appariement des élèves par l'expérience de l'enseignant. Toutefois, l'omission de cette variable n'affecte probablement pas les résultats présentés ici, dans la mesure où des travaux antérieurs ont montré que l'expérience professionnelle du maître n'a pas d'impact significatif sur les progrès des élèves malgaches (Lassibille et Tan 2003).

9 Différentes méthodes d'appariement ont été testées ; celle du score de propension le plus proche produit les sous-échantillons les plus semblables.

Tableau 2: Caractéristiques des échantillons non appariés et appariés d'élèves.

	Echantillon	Moyenne		% biais	% réduction du |biais|	t-test	
		Elèves de quatrième (groupe de traitement)	Elèves de troisième (groupe de contrôle)			t	p > t
Caractéristiques des élèves							
Age (en années)	Non apparié	11,601	11,275	21,1		11,70	0,000
	Apparié	11,597	11,596	0,1	99,6	0,06	0,950
Fille (%)	Non apparié	48,437	43,663	9,6		5,40	0,000
	Apparié	48,559	48,682	-0,2	97,4	-0,19	0,853
Mère alphabète (%)	Non apparié	90,224	87,885	7,5		4,28	0,000
	Apparié	90,26	90,444	-0,6	92,1	-0,47	0,638
Père alphabète (%)	Non apparié	87,339	84,605	7,9		4,53	0,000
	Apparié	87,571	87,632	-0,2	97,8	-0,14	0,888
Ménage disposant de l'électricité (%)	Non apparié	5,124	4,156	4,6		2,54	0,011
	Apparié	5,133	4,817	1,5	67,4	1,10	0,273
Caractéristiques des enseignants							
Fonctionnaire (%)	Non apparié	51,589	50,24	2,7		1,52	0,128
	Apparié	51,546	51,546	0,0	100,0	-0,00	1,000

Les élèves de troisième n'ayant pas été exposés au programme de la classe de quatrième, ils devraient normalement obtenir de faibles résultats au test de quatrième. Que montre l'évidence empirique à cet égard ? Le Tableau 3 ci-après reproduit les pourcentages de réponses correctes aux différents items, calculés pour les élèves de chaque niveau d'études qui composent l'échantillon apparié. Plusieurs faits méritent d'être soulignés. Comme on pouvait s'y attendre, les élèves de troisième ont en moyenne de moins bons résultats que leurs homologues. Toutefois, les différences entre chaque catégorie d'élèves sont relativement faibles. En mathématiques et en malgache, les élèves de quatrième ont un avantage sur les élèves de troisième de 18 et 14% respectivement ; en français, cet écart n'est que de 4%[10]. Si l'on considère à présent les scores combinés dans les trois matières, les élèves de quatrième ont en moyenne un avantage de 12 points seulement sur les élèves de la classe de troisième.

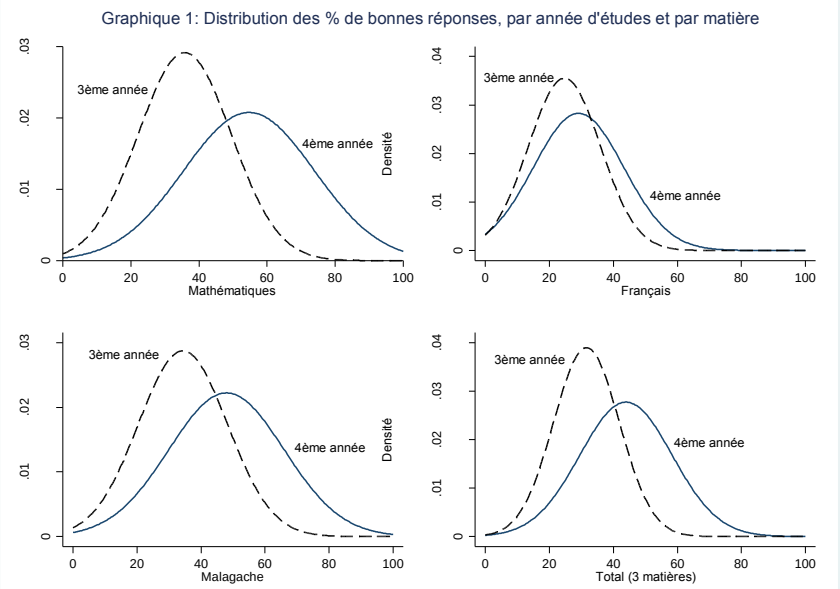
10 Les résultats auraient été différents si l'on n'avait pas utilisé la méthode du score de propension.

Tableau 3: Réponses correctes au test de quatrième.

	Nombre d'élèves	Réponses correctes (%)			
		Mathématiques	Français	Malgache	Total
Elèves de troisième	4.249	35,8	24,8	34,3	31,6
Elèves de quatrième	11.417	54,7	29,2	48,0	44,0
Différence	—	18,9*	4,4*	13,7*	12,4*

* = significatif à 1%.

Le Graphique 1 illustre mieux les différences de résultats au test, en reproduisant la distribution des pourcentages de bonnes réponses obtenues par les deux catégories d'élèves dans les trois matières. Comme on peut clairement l'apprécier, les différentes distributions se chevauchent de façon tout à fait considérable, laissant a priori penser que le test est peu sensible au contenu des programmes (voir notre discussion ci-dessous). En mathématiques et en malgache, on estime qu'environ 60% des deux distributions se chevauchent. Le recouvrement est même plus important en français, puisque selon nos calculs 80% des deux distributions se chevauchent[11].



Comment peut-on expliquer ces résultats quelque peu intrigants? Chaque élève de quatrième a été apparié à un élève de troisième ayant le même score de propension. Comme l'a montré le Tableau 2, l'appariement qui a été réalisé est de bonne qualité et les deux catégories d'élèves ont des caractéristiques pratiquement identiques. Autrement dit, les différences de résultats que l'on observe entre les

11 Dans la mesure où la plupart des tests portent sur de larges ensembles de connaissances, les distributions se recouvrent toujours. Cependant, dans le cas présent, et c'est sur quoi nous voulons insister, le chevauchement est très important.

élèves de troisième et ceux de quatrième ne peuvent s'expliquer par des effets d'élèves ou d'enseignants qui seraient aléatoires. On pourrait plus simplement attribuer les bons résultats des élèves de troisième à la chance qu'ils ont eue lors de la passation du test. Cette hypothèse n'est toutefois guère plausible. Comme cela a été dit, chaque item à choix multiple du PASEC comporte en moyenne quatre réponses possibles, et par conséquent la probabilité d'obtenir un bon résultat au test sans aucune connaissance préalable est très faible. Des différences dans les dotations des écoles en ressources physiques et pédagogiques directement liées à l'enseignement en classe pourraient expliquer en partie les résultats obtenus. Mais là encore, cette hypothèse est peu vraisemblable. Comme le montre le Tableau 4 ci-dessous, les deux groupes d'élèves fréquentent des établissements aux caractéristiques similaires. Dans chaque sous-échantillon, les écoles sont dotées en moyenne d'un même rapport élèves-maître. En ce qui concerne les ressources pédagogiques, les élèves de troisième et ceux de quatrième disposent d'un même nombre de manuels scolaires. Enfin, la qualité des infrastructures scolaires est quant à elle semblable dans chaque sous-échantillon d'élèves[12].

Tableau 4: Ressources des écoles (moyenne).

	Elèves de troisième	Elèves de quatrième	Différence
Rapport élèves-maître	43,9	43,7	-0,2 (0,5)
Livres par élève	1,0	1,0	0,0 (0,0)
Indice d'équipement de l'école[a/]	99,8	100,9	1,1 (1,5)

Note: écart-types entre parenthèses.
a/ Pour la construction de l'indice, voir note 12.

[12] La qualité des équipements est mesurée au travers d'un indice qui est construit au moyen d'une analyse en composantes principales ; l'indice prend en compte les éléments suivants : la structure de l'école est permanente, le nombre de classes est suffisant, l'école dispose d'électricité, d'eau, de latrines et de chaises pour tous les élèves. L'indice varie entre 126 pour les écoles qui sont dotées de tous ces équipements, et 75 pour celles qui n'en disposent d'aucun.

Les bons résultats obtenus par les élèves de troisième pourraient s'expliquer par le fait que l'enseignement donné en classe de quatrième n'est pas aligné sur le programme de quatrième. Toutefois, il n'y a aucune raison objective de penser que les enseignants de la classe de quatrième suivent les programmes de façon moins stricte que leurs homologues de troisième. Par ailleurs, les responsables des sous-circonscriptions scolaires inspectent fréquemment les écoles à Madagascar et suivent de près les aspects liés à la pédagogie. D'après l'enquête menée par le ministère de l'éducation auprès des écoles impliquées dans le test, près de 90% d'entre elles ont été inspectées au cours de l'année précédant l'enquête ; environ 75% de ces inspections avaient pour objet des aspects pédagogiques.

Selon toute vraisemblance, les bons résultats qu'obtiennent les élèves de troisième au test de quatrième s'expliquent par le fait que ce test est mal calibré et que son aptitude à appréhender les aspects pour lesquels il a été théoriquement conçu est faible. En d'autres termes, tout laisse penser que le test est imparfaitement aligné sur le contenu des programmes scolaires enseignés à en classe de quatrième, et qu'il ne permet guère d'apprécier les connaissances qui ont réellement été acquises à cette étape du système d'éducation primaire. Certes, on pourrait objecter à cette conclusion que a) certaines notions étant enseignées à la fois dans les deux niveaux d'études, le test de quatrième peut naturellement inclure certains aspects du programme de troisième ; b) les élèves de troisième – tout du moins les plus talentueux – peuvent répondre correctement à certains items du test de quatrième. Si ces arguments expliquent bien pourquoi certains élèves de troisième peuvent obtenir des résultats aussi bons – voire meilleurs – que d'autres élèves de quatrième, ils ne peuvent certainement pas justifier le chevauchement aussi important (jusqu'à 80%) que l'on observe dans les distributions de bonnes réponses apportées par les deux catégories d'élèves.

Conclusion

Dans cet article nous avons brièvement passé en revue les évaluations conduites dans les pays en développement ou à revenus intermédiaires, qui sont pour la très grande majorité d'entre eux exclus du champ des évaluations « gold standard » que sont PISA, TIMSS ou PIRLS. Cette revue a montré que la construction des instruments d'évaluation dans ces pays-là était très mal documentée et que rien ne permettait d'apprécier l'aptitude des tests à appréhender les aspects pour lesquels ils sont théoriquement conçus. Même les programmes les plus élaborés ne s'attardent pas à vérifier que les instruments sont alignés sur les curricula, les utilisateurs de ces évaluations devant au mieux se contenter de commentaires publiés dans les rapports techniques qui assurent, mais sans le justifier, que les tests ont été construits après des analyses détaillées des programmes et qu'ils coïncident avec le contenu de ceux-ci.

Profitant de données inédites, nous avons testé ex-post la sensibilité d'un instrument conçu dans le cadre du PASEC. A l'aide de la méthode dite du score de propension, nous avons ainsi comparé les résultats obtenus par des élèves de troisième et de quatrième année du primaire à un test de la classe quatrième qui a été administré à tous à la fin de la même année scolaire. Les résultats ont montré un recouvrement jusqu'à 80% des distributions de bonnes réponses apportées par les deux catégories d'élèves. Le fait qu'un nombre aussi important d'élèves de troisième obtient un si bon score à un test conçu pour mesurer les acquis à l'issue d'un programme de quatrième pose à l'évidence d'importantes questions sur la validité de contenu du test. Nos conclusions remettent profondément en cause les résultats d'un grand nombre de travaux empiriques qui analysent l'efficacité interne des systèmes d'enseignement à l'appui de ce type de données, et jettent assurément des doutes sur les recommandations en matière de développement de la politique éducative qui ont pu être formulées à partir de celles-ci. Nos résultats étant limités à un seul pays en développement, ils ne sont naturellement pas généralisables à d'autres contextes, ou à d'autres instruments internationaux d'évaluation. Toutefois, deux leçons sont à tirer du test que nous avons mené ici.

D'une part, les économistes de l'éducation, qui utilisent de façon intensive les évaluations internationales dans leurs travaux, devraient être plus prudents à l'heure d'interpréter leurs résultats. D'autre part, les concepteurs de tests devraient s'assurer que leurs instruments sont bien appariés aux intentions des curricula. Afin d'éviter des doutes et des erreurs qui peuvent être lourds de conséquences pour la définition des politiques éducatives, des études d'alignement devraient être systématiquement entreprises et leurs résultats devraient être publiés et accessibles aux utilisateurs potentiels de ces batteries de tests.

Références bibliographiques

BEATON, A.E., MARTIN, M.O., MULLIS, I.V.S, GONZALEZ, E.J., SMITH, T.A., ET KELLY, D.L. (1996). *Science Achievement in the Middle School Years: IEAs' Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: Boston College, TIMSS International Study Center.

BOURDON, J. (2005). Les apports des études internationales pour évaluer l'efficacité de l'école dans les pays en développement. In M. Demeuse, A. Baye, M.H. Straeten, J. Nicaise, A. Matoul (Eds). *Vers une école juste et efficace. 26 contributions sur les systèmes d'enseignement et de formation* (pp. 73-93). Bruxelles: De Boeck.

CHINAPHA, V. (1995). The Monitoring Project Moving Ahead. Studies and Working Document n°9. Division of Basic Education. Paris: UNESCO.

DEMEUSE, M. (2002). Mesurer le capital humain : qu'y a-t-il dans la « boîte noire » ? . In D. de la Croix, F. Docquier, C. Mainguet, S. Perelman et E. Wasmer (Eds). *Capital humain et dualisme sur le marché du travail* (239-258.). Bruxelles : De Boeck.

ERCIKAN, K., ARIM, R., OLIVERI, M. ET SANDILANDS, D. (2008). *Evaluation of Dimensions of the Work of the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) and of its Programme of Cooperation with the International Institute for Educational Planning (IIEP).* Paris: UNESCO.

GREANEY, V. ET KELLAGHAN T. (2008). *Assessing National Achievement Levels in Education*. Washingthon DC: The World Bank.

HANUSHEK, E. (1986). The economics of schooling: production and efficiency in public schools. *Journal of Economic Literature*, 24, 1141–77.

HANUSHEK, E. (1995). Interpreting recent research on schooling in developing countries. *World Bank Research Observer*, 10 (2), 227–246.

IIPE (Institut International de Planification de l'Education de l'UNESCO) (2006). *Lettre d'information de l'IIPE, n°1*, janvier-mars.

KREMER, M. ET HOLLA, A. (2009). Improving Education in the Developping World : What Have We Learned from Randomized Evalautions ? *Annual Review of Economics*, 1, 513-542.

LASSIBILLE, G., ET TAN, J.P. (2003). Student learning in public and private primary schools in Madagascar. *Economic Development and Cultural Change*, 51 (3), 699–717.

MONSEUR, C. ET DEMEUSE, M. (2004). Quelques réflexions méthodologiques à propos des enquêtes internationales dans le domaine de l'éducation. *Politiques d'Education et de Formation*, 11(2), 37-54.

OREALC (Oficina Regional de Educación para América Latina y el Caribe) (2008). *Primer Reporte de Resultados del Segundo Estudio Regional Comparativo y Explicativo (SERCE).* Santiago: UNESCO.

PRITCHETT, L. ET FILMER, D. (1999). What education production functions really show: a positive theory of education expenditures. *Economics of Education Review*, 18 (3), 223–239.

ROSENBAUM P.R. ET RUBIN, D.B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39* (1), 33–38.

WINTER, S.J. (1998). International comparisons of student achievement: can they tell us which nations perform best and which education systems are the most successful? *Education 3 to 13*, 26(2), 26-32.

Annexe: Probabilité d'être en classe de quatrième
(modèle probit)

	Coefficient	Statistique z
Caractéristique de l'élève		
Age (en années)	0,0949	13,31**
Fille	0,1483	6,80**
Mère alphabète	0,1682	4,76**
Père alphabète	0,1286	4,03**
Ménage disposant de l'électricité	0,1432	2,78*
Caractéristique de l'enseignant		
Fonctionnaire	0,0429	1,98*
Constante	-0,8363	-8,55**
Nombre d'observations	15,668	—
LR chi²	241,68**	—
Pseudo R²	0,0132	—

** = significatif à 1%; * = significatif à 5%,