A SAVOIR
16

Décembre 2012

Renforcer la mesure sur la qualité de l'éducation

Analyse comparative des évaluations sur les acquis des élèves au sein des pays en développement

Nadir ALTINOK
BETA (Bureau d'économie théorique et appliquée),
IREDU (Institut de recherche sur l'éducation), Université de Lorraine

Jean BOURDON
IREDU (Institut de recherche sur l'éducation),
CNRS (Centre national de la recherche scientifique), Université de Bourgogne

afd
AGENCE FRANÇAISE DE DÉVELOPPEMENT

Renforcer la mesure sur la qualité de l'éducation

Analyse comparative des évaluations sur les acquis des élèves au sein des pays en développement

Nadir ALTINOK

BETA (Bureau d'économie théorique et appliquée),
IREDU (Institut de recherche sur l'éducation), Université de Lorraine.
nadir.altinok@univ-lorraine.fr

Jean BOURDON

IREDU (Institut de recherche sur l'éducation),
CNRS (Centre national de la recherche scientifique), Université de Bourgogne.
jbourdon@u-bourgogne.fr

CONTACTS

Véronique SAUVAT

Division Recherche économique et sociale, AFD
sauvatv@afd.fr

Valérie TEHIO

Division Education et formation professionnelle, AFD
tehiov@afd.fr

À Savoir

Créée en 2010 par le département de la Recherche de l'AFD, la collection À Savoir rassemble des revues de littérature ou des états des connaissances sur une question présentant un intérêt opérationnel.

Alimentés par les travaux de recherche et les retours d'expériences des chercheurs et opérateurs de terrain de l'AFD et de ses partenaires, les ouvrages de cette collection sont conçus comme des outils de travail. Ils sont destinés à un public de professionnels, spécialistes du thème ou de la zone concernés.

Retrouvez toutes nos publications sur http://recherche.afd.fr

Précédentes publications de la collection (voir page 179).

[Avertissement]

Les analyses et conclusions de ce document sont formulées sous la responsabilité de ses auteurs. Elles ne reflètent pas nécessairement le point de vue de l'AFD ou de ses institutions partenaires.

Directeur de la publication :
Dov ZERAH
Directeur de la rédaction :
Alain HENRY

FSC
Conception et réalisation : Ferrari / Corporate – Tél . : 01 42 96 05 50 – J. Rouy / Coquelicot
Imprimée en France par : STIN

Remerciements

Ce rapport a été financé par l'Agence Française de Développement (AFD). Les auteurs tiennent à remercier l'AFD pour son soutien, et tout particulièrement, Jean-Claude Balmès, Nicolas Gury, Thomas Mélonio ainsi que Valérie Tehio pour l'ensemble de leurs commentaires et suggestions sur le contenu de ce rapport.

D'autres collègues ont également beaucoup aidé. Nous remercions en particulier :

- l'équipe du *Southern and Eastern Africa Consortium for Monitoring Educational Quality* (SACMEQ), dont Stéphanie Dolata, Kenneth Ross et Mioko Saito ;
- l'équipe du Programme d'analyse des systèmes éducatifs des pays de la CONFEMEN (PASEC), dont Jacques Boureima KI, Antoine Marivin et Vanessa Sy ;
- Jean-Marc Bernard pour sa disponibilité afin de nous éclairer sur l'historique du PASEC ;
- l'équipe de l'*International Association for the Evaluation of Educational Achievment* (IEA), dont Barbara Malak, David Ebbs ainsi que Pierre Foy pour leur aide ;
- Gabrielle Bonnet et Patrick Montjourides pour leur appui sur l'*enquête Latin American Laboratory for Assessment of the Quality of Education – Second Regional Comparative and Explanatory Study* (LLCE – SERCE)
- l'équipe de *Research Triangle International* (RTI International), et en particulier Amber Gove et Souhila Messaoud Galusi pour nous avoir fourni une base de données issue d'une évaluation *Early Grade Reading Assessment* (EGRA) ;
- Luis Crouch, pour son aide dans l'accès aux données EGRA.

Sommaire

Résumé

Ce rapport a pour ambition de présenter l'ensemble des évaluations existantes sur les acquis des élèves au sein des pays en développement. Pour ce faire, deux parties permettent d'éclairer l'analyse des évaluations sur les acquis.

La première partie a pour principal objectif de distinguer les différentes approches disponibles dans la mesure des acquisitions et compétences, ainsi que de recenser le niveau de participation des pays. Plusieurs types d'évaluations sont présentés : les évaluations nationales, régionales, internationales, hybrides, ainsi que celles concernant les populations adultes. Pour chaque évaluation, notre analyse présente la nature de celle-ci, des populations évaluées ainsi que les principaux résultats. Nous montrons que la plupart des pays effectuent des évaluations nationales. Par ailleurs, si la participation des pays en développement, dans les tests régionaux et internationaux, s'accroît dans le temps, il subsiste cependant un noyau dur de pays ne participant à aucune évaluation régionale ni internationale, en particulier parmi les pays les plus peuplés de la planète (Chine [1], Inde, Nigeria [2], Bangladesh), ceci même si l'évaluation scolaire y est active.

Dans la deuxième partie, nous présentons les principales caractéristiques des évaluations sur les acquis des élèves dans une approche comparative. Pour chaque évaluation, notre analyse détaille la nature de celle-ci, en distinguant plusieurs caractéristiques. Après la présentation des caractéristiques techniques des tests (méthode d'estimation, échantillonnage), nous nous focalisons sur leur nature et leur contenu. L'utilisation possible des évaluations par le biais de comparaisons inter- et intranationales est relatée. De nombreuses différences apparaissent entre les évaluations étudiées dans ce rapport. Bien que l'on observe une tendance à l'homogénéisation des enquêtes, des caractéristiques de l'enquête PASEC [3] marquent un retard sur un certain nombre de domaines. Par exemple, le non-recours aux méthodes modernes de l'estimation d'un score basé sur des compétences ressort comme la principale faiblesse de cette évaluation. Au contraire, l'approche en termes de valeur ajoutée est le principal atout du PASEC.

[1] À l'exception de la participation des régions à statut spécial de Hong Kong et Macao au *Program for International Student Assessment* (PISA).

[2] Le Nigeria a fait l'objet d'une analyse par une enquête de type *Monitoring Learning Achievement* (MLA).

[3] Programme d'analyse des systèmes éducatifs des pays de la Conférence des ministres de l'Éducation des pays ayant le français en partage (CONFEMEN).

Synthèse

La mesure de la performance des systèmes éducatifs est devenue monnaie courante. Réponse à la volonté croissante d'évaluer les politiques éducatives, elle illustre aussi la logique d'obligation de résultats qui se substitue de plus en plus à celle d'obligation de moyens. La mesure des acquis des élèves et des adultes sert ainsi à évaluer la performance des systèmes éducatifs. S'il va de soi que ces enquêtes ne mesurent qu'une seule dimension de la qualité des acquis scolaires, elles constituent néanmoins une base solide sur laquelle il est possible de s'appuyer pour mettre en parallèle ces acquis et les moyens mis en œuvre pour les obtenir. Par ailleurs, comme le soulignent Duru-Bellat et Jarousse (2001), la référence à des « produits » de l'éducation – et donc ici à la performance dans les acquis scolaires – est une véritable nécessité dans un débat où dominent le plus souvent des arguments rhétoriques. Plus encore dans le contexte de pays en développement, il est souvent fait état d'une *« trappe de qualité »,* suivant laquelle le développement des systèmes éducatifs est contraint non pas uniquement par des facteurs d'offre, construire de écoles et recruter des maîtres, mais aussi des facteurs de demande. Les familles jugeant la qualité de l'éducation dispensée insuffisante ne seraient pas toujours promptes à scolariser leurs enfants (Prichett, 2001). Or, en rapport à ce besoin de connaissances pour pouvoir agir dans le domaine des politiques publiques, l'offre d'évaluations des acquis scolaires est très variée.

Dans une première partie de ce travail, une synthèse sur ces enquêtes nous conduit à les regrouper en cinq grandes catégories :

- des évaluations nationales sur les acquis des élèves (ENAE) qui consistent à évaluer le niveau de connaissances (savoirs et/ou compétences) dans une optique nationale ; elles se réfèrent la plupart du temps au curriculum du pays ;
- des évaluations régionales sur les acquis des élèves (ERAE) qui regroupent des pays d'une même région ou d'une même langue ;
- des évaluations internationales sur les acquis des élèves (EIAE) qui consistent à évaluer des élèves et/ou adultes dans une optique internationale et comparative ;
- des évaluations hybrides sur les acquis des élèves (EHAE). Celles-ci ne sont pas conçues explicitement pour la comparaison internationale, et ne sont pas non plus des enquêtes régionales ;
- des évaluations internationales sur les compétences des adultes (EICA). Au contraire des autres évaluations, certaines enquêtes concernent exclusivement les adultes ; elles s'intéressent davantage au domaine des compétences qu'à celui des savoirs.

En premier lieu, les évaluations nationales sur les acquis des élèves regroupent les tests à l'échelle d'un pays ou d'une région au sein d'un pays. Notre partie présente un recensement des différentes évaluations nationales existant au sein des pays en développement. À l'image des tests comparatifs internationaux, les différents tests nationaux peuvent être divisés en trois catégories :

- la première concerne les tests permettant de dresser un bilan des acquis des élèves à la fin d'une année scolaire ou à la fin d'un cycle d'enseignement distinctif, et comportant des enjeux importants pour le parcours individuel de l'élève. Ces tests sont qualifiés de tests sommatifs ou « d'évaluation de l'apprentissage » ;
- le deuxième groupe de tests nationaux vise surtout à piloter et évaluer les établissements et/ou le système éducatif dans son ensemble ;
- le troisième groupe de tests nationaux a pour principal but de contribuer au processus d'apprentissage des élèves à titre individuel en identifiant leurs besoins spécifiques et en préconisant des opérations de suivi ciblées.

Ce recensement des enquêtes nationales représente la base la plus complète existant à ce jour pour les pays en développement. De leur côté, les enquêtes internationales sur les acquis des élèves sont les évaluations les plus médiatisées de par le monde, en particulier au sein des pays développés. Les résultats de l'enquête PISA de l'Organisation de coopération et de développement économiques (OCDE) sont aujourd'hui repris dans pratiquement tous les grands médias internationaux, ce qui témoigne d'une volonté politique de mieux mesurer ce que « produit » le service éducatif. Cependant, les pays en développement participent aussi, même s'ils restent minoritaires, à ces tests internationaux. Les enquêtes internationales analysées dans le rapport sont les suivantes :

- le cycle d'enquête *Trends in International Mathematics and Science Study* (TIMSS), dont l'objectif central est d'évaluer le niveau des élèves en mathématiques et en sciences, ainsi que de décrire le contexte dans lequel les élèves apprennent. Par ce second objectif, les fondateurs de l'enquête TIMSS ont résolument adopté une approche en termes de finalité politique ;
- l'enquête *Progress of International Reading Literacy Study* (PIRLS) évalue les capacités en lecture au niveau primaire. PIRLS a été effectuée à trois reprises jusqu'à aujourd'hui (2001, 2006 et 2011) ; seuls les élèves du grade 4 ont été évalués, à un âge moyen de 9 ans ;
- l'OCDE a lancé le Programme international pour le suivi des élèves (*Progress for International Student Assessment* – PISA) en 1997 pour répondre au besoin de données sur la performance des élèves qui soient comparables au niveau interna-

tional. À la différence des enquêtes de l'IEA, l'enquête PISA évalue des élèves qui ont tous 15 ans, quel que soit leur niveau de grade.

Pour souligner leur spécificité, trois types d'enquêtes ont lieu dans des régions distinctives :

- le consortium *Southern and Eastern Africa Consortium for Monitoring Educational Quality* (SACMEQ) organise les enquêtes SACMEQ en Afrique anglophone australe. Trois évaluations successives du SACMEQ ont eu lieu jusqu'à aujourd'hui (1995, 2000 et 2007). La principale originalité de la troisième vague a consisté à évaluer l'impact de l'éducation face à l'épidémie de VIH-Sida : une partie des questionnaires adressés aux élèves, aux enseignants ainsi qu'aux directeurs d'école intégrait des questions relatives à ce virus ;
- les enquêtes issues du Programme d'analyse des systèmes éducatifs (PASEC) de la Conférence des ministres de l'Éducation des pays ayant le français en partage (CONFEMEN) concernent les pays francophones d'Afrique subsaharienne. L'enquête PASEC vise à évaluer en début et en fin d'année les élèves du grade 2 et du grade 5. Par exemple, le test de mathématiques au grade 5 inclut des items qui évaluent les connaissances dans les propriétés des nombres et l'habilité des élèves à effectuer des calculs simples (addition et soustraction). Les tests incluent également des items qui demandent aux élèves d'utiliser l'addition, la soustraction, la multiplication et la division dans la résolution de problèmes. Pour tenter de bien différencier les impacts de l'enseignant et des effets de classe, ces enquêtes sont réalisées dans une logique de valeur ajoutée en comparant un test de fin d'année à un premier réalisé en début d'année ;
- le réseau des systèmes éducatifs nationaux des pays d'Amérique latine et des Caraïbes, appelé *Latin American Laboratory for Assessment of the Quality of Education (LLECE)* a été créé en 1994 et est coordonné par l'Office régional de l'*Organisa*tion des Nations unies pour l'éducation, la science et la culture (UNESCO) en Amérique latine et aux Caraïbes. Deux évaluations ont eu lieu jusqu'à aujourd'hui (en 1997 et en 2007).

Ces enquêtes paraissent encore trop techniques pour certains pays, du fait des compétences à mobiliser pour les mettre en œuvre, et nécessitent la disponibilité d'experts, ce qui n'est pas toujours simple à garantir. Souvent, les lacunes du système d'information scolaire national conditionnent la qualité des échantillons retenus dans les enquêtes. Des enquêtes hybrides ont été lancées, il y a quelques années, avec un coût moindre, afin de diagnostiquer le niveau de base en lecture ou en mathématiques. L'élaboration d'une évaluation des compétences fondamentales en lecture (*Early Grade Reading Assessment* – EGRA) a commencé, sous l'égide de l'Agence améri-

caine pour le développement international (USAID). L'objectif fondamental de cette évaluation à été de tester les premiers pas des élèves dans l'apprentissage de la lecture : la reconnaissance des lettres de l'alphabet, la lecture de mots simples et la compréhension des phrases et des paragraphes. Enfin, on peut citer les évaluations sur les compétences des adultes qui devraient concerner à la fois les pays développés et les pays en développement. La récente initiative de l'Institut de statistiques de l'UNESCO (UIS-UNESCO), avec l'évaluation *Literacy Assessment and Monitoring Program* (LAMP), vise ainsi à évaluer les compétences des adultes sur une même base que l'enquête sur la littératie et les compétences des adultes (ELCA) de l'OCDE. Il sera ainsi possible, d'ici quelques années, de comparer les compétences d'adultes de plusieurs dizaines de pays de par le monde, indépendamment de leur niveau de développement.

Dans une deuxième partie, notre travail consiste à comparer les différentes évaluations sur les acquis des élèves. Dans les dimensions étudiées, les principaux résultats sont présentés ci-après.

- **Approche analytique des tests :** la quasi-totalité des tests utilise les techniques modernes d'échantillonnage ainsi que de production des scores. La méthode dite *Item Response Theory* (théorie de la réponse aux items) basée sur l'application de la méthode de Rasch est utilisée, mais à des degrés divers, dans toutes les enquêtes sauf PASEC et EGRA. Cependant, l'évaluation PASEC adopte ces approches depuis 2012. De manière générale, ces méthodes tentent de retracer, à partir de la passation de tests par un échantillon d'individus, une échelle de mesure qui puisse décrire au travers d'une fonction continue le niveau des compétences de chacun en fonction des réponses aux items du test.
- **Échantillonnage :** la plupart des enquêtes définissent la population désirée selon leur grade. Cependant, l'enquête PISA a pour cible les enfants de 15 ans, quelle que soit leur position dans le cursus scolaire. Les enquêtes régionales sont ciblées sur le niveau primaire (PASEC, SACMEQ, LLECE), mais les grades testés varient selon les évaluations. L'enquête PIRLS évalue également le primaire (grade 4[4]). En règle générale, les évaluations internationales se focalisent plus sur le niveau secondaire, tandis que les évaluations régionales testent le niveau primaire. En ce qui concerne la taille de l'échantillon, c'est dans le PISA que le nombre minimum d'élèves est le plus élevé (plus de 5 000) ; au contraire, l'évaluation EGRA ne pose que la limite de 400 élèves.

[4] L'enquête PrePIRLS, lancée en 2011, était un test plus simple que le test PIRLS. Les grades qui pouvaient être évalués étaient : 4, 5 ou 6 (selon les pays).

- **Collecte des données :** la quasi-totalité des enquêtes se déroulent autour des tests écrits, seule l'évaluation EGRA est orale. Par ailleurs, la spécificité de l'enquête EGRA est d'être une évaluation rapide (moins d'un quart d'heure) tandis que les autres peuvent durer jusqu'à plusieurs heures. Ces différences sont importantes, et dépendent de la nature des besoins des différents ministères de l'Éducation. De façon innovante, PISA évalue depuis 2009 une partie des élèves par le biais d'une passation des tests sur ordinateurs. Au contraire d'EGRA, le test PASEC du grade 2 s'effectue en grande partie à l'écrit. Or, une certaine proportion d'élèves semble répondre au hasard à l'intégralité du test, étant donné la faiblesse de leurs scores. Ainsi, il apparaît impossible de savoir si l'élève a mal répondu à la question, ou s'il ne l'a tout simplement pas comprise, faute de maîtrise suffisante de la langue. Si l'évaluation au début de la scolarité obligatoire semble primordiale pour détecter les élèves en difficulté, il importe de mieux redéfinir les méthodes d'évaluation de ces savoirs initiaux.
- **Contenu des tests** : certains tests ont une approche orientée vers la mesure des compétences des élèves/adultes (*PISA, International Adult Literacy Survey* – IALS), tandis que d'autres s'orientent davantage sur les contenus scolaires (PIRLS, SACMEQ, PASEC, LLECE). Enfin, l'évaluation EGRA se base sur les savoirs fondamentaux, c'est-à-dire la lecture, alors que l'on assiste, depuis quelques années, à une convergence du contenu des items, tendant à standardiser la procédure d'évaluation des élèves entre les enquêtes. Tandis que la plupart des enquêtes tendent à évaluer les élèves en mathématiques, ce n'est pas le cas pour la lecture ou encore des sciences. La plupart des enquêtes recensées – à savoir LLECE, TIMSS, PISA, PIRLS, PASEC, SACMEQ – ont recours à un test en mathématiques. Cette matière est perçue comme étant le domaine le plus standardisé, facilitant ainsi l'optique de comparaison internationale.
- **Variables de marginalisation :** la marginalisation peut être mesurée par différents facteurs, tels que l'âge de l'élève, son genre, son origine, la langue parlée à la maison ou encore sa zone d'habitation. La marginalisation a été le thème central du rapport *Education for All Global Monitoring Report (GMR) 2010.* Dans le cadre du rapport 2010, l'équipe du GMR prend en compte les cinq dimensions principales de la marginalisation tout juste évoquées. Si la plupart des enquêtes proposent de telles variables, certaines évaluations, comme SACMEQ, semblent plus appropriées pour ce type d'analyse. En effet, la possibilité de comparer les élèves entre régions à l'intérieur de pays peut permettre une meilleure évaluation de la marginalisation. Toutes les enquêtes n'ont pas une mesure adéquate du niveau socioéconomique des élèves. Ce sont surtout PISA et LLECE-SERCE *(Second Regional Comparative and Explanatory Study)* qui apparaissent en pointe dans ce domaine.

- **Qualité des tests :** les enquêtes sont souvent jugées selon leur degré de qualité, qui englobe la validité ainsi que la consistance des tests. La validité d'un test souligne le degré de cohérence entre ce qui devrait être mesuré et les stratégies de collecte de données et des instruments de mesure. Bien que la validité des tests soit primordiale, seules les évaluations internationales ont été évaluées sur ce point par des spécialistes de l'éducation. Or, nous soulignons que les enquêtes internationales sont souvent critiquées au regard de leur validité. Par exemple, l'enquête PISA est censée mesurer, à l'âge de l'adolescence, les compétences nécessaires dans la vie de tous les jours. Or, les questions posées au test ne relèvent pas toujours de ce domaine précis en reprenant des domaines des curricula. Par ailleurs, la participation de pays en développement aux enquêtes internationales apparaît également critiquable, au regard de la validité : très souvent, les programmes scolaires ne sont pas concordants avec la structure des tests (en raison de la complexité des questions posées). Même si les enquêtes internationales jouissent d'une crédibilité supérieure aux enquêtes régionales, ces dernières apparaissent plus cohérentes, dans la mesure où les contenus scolaires évalués sont plus proches des curricula nationaux que ceux présents dans les enquêtes internationales.
- **Indicateurs normatifs et disponibilité de mesures de référence *(benchmarks)* :** au-delà de la simple évaluation des élèves dans des domaines de compétences variés, les tests récents développent des critères précis permettant d'évaluer le niveau de chaque élève. De façon générale, il est possible de distinguer deux types de tests : les tests à référence normative[5] et les tests à référence critériée[6]. La plupart des tests sont aujourd'hui des tests à référence critériée. Parallèlement à cette approche, les enquêtes développent des *benchmarks* délimitant les performances des élèves. Ceux-ci sont élaborés par des spécialistes de psychométrie. En particulier, les enquêtes PIRLS, PISA, LLECE ainsi que SACMEQ définissent des seuils précis au-delà desquels les élèves sont censés maîtriser des savoirs et compétences particuliers. Notons l'absence explicite de *benchmarks* pour le PASEC, en dehors d'une référence implicite à un seuil de décrochage, et la référence exclusive des *benchmarks* de l'enquête EGRA aux référentiels des États-Unis.

[5] Un test à référence normative *(norm-referenced)* est un test, une enquête ou une évaluation qui fournit une estimation de la position de l'individu testé par rapport à une population prédéfinie, en relation avec la dimension mesurée. Par exemple, l'objectif d'un test à référence normative peut être de voir si les élèves peuvent prononcer un certain nombre de mots prédéfini à la minute.

[6] Un test à référence critériée *(criterion-reference)* est un test qui permet de déduire, à partir des scores réalisés aux tests, si la personne évaluée a acquis ou non les connaissances ou compétences désirées.

- **Comparabilité des enquêtes :** de plus en plus, les ministères de l'Éducation demandent à connaître non seulement le score moyen des élèves mais également les divergences possibles entre les régions constituant le pays, ou encore entre les différents groupes de populations. Les enquêtes peuvent permettre une comparaison internationale, intranationale, mais aussi temporelle. Cependant, seule l'enquête SACMEQ permet ces trois comparaisons simultanément. Peu d'enquêtes offrent la possibilité d'effectuer des comparaisons intranationales, qui est pourtant souvent un objectif recherché par les ministères de l'Éducation. Du fait d'une volonté initiale d'analyser le niveau des connaissances à l'intérieur du pays, les enquêtes telles que PASEC et EGRA permettent difficilement de comparer la performance de tous les pays entre eux. Les enquêtes internationales telles que PISA et PIRLS permettent les comparaisons internationale et temporelle, mais avec des limites parfois fortes. L'évaluation LLECE-SERCE ne permet qu'une comparaison internationale, réduisant d'autant la portée de cette étude.

Introduction

Le contexte dans les pays en développement et le défi de la croissance

Le Forum mondial sur l'éducation s'est tenu à Dakar du 26 au 28 avril 2000. Il a constitué l'événement culminant de la décennie de l'Éducation pour tous (EPT), initiée en 1990 à Jomtien (Thaïlande), et plus particulièrement marquée par le Bilan de l'EPT à l'an 2000. Cette évaluation de l'éducation fut la plus importante jamais entreprise (UNESCO, 2000) sur l'état de l'éducation à la veille du forum. Cette conférence a notamment affirmé que l'objectif d'EPT devait être atteint au plus tard en 2015. Plusieurs objectifs ont été affirmés (voir encadré 1). L'objectif 6 appuyait notamment la volonté d'*« améliorer sous tous ses aspects la qualité de l'éducation dans un souci d'excellence, de façon à obtenir pour tous des résultats d'apprentissage reconnus et quantifiables – notamment en ce qui concerne la lecture, l'écriture et le calcul et les compétences indispensables dans la vie courante »* (UNESCO, 2000).

Dans la plupart des pays, les systèmes éducatifs sont actuellement engagés dans une recherche de « qualité » et d'« efficacité ». Ces notions renvoient le plus souvent aux résultats des élèves dans des tests standardisés. Ces comparaisons internationales ont été utilisées pour légitimer des recommandations sur l'état des systèmes nationaux d'éducation.

Les études économiques de comparaison internationale ont montré que de nombreuses variables éducatives étaient un facteur déterminant de la croissance du produit intérieur brut (PIB) par tête des pays (Barro, 1991 ; Mankiw *et al.*, 1992). Cependant, les problèmes relatifs aux données ont apporté de nombreuses limites : les variables éducatives, telles que les taux de scolarisation ou le nombre moyen d'années scolaires sont des indicateurs imprécis de la mesure du capital humain acquis par l'éducation (Benhabib et Spiegel, 1994 ; Gurgand, 2000 ; Pritchett, 2001). En cela, ils sont davantage une mesure quantitative de l'affectation de moyens à l'éducation qu'une mesure de résultats par l'évaluation des compétences acquises dans la formation initiale. De plus, la littérature de recherche reste toujours largement dans l'expectative pour lier les moyens consacrés à l'éducation et les résultats qui en ressortent (Hanushek, 2006). L'analyse des différences internationales des taux de croissance du PIB montre que les connaissances en mathématiques et sciences sont des composantes essentielles

du capital humain incorporé à la force de travail (Hanushek et Kimko, 2000). Pour autant, ces compétences ne sont pas toujours précisément corrélées avec les mesures quantitatives de l'éducation, ou encore des indicateurs de financement de l'éducation. D'où l'intérêt de rechercher une mesure plus précise de la qualité de l'éducation, même si chaque intervenant est conscient que les savoirs acquis ne le sont pas uniquement dans le système d'enseignement.

Les enquêtes qui vont être analysées retiennent cette vision économique de la « production d'école ». L'enquête sur les acquisitions scolaires retient que l'acquisition des compétences s'explique par une offre d'éducation dans la classe et l'école, d'une part et, d'autre part, par le contexte socioéconomique de l'élève. Ainsi, ces acquisitions se situent dans une logique d'explication dite de la « fonction de production d'école ». Toutefois, cette pierre angulaire de la recherche en éducation, si elle est reconnue comme schéma de réflexion, reste largement discutée au niveau de sa validation (Hanushek, 2003). La comparaison internationale de ce type d'analyses, en particulier si l'on se base sur le travail comparatif mené par Hanushek et Luque (2003) sur les enquêtes TIMSS, n'a pas validé l'idée dominante d'un plus fort impact dans les pays les moins avancés (PMA) des ressources des écoles sur la qualité des apprentissages. Ce constat initial venait des conclusions, dans le début des années 1980, de Heyneman et Loxley (1983) qui, après avoir examiné les effets de la situation socioéconomique et des facteurs scolaires sur les acquisitions des élèves à l'école primaire dans seize pays à faible revenu et treize pays à revenu élevé, avaient observé que l'influence des antécédents familiaux sur les acquisitions variait considérablement suivant le développement économique entre les pays, et que le pourcentage de la variance expliquée par la dotation des écoles était négativement corrélé avec le niveau de développement d'un pays. Ceci a donc été infirmé par Hanushek et Luque (2003), mais aussi par Chudgar et Luschei (2009). En suivant une approche d'héritage social, Gameron et Long (2007) soulignent, pour les pays en développement, une rapide décroissance des effets-école dans l'acquisition à mesure que l'éducation croissante des parents va permettre d'amplifier les effets d'héritage entre génération et, donc, relativement diminuer les effets purement scolaires.

Encadré 1 *Les engagements de Dakar (2000)*

Objectif 1. Développer et améliorer sous tous leurs aspects la protection et l'éducation de la petite enfance, et notamment des enfants les plus vulnérables et défavorisés.

Objectif 2. Faire en sorte que, d'ici à 2015, tous les enfants, en particulier les filles, les enfants en difficulté et ceux qui appartiennent à des minorités ethniques, aient la possibilité d'accéder à un enseignement primaire obligatoire et gratuit, de qualité, et de le suivre jusqu'à son terme.

Objectif 3. Répondre aux besoins éducatifs de tous les jeunes et de tous les adultes en assurant un accès équitable à des programmes adéquats ayant pour objet l'acquisition de connaissances ainsi que de compétences nécessaires dans la vie courante.

Objectif 4. Améliorer de 50 % les niveaux d'alphabétisation des adultes, et notamment des femmes, d'ici à 2015, et assurer à tous les adultes un accès équitable aux programmes d'éducation de base et d'éducation permanente.

Objectif 5. Éliminer les disparités entre les sexes dans l'enseignement primaire et secondaire d'ici à 2005 et instaurer l'égalité dans ce domaine en 2015, en veillant notamment à assurer aux filles un accès équitable et sans restriction à une éducation de base de qualité avec les mêmes chances de réussite.

Objectif 6. Améliorer sous tous ses aspects la qualité de l'éducation dans un souci d'excellence, de façon à obtenir pour tous des résultats d'apprentissage reconnus et quantifiables – notamment en ce qui concerne la lecture, l'écriture, le calcul et les compétences indispensables dans la vie courante.

Source : UNESCO (2000).

Définition de la qualité de l'éducation

La qualité des acquis scolaires échappe aux définitions contextuelles de satisfaction aux objectifs des programmes scolaires et prend plus appui sur la maîtrise de compétences permettant la performance à l'âge adulte. *« L'ancienne notion de qualité est devenue obsolète. En dépit des différents contextes, il existe de nombreux points communs dans la recherche de l'éducation de qualité, qui devraient permettre à chaque individu, femme et homme, d'être des membres actifs à part entière de leurs communautés ainsi que des citoyens du monde. »* (UNESCO, 2003, p. 1).

Très souvent, les comparaisons internationales se basent sur des indicateurs mesurant uniquement la quantité d'éducation (comme les taux de scolarisation ou encore le

nombre moyen d'années passées dans le système scolaire). Est-il légitime de supposer qu'une année d'éducation dans un pays *i* est similaire à une année d'éducation dans un pays *j* ? Le rendement de l'éducation est-il similaire dans l'ensemble des pays ? Peut-on considérer l'éducation comme un bien homogène, à l'image du capital physique ?

Il est aujourd'hui indéniable de parler de l'existence de différences dans la qualité des systèmes éducatifs. Parler de la qualité de l'éducation implique avant tout de réfléchir à la définition que l'on propose pour la mesurer. Plusieurs sont possibles, mais deux d'entre elles apparaissent comme étant les plus caractéristiques. La définition la plus classique est celle que donne Coombs (1985) dans son ouvrage *Les crises mondiales dans l'éducation dans les années 1980,* quand il souligne : « [...] *la dimension qualitative signifie bien davantage que la qualité de l'éducation telle qu'elle est habituellement définie et jugée par la performance des élèves en termes traditionnels de programmes et de normes. La qualité* [...] *dépend également de la pertinence de ce qui est enseigné et appris – comment ceci répond aux besoins actuels et futurs des apprenants concernés, compte tenu de leurs circonstances et perspectives particulières. Elle fait également référence aux changements significatifs apportés au système éducatif lui-même, à la nature de ses apports (étudiants, enseignants, infrastructures, équipement et matériel) ; ses objectifs, les technologies éducatives et de programmes ; et son environnement socioéconomique, culturel et politique* » (p. 105).

En 1995, dans son rapport intitulé *Priorités et stratégies pour l'éducation,* la Banque mondiale émettait les observations suivantes au sujet de la qualité de l'éducation : *« La qualité dans l'éducation est aussi difficile à définir qu'à mesurer. Une définition adéquate doit inclure les résultats des élèves. La plupart des éducateurs aimeraient aussi y inclure la nature de l'expérience éducative aidant à produire de tels résultats – l'environnement de l'apprentissage »* (p. 46).

Ces deux définitions insistent particulièrement sur les résultats des élèves aux tests de performance. Cependant, elles soulignent également la nécessité de s'appuyer sur d'autres dimensions. L'idéal serait alors de construire des mesures multidimensionnelles de la qualité de l'éducation. Il reste néanmoins très difficile, dans une perspective de comparaison internationale, de prendre en compte, de manière homogène, toutes les dimensions de la qualité de l'éducation. Pour cette raison, nous supposerons qu'un système éducatif est de bonne qualité lorsque les élèves y étudiant ont des scores relativement élevés à des tests d'acquisition standardisés. Les indicateurs construits mesurent donc une seule dimension de la qualité de l'éducation, et non celle-ci de façon globale.

L'importance des tests sur les acquis des élèves

Il importe par ailleurs de se demander quel est l'intérêt de recourir aux tests sur les acquis des élèves/adultes. Selon Vegas et Petrow (2008), cinq raisons peuvent être invoquées pour justifier l'usage des enquêtes sur les acquis des élèves ou adultes.

- La première raison renvoie à l'idée selon laquelle l'opportunité d'apprendre peut être considérée comme un droit humain. L'éducation a été reconnue comme un droit fondamental d'après la Déclaration universelle des droits de l'Homme de 1948 ; rappelé lors des déclarations de Jomtien et de Dakar, ce droit a été incorporé dans la quasi-totalité des constitutions nationales. Par ailleurs, la Convention des droits de l'enfant va au-delà de cette garantie en décrivant l'objectif assigné à l'éducation qui inclut notamment le développement des talents ou encore des habiletés mentales et physiques des enfants (article 29). L'éducation pour tous est indispensable pour réduire la pauvreté. Néanmoins, assurer l'universalisation de l'éducation passe surtout par l'équité devant la dimension qualitative de l'enseignement fourni. Au-delà de la simple scolarisation universelle, les enfants ont droit à une éducation de qualité supérieure et identique, quelle que soit leur origine socioéconomique.
- La deuxième raison qui nous invite à nous référer aux enquêtes sur les acquis des élèves concerne l'effet désormais reconnu de la qualité de l'éducation sur les revenus des travailleurs. Jusqu'à très récemment, la plupart des études se sont attachées à démontrer que la quantité d'éducation – mesurée généralement par le nombre d'années d'études – avait un impact sur les revenus. S'appuyant sur les travaux pionniers de Mincer (1974), ces recherches montrent qu'en moyenne, une année supplémentaire d'éducation est associée à un gain de 10 % des revenus, mais également que les rendements de l'éducation diffèrent significativement entre les pays et les niveaux de revenus (Psacharopoulos et Patrinos, 2004). Elles montrent également que, pour expliquer la diversité salariale à un niveau donné, certaines compétences non cognitives non acquises dans les programmes scolaires devaient être prises en compte (Boissière *et al.*, 1985). Par ailleurs, la maîtrise – ou non – de ces éléments non cognitifs influence significativement le rythme des acquisitions cognitiques (Borghans *et al.*, 2008). Depuis quelques années, de nouvelles études démontrent une relation positive entre les acquis des élèves et les revenus des travailleurs (*cf.* UNESCO, 2004, chapitre 2, pour une revue de littérature). Ces recherches utilisent généralement les scores des élèves aux tests sur les compétences en mathématiques, en sciences ou en lecture. Trois études menées aux États-Unis montrent qu'une augmentation d'un écart-type du score des élèves en mathématiques est associée à une hausse de 12 % des revenus (Mulligan, 1999 ;

Murname *et al.*, 2000 ; Lazear, 2003). En utilisant les données d'une enquête sur la littératie des adultes (enquête IALS qui concerne 15 pays, dont le Canada, le Chili, les États-Unis ainsi que 12 pays européens), Leuven *et al.* (2004) montrent que les capacités cognitives des travailleurs – mesurées par leurs niveaux en lecture – agissent significativement sur leurs salaires. Ces effets persistent même si les auteurs introduisent les années d'éducation, ce qui renforce l'importance que l'on doit accorder à la qualité de l'éducation. En ce qui concerne les pays en développement, Hanushek et Woessmann (2007) soulignent que les rendements de l'éducation seraient supérieurs pour ces groupes de pays en comparaison aux pays développés. Par ailleurs, en utilisant les données du Chili dans l'enquête IALS sur la littératie des adultes, Sakellariou (2006) montre que l'augmentation d'un écart-type du niveau des travailleurs en lecture induit une hausse de 15 à 20 % sur leurs revenus.

- Troisième raison : il faut également souligner l'effet de la qualité de l'éducation sur la société prise dans sa totalité. Au-delà des seuls gains économiques de l'éducation, il est reconnu que celle-ci permet des gains extra-économiques, notamment en matière de santé (par exemple sur les mères et leurs enfants), dans la réduction de la mortalité infantile, des migrations, de l'âge des mariages, de la violence civile, de la citoyenneté, etc. Les rendements sociaux de l'éducation – qui incluent l'ensemble de ces aspects – dépassent largement les rendements privés. Plus précisément, la qualité de l'éducation apporte de nombreuses améliorations sociales. À titre d'exemple, de meilleurs scores en lecture et mathématiques sont associés à une baisse des taux de fertilité au Ghana (Oliver, 1999) et en Afrique du Sud (Thomas, 1999).
- Les effets de l'éducation sur le développement économique sont la quatrième raison. La relation entre éducation et croissance économique a souvent été évaluée à partir d'indicateurs quantitatifs de l'éducation. La plupart de ces travaux ont souligné que l'éducation était un facteur clé pour le développement économique. Pour autant, Pritchett (2001) a montré que l'effet de l'éducation sur la croissance n'était pas aussi évident : en soulignant l'importance de la qualité de l'éducation dans le processus de développement économique, l'auteur montre qu'un enseignement de piètre qualité peut être associé à un faible développement économique. En effet, cette piètre qualité envoie un signal qui restreint la demande d'éducation des familles, ce qui cantonne la population dans la pauvreté. En soi, la dimension qualitative de l'éducation peut être un accélérateur fondamental de la croissance économique. Les récents travaux associant la qualité de l'éducation et la croissance économique tendent tous à souligner l'importance de la dimension qualitative de l'éducation (Lee et Lee, 1995 ; Hanushek et Kimko, 2000 ; Barro, 2001 ; Coulombe et Tremblay,

2006 ; Hanushek et Woessmann, 2007). Le travail le plus influent montre notamment que l'augmentation d'un écart-type du score des élèves, au primaire, est associée à un gain de 1 % du taux de croissance annuel du PIB par habitant (Hanushek et Kimko, 2000).

- Enfin cinquième et dernière raison : toujours en suivant Vegas et Petrow (2008), il est nécessaire de souligner l'impact du niveau des acquis des élèves sur les inégalités. La relation entre l'éducation et les inégalités est complexe : si, d'un côté, l'objectif de scolarisation peut conduire à une baisse des inégalités, il ne faut pas négliger le pouvoir qu'a le système éducatif de légitimer les inégalités sociales (Bourdieu et Passeron, 1964). De récentes études montrent que la qualité de l'éducation peut être responsable de l'aggravation des inégalités de revenus, et que l'augmentation de la qualité de l'éducation pour les populations pauvres peut ainsi réduire ces inégalités. L'accès à l'éducation primaire a augmenté dans la quasi-totalité des pays du monde depuis une vingtaine d'années. Parallèlement à cette évolution, la question de la qualité de l'éducation reste fondamentale : la généralisation de la scolarisation permet-elle réellement une baisse des inégalités, ou au contraire, les renforce-t-elle ? Comme le souligne Reimers (2000) pour l'Amérique latine, s'assurer que les élèves acquièrent des savoirs et des compétences est une condition indispensable pour garantir l'égalité des chances. Il convient donc de se focaliser sur la qualité de l'éducation et non seulement la quantité d'éducation. Pour autant, il faudrait aussi évaluer l'importance de la qualité de l'éducation comme variable d'efficacité économique, dans la mesure où l'éducation permettrait aux économies de travailler au plus près possible de la frontière d'efficacité technologique dans une optique de croissance.

1. Présentation des tests sur les acquisitions et les compétences

1.1. Une typologie des tests

Dans cette section, nous présentons les différentes évaluations sur les acquis des élèves et des adultes existant à ce jour. Il est possible de recenser plusieurs groupes d'enquêtes qui concernent près de 110 pays à travers le monde. Le point sensible de l'évaluation de la qualité de l'éducation, par la mesure sur les acquis et les compétences, est de faire la distinction entre une performance et une compétence. Il faut avoir toujours à l'esprit que l'exercice est réducteur : on observe une performance d'un sujet à une épreuve et l'on infère des conclusions sur sa compétence (Mislevy, 1994). Ainsi, les évaluations sur les connaissances des élèves ou adultes peuvent être scindées en cinq groupes.

- *Des évaluations nationales sur les acquis des élèves* (ENAE), qui consistent à évaluer le niveau de connaissances (savoirs et/ou compétences) dans une optique nationale ; elles se réfèrent la plupart du temps au curriculum du pays. On peut citer le cas de l'enquête *Sistema de Medición de la Calidad Educatión* (SIMCE), un test annuel sur l'ensemble de la population scolarisée, qui évalue – sur différentes années – les grades 4, 8 et 10. Ces tests sont basés sur le curriculum chilien et ont une importance capitale dans l'attribution des aides gouvernementales, suivant un mécanisme de pseudo marché. En effet, les résultats sont rendus publics afin d'informer les parents de la performance des écoles publiques et privées. Les aides sont parfois versées par rapport au classement qu'a obtenu l'école (Vegas et Petrow, 2007, pp. 36-54)[7].
- *Des évaluations régionales sur les acquis des élèves* (ERAE), qui regroupent des pays d'une même région ou d'une même langue, et dont l'objectif est de standardiser les tests de pays ayant des caractéristiques communes. Elles permettent aussi, la plupart du temps, des comparaisons entre pays participants. Ces enquêtes sont SACMEQ, PASEC et LLECE.

[7] Pour une comparaison entre enquêtes nationales et examens publics, voir notamment Greeney Kellaghan (2008), pp.14-16.

- *Des évaluations internationales sur les acquis des élèves* (EIAE), qui consistent à évaluer des élèves et/ou adultes dans une optique internationale et comparative. Celles-ci acceptent de plus en plus des pays issus de différents continents et de différents niveaux économiques. L'optique est ici davantage axée sur un classement des pays *(league tables)* et se base sur la possibilité d'atteindre des standards de qualité élevés au niveau international. Ces enquêtes sont PISA, PIRLS et TIMSS.
- *Des évaluations hybrides sur les acquis des élèves* (EHAE). Celles-ci ne sont pas destinées explicitement la comparaison internationale, et ne sont pas non plus des enquêtes régionales. Mises en place en 2007, les enquêtes EGRA concernent l'évaluation des connaissances en lecture/écriture d'élèves du primaire. Cependant, les tests se déroulent indépendamment dans chaque pays, ce qui n'en fait pas une enquête internationale similaire à PIRLS ou TIMSS, où les pays sont testés à une période similaire. Comme pour EGRA, on présentera également l'évaluation *Early Grade Mathematics Assessment* (EGMA) ainsi que *Snapshot of School Management Effectiveness* (SSME). On peut aussi citer l'enquête MLA. Bien que celle-ci ait eu pour objectif initial l'évaluation comparative internationale, les items constituant les tests ont beaucoup varié d'enquête en enquête.
- *Des évaluations internationales sur les compétences des adultes* (EICA). Au contraire des autres évaluations, certaines enquêtes concernent exclusivement les adultes et se basent ainsi plus sur le domaine des compétences que des savoirs. Ces enquêtes sont l'Enquête internationale sur l'alphabétisation des adultes (EIAA) ou *International Adult Literacy Study (IALS),* l'ELCA (ou *Adult Literacy and Life Skills Survey,* ALL) et l'enquête Programme pour l'évaluation internationale des compétences des adultes (PIAAC). Seuls les pays développés sont concernés par ces enquêtes. Cependant, nous montrerons que le projet LAMP – Programme d'évaluation et de suivi de l'alphabétisation – vise à développer une évaluation des compétences des populations adultes dans les pays en développement. Ces évaluations sont nationales et internationales mais, comme elles se réfèrent à un contexte d'efficacité du travailleur en situation d'emploi, elles testent des questions macroéconomiques de compétitivité et permettent une autre lecture de la qualité des systèmes éducatifs : quels sont les systèmes qui permettent le mieux les compétences utiles en situation de travail ?

De cette typologie, on peut retenir des finalités très différentes. L'enquête nationale d'évaluation peut raisonnablement se fixer comme mesure de la qualité, le taux d'assimilation du programme scolaire correspondant au niveau testé. Dans une enquête régionale, des points communs existent souvent dans les systèmes, comme une langue d'enseignement commune ou en partie partagée à un niveau du curriculum, ainsi le

test s'articulera sur le plus grand dénominateur commun des programmes[8]. L'enquête internationale n'a pas les mêmes points d'ancrage : elle tentera de déterminer une analyse synchrone des acquisitions sur des terrains où les langues d'enseignement sont différentes, tout autant que les programmes. Aussi, on constate un éloignement entre une mesure des acquisitions, en rapport à un programme, et ces enquêtes internationales, qui paraissent plus s'orienter vers la mesure de la validation, par les élèves, de macrocompétences qui dépassent la validation de compétences ponctuelles. La qualité se réfère alors plus à la capacité, acquise durant la scolarité, à se préparer à la vie d'adulte. Ceci entraîne un débat sur les finalités des enquêtes d'acquisition qui reste ouvert (Vrignaud, 2006).

1.2. Les enquêtes nationales

1.2.1. *Présentation*

Les évaluations standardisées des élèves jouent dans les systèmes éducatifs nationaux un rôle croissant en tant qu'instrument de mesure et de pilotage de la qualité de l'enseignement, mais aussi de diagnostic des structures. Ces tests nationaux ou enquêtes nationales standardisées répondent surtout à l'objectif de pouvoir disposer d'un outil unique. Cette standardisation des formes d'évaluation des élèves, à savoir celle administrée et organisée au niveau central, est donc liée à l'organisation de chaque système éducatif national. Il s'agit de tests standardisés par les autorités éducatives nationales ou – dans le cas d'États où l'éducation est décentralisée – par l'autorité en charge de ce secteur.

Cette notion de centralité ou de relative déconcentration est en elle-même relativement intéressante. En général, l'éducation suivant le principe de subsidiarité, est souvent organisée à un échelon décentralisé. Ainsi, en Europe, des pays à forte tradition de délégation des pouvoirs à l'initiative locale font que l'éducation, comme d'autres politiques sociales, sont des initiatives de proximité. Bien souvent, l'éducation de base est donc placée sous la tutelle de l'entité la plus déconcentrée[9]. Toutefois, ces dernières années, en particulier suite à la diffusion d'évaluations internationales (surtout celles

[8] Ainsi les programmes sur l'Afrique subsaharienne, SACMEQ pour la zone anglophone australe et PASEC pour la zone francophone, ont relativement trouvé ce dénominateur en fonction du suivi des compétences de base en lecture/écriture et mathématiques.

[9] Si l'on croise, par ailleurs, les champs de compétences dans l'éducation et le niveau de déconcentration, on arrive vite à un grand nombre de combinaisons possibles. Ainsi, la France voit la question des infrastructures scolaires faire l'objet d'un mouvement de déconcentration au niveau le plus fin des territoires. Il est nécessaire de parler de déconcentration, et non de décentralisation, car des normes fortes du niveau central encadrent l'initiative locale. À l'inverse, les questions liées au personnel enseignant restent traitées au niveau national.

de PISA) et à leur retentissement médiatique, le principe de subsidiarité maximale n'est plus retenu comme la forme optimale d'organisation. Dans le cadre de la Confédération helvétique, par exemple, les dispositions constitutionnelles sur l'éducation en vigueur depuis la votation populaire du 21 mai 2006 fixent un principe général : la Confédération et les cantons, dans les limites de leurs compétences respectives, veillent ensemble à la qualité et à la perméabilité de l'espace suisse de formation. Ce choix est très explicite : même si l'obligation de moyens reste du domaine de la subsidiarité, la Constitution fédérale doit garantir le résultat de la qualité d'acquisition de compétences de base homogènes sur l'étendue de la confédération.

Aux États-Unis, la loi *No Child Left Behind* (NCLB) est née d'une inquiétude sur la perte des avantages des États-Unis, en termes de niveau et d'homogénéité des acquisitions du socle de compétences (en particulier la numératie de base), par rapport aux pays d'Asie. Elle a donc eu pour objectif de vérifier que l'initiative des États dans le domaine de l'éducation permettait d'obtenir une garantie de qualité des acquisitions et de leur homogénéité au sein de la population. Même si cette loi n'est pas détaillée dans cette étude, retenons que l'évaluation internationale a entraîné plus d'évaluations nationales. L'évaluation internationale des apprentissages scolaires pose la question de l'étendue de la variété des résultats entre les pays, suivant la position qui caractérise ces pays le long de l'échelle du développement économique. On peut donc s'attendre, si l'on admet logiquement un lien entre qualité des apprentissages et richesse économique, à voir la variabilité des apprentissages se réduire s'ils sont contrôlés en rapport avec la richesse économique. Le système national d'évaluation a ainsi une obligation de résultats sous le double critère de l'efficacité et de la justice.

L'évaluation des élèves constitue, surtout dans sa dimension internationale, un système complexe incluant toute une gamme d'outils et de méthodes souvent menés en externe (même si des services d'études des ministères nationaux sont impliqués). Dans un contexte national, le champ est plus vaste : l'évaluation pouvant être interne ou externe, formative ou sommative, avec une diversité dans le choix des instruments. Malgré les différences d'approches en matière d'évaluation des élèves, liées à la variété de la structure des systèmes, l'évaluation des acquis de l'apprentissage fait partie de la structure globale des systèmes éducatifs nationaux : dans chaque pays, l'évaluation fait partie de l'enseignement et de l'apprentissage, avec pour objectif l'amélioration de la qualité de l'éducation.

Cette intégration au système et à ses normes fait que le processus d'évaluation des élèves est généralement normé par des textes et directives. L'outil découle du programme national, de sa traduction dans les manuels et dans les pratiques enseignantes. Ces réglementations contiennent les principes fondamentaux de l'évaluation :

les objectifs donnés à l'école sont-ils atteints ? Quelles inégalités sont identifiées parmi les élèves, suivant les régions et les contextes de vie ? L'évaluation nationale est, de ce fait, proche des normes et textes nationaux relatifs aux questions liées à la notation des élèves, aux critères de correction et de restitution des résultats. Dans un sens, l'évaluation nationale peut se rattacher à l'évaluation continue qui mesure la participation des élèves en classe, leurs devoirs et les autres travaux. Ainsi l'évaluation nationale peut-elle s'identifier à l'évaluation formative, activité permanente des enseignants, qui vise à contrôler et à adapter le processus d'apprentissage au fil des jours. Elle a lieu tout au long de l'année scolaire, fait partie intégrante des activités d'enseignement et fournit un retour direct d'informations aux enseignants comme aux élèves. Cette évaluation peut être compilée pour servir, au niveau de la classe ou de l'école, à effectuer des choix (passage ou non en classe supérieure ; orientation ; etc.). Ainsi, nous verrons que, pour de nombreux pays en développement, la culture d'évaluation est née de la synthèse et de la mise en cohérence du contrôle des acquisitions des élèves.

Les écoles peuvent souvent adjoindre à ces évaluations des outils de tests nationaux basés sur des procédures définies au niveau central. Ceux-ci sont utilisés afin de garantir la comparabilité des performances des élèves et d'identifier les problèmes d'acquisition. Les résultats de ces tests nationaux peuvent être comparés à différents niveaux. Ils fournissent aux élèves des informations sur les connaissances qu'ils ont acquises, et celles-ci peuvent être comparées à celles de leurs pairs et par rapport aux moyennes du secteur géographique et nationales. Souvent, comme c'est le cas pour les évaluations françaises en troisième année du primaire, un ensemble d'outils accompagne ces tests pour aider l'enseignant dans son approche pédagogique, ou pour remédier à des déficiences de l'élève ou du groupe pédagogique. En général, ces évaluations généralisées servent à alimenter un échantillon représentatif de l'ensemble des écoles du pays afin d'apprécier la variété des acquisitions et, à travers une collecte longitudinale, de dégager les tendances sur le moyen terme. Les pays qui utilisent depuis plus longtemps ces tests nationaux pour aider les enseignants et les établissements à évaluer les connaissances, les aptitudes et les compétences des élèves, élaborent souvent des politiques et des stratégies spécifiques visant à un meilleur équilibre entre l'évaluation formative et les tests et examens nationaux. Parfois, mais ceci ne concerne pas le secondaire, ces tests servent également à organiser l'orientation.

Si l'on examine les tendances des pays, comme en Europe, où les tests nationaux sont déjà anciens, on note un recours fréquent à ces derniers comme objectifs de référence afin de soutenir le processus d'apprentissage des élèves à titre individuel en identifiant les besoins d'apprentissage et en adaptant l'enseignement du groupe en conséquence. Ces évaluations nationales standardisées occupent une place de choix

dans le pilotage du système éducatif. Il semble exister une certaine convergence entre ces opérations nationales et les évaluations internes et/ou autoévaluations réalisées par les établissements. Implicitement se précise une combinaison de modèles descendants de pilotage et d'évaluation avec des approches ascendantes (de l'école vers le système) d'évaluation des élèves à l'initiative des établissements pour une amélioration de la qualité de l'enseignement. Deux débats peuvent modifier ces tendances : l'un est le recours aux résultats de ces évaluations pour établir la performance des établissements (si ce n'est individuellement celle des enseignants), l'autre est une réserve exprimée par les parents. Dans la mesure où ces évaluations conditionneraient l'orientation et la sélection des enfants, l'école pourrait (si elle se base sur ces évaluations) amplifier les inégalités préexistantes.

Enfin, dans un schéma de libéralisation et de mise en compétition des établissements, il est évident que les enquêtes nationales pourraient être utilisées pour mesurer la qualité des écoles et, donc, constituer une information sur les actions à mener sur celles-ci. La fameuse réforme suédoise de 1992, associant la liberté de choix des établissements à une subvention à la famille de l'élève, a été menée sans s'appuyer sur un étalonnage des établissements. À l'inverse, la réforme des *Grant Maintained Schools* (1988), en Grande Bretagne, s'est appuyée sur des tests d'acquisition qui permettaient justement de prendre en compte l'effet école et sa « valeur ajoutée » en rapport à son contexte socioéconomique pour déterminer le montant de l'allocation de l'Etat à chaque école (Harlen, 2007).

Si l'on suit la synthèse du réseau Eurydice [10] (2009), les enquêtes d'évaluation nationale sont systématiquement utilisées pour l'évaluation externe des écoles en Hongrie, Lituanie, Suède, Slovénie et Roumanie. De même, dans l'espace européen, ces enquêtes seraient utilisées partiellement comme source d'évaluation en Belgique, Bulgarie, Estonie, Grande-Bretagne, Islande, Lettonie, aux Pays-Bas, au Portugal et en Slovénie.

Les tests nationaux sont organisés sous l'autorité d'un organe national/centralisé, donc à l'exception de systèmes éducatifs décentralisés où ces tests sont organisés au niveau d'une province ou d'un État, tous les élèves passent le test dans des conditions les plus similaires possible. Ne peuvent être pris en compte dans cette étude les tests

[10] Le réseau Eurydice fournit de l'information sur les systèmes éducatifs européens ainsi qu'une analyse de ces systèmes et des politiques menées en la matière. En 2012, il est constitué de 38 unités nationales basées dans les 34 pays qui participent au programme de l'Union européenne dans le domaine de l'éducation et de la formation tout au long de la vie (les États membres de l'Union européenne (UE), les pays de l'Association européenne de libre-échange (AELE), la Croatie, Serbie et la Turquie) ; il est coordonné et géré par l'Agence exécutive « Éducation, Audiovisuel et Culture » de l'UE, située à Bruxelles, qui élabore ses publications et bases de données (Cf. http://eacea.ec.europa.eu/education/eurydice/index_fr.php)

(i) ayant pour but de détecter les problèmes de développement de l'enfant, ou menés *(ii)* dans certaines filières spécialisées ou *(iii)* des filières dédiées à des minorités. Les outils, en particulier ceux utilisant les nouvelles technologies de l'information et de la communication (NTIC), destinés à aider les enseignants à mener des contrôles continus ou à mesurer l'impact de leurs enseignements, ne sont pas repris ici. Les informations que nous utilisons sont, comme pour les autres sections, les niveaux d'enseignement de classification internationale type de l'éducation (CITE) niveau 1 (primaire) et 2 (secondaire inférieur). Autant que possible, nous avons repris tous les types de tests nationaux, qu'ils aient une finalité sommative ou formative, qu'ils soient basés sur l'exhaustivité ou les sondages.

À l'image des tests comparatifs internationaux, les différents tests nationaux peuvent être divisés en trois catégories :

- la première concerne les tests permettant de dresser un bilan des acquis des élèves à la fin d'une année scolaire ou à la fin d'un cycle d'enseignement distinctif, et comportant des enjeux importants pour le parcours individuel de l'élève. Ces tests sont qualifiés de tests sommatifs ou d' « évaluation de l'apprentissage ». Leurs effets peuvent être repris pour délivrer des certificats et/ou prendre des décisions pédagogiques importantes liées à l'orientation, au choix d'un établissement, au passage en classe supérieure, etc. ;
- le deuxième groupe de tests nationaux vise surtout à piloter et évaluer les établissements et/ou le système éducatif dans son ensemble. Le pilotage et l'évaluation renvoient ici au mouvement de collecte et d'analyse d'informations afin de contrôler les performances obtenues par rapport à des objectifs et de prendre des mesures rectificatives en cas de besoin. Les résultats de ces tests nationaux servent d'indicateurs de la qualité de l'enseignement et peuvent donc conduire à des choix sur les restructurations d'établissements ou les modulations de programmes ;
- le troisième groupe de tests nationaux a pour principal but de contribuer au processus d'apprentissage des élèves à titre individuel en identifiant leurs besoins spécifiques et en préconisant des opérations de suivi ciblées. On retiendra ici l'idée d'évaluation formative difficilement séparable du contexte de la classe ou de l'établissement.

Sur le mode d'observation, on retrouvera aussi des méthodes de mesure différentes, entrevues pour les enquêtes internationales. Si le test en fin de période scolaire domine, on constate également l'utilisation de la méthode de valeur ajoutée, qui vise à mesurer le progrès des élèves entre une observation en début et en fin de période scolaire (souvent une année) afin de directement relier le progrès des apprentissages (valeur

ajoutée) aux moyens mis en œuvre. Une variante croisant la démarche de valeur ajoutée et l'évaluation sommative reviendra à observer les progrès des élèves au cours de l'ensemble d'un cycle scolaire. On évoque alors une enquête en suivi de cohorte, notion intimement liée au développement des mesures répétitives de l'évaluation. Le problème que posent ces enquêtes concerne l'inertie des effets : s'ils sont plutôt neutres au niveau de l'élève, les effets maître et école seraient, quant à eux, relativement biaisés, car il semble délicat de lier la variation de la performance des élèves d'une classe entre le début et la fin d'année, au seul maître de l'année. Ainsi, Rothstein (2010) montre aisément, pour la valeur ajoutée du groupe à l'année t, un large impact lié aux caractéristiques des maîtres qui ont enseigné le groupe les années antérieures. Dans les exemples d'enquêtes nationales qui suivront, nous verrons que les évaluations se concentrent de plus en plus sur des moments précis du cursus.

De fait, surtout dans une logique d'évaluation formative, le travail pédagogique de l'enseignant, qui consiste à évaluer la participation quotidienne des élèves en classe, leurs devoirs, les tests et travaux écrits et oraux, répond déjà à ce principe d'évaluation qui vise à contrôler et à améliorer le processus pédagogique. Dans l'histoire de l'éducation européenne, on a remarqué que des consignes liées au programme avaient pour objet d'établir une certaine normalisation dans les écoles des devoirs et tests proposés aux élèves. Par ailleurs, surtout dans les pays du nord de l'Europe, le travail des équipes pédagogiques dans les écoles est plus complet, à la fois pour arbitrer devoirs et tests de connaissances mais aussi, et surtout, pour assurer une cohérence d'appréciation entre les classes. Dans des systèmes proches de ceux de l'Europe, Vaniscotte (1996) montre que certains, plutôt nordiques, qui n'avaient pas développé la fonction d'inspection, ont été plus prompts à standardiser les modes d'évaluation des connaissances acquises.

Une dernière dimension à évoquer est celle de l'indépendance du système d'évaluation. Souvent, en particulier pour des pays décentralisés où plusieurs secteurs de l'éducation coexistent, l'activité de certification des acquisitions scolaires est dévolue à une autorité autonome et indépendante. À l'origine en charge de l'organisation des examens nationaux, cette autorité a naturellement assumé la responsabilité de l'activité d'évaluation.

1.2.2. *Quelques grandes options des pays clefs*

Nous rechercherons ici à présenter rapidement les grandes caractéristiques des évaluations des élèves au primaire, notre but n'étant pas l'exhaustivité ni la lecture parallèle suivant les pays, mais plutôt celui d'appréhender comment, dans des pays où la culture d'évaluation s'est répandue en premier, des points fixes caractérisant les méthodes d'évaluation se sont forgés.

États-Unis

Le *Scholastic Aptitude Test* (SAT) était, à sa création (1926), un outil destiné à favoriser l'égalité des chances par la méritocratie. Il s'apparente à un test national, essentiellement sous forme de questionnaire à choix multiples (QCM), passé par les lycéens, et qui vise à établir un classement et une répartition d'étudiants potentiels dans les universités. Il reste plus un outil de filtrage des individus qu'un outil d'évaluation du système. Son caractère méritocratique est souligné par Goastellec (2003), qui montre la nécessité de répondre à la course à l'excellence de certains collèges universitaires dans un contexte de forte différenciation.

Dans les années 1930, une ébauche d'évaluation du programme et des acquisitions fondamentales se met en place aux États-Unis. Sous l'influence de R. Tyler, un système d'évaluation des élèves américains, le *National Assessment of Educational Progress* (NAEP), fait l'objet d'un rejet par les États, qui ne veulent pas d'ingérence fédérale afin de préserver leur autonomie de politique éducative énoncée dans la Constitution. Le travail est resté informel et du domaine de la pure recherche pédagogique, avant son émergence en 1964 sous l'administration Johnson. Le NAEP vise à mettre en place des évaluations des élèves dans les compétences de base principales : anglais, mathématiques, sciences, histoire-géographie, instruction civique. Il concerne tous les États, avec un dessin d'échantillonnage adéquat et permet de comparer les compétences et de fournir une vision dynamique (dans le temps) des progrès d'une cohorte/classe d'âge. Avec le temps s'est instauré un consensus assez large entre les administrateurs de l'éducation, les parents d'élèves, les syndicats d'enseignants sur l'utilité de cet outil d'évaluation.

En 1966, l'enquête *Equality of Educational Opportunity Survey* (EEOS), dirigée par Coleman à partir des toutes premières enquêtes NAEP, indique que les différences constatées à l'entrée de l'école se retrouvent à la sortie. Il apparaît aussi que l'école renforce les désavantages rencontrés par certaines minorités stigmatisées dans l'acquisition de savoirs. Le rapport montre aussi que de nombreux établissements ont sciemment laissé les meilleurs élèves se concentrer dans certains établissements. L'école américaine n'a donc pas le pouvoir de réduire, entre élèves, les différences liées à leur origine sociale. Quelques années après la publication du rapport Coleman, le NAEP – actuellement le *Nation's Report Card* – est mis en place afin de recueillir, en suivant des élèves durant leur scolarité de base, des « données sur le cheminement éducatif des étudiants américains ». Le détail des évolutions depuis la période initiale est donné par Mullis et Jenkins (1990). Il est utilisé dans la conduite des politiques d'éducation comme référence par chaque acteur, même si la Loi NCLB a réactivé

certaines peurs liées à une trop grande intrusion du pouvoir fédéral. Pour respecter les volontés de déconcentration, l'organisation des tests NAEP est en grande partie administrée par les États. En 1988, le Congrès fédéral crée, *via* le *Hawkins-Stafford Act,* le Conseil de direction d'évaluation nationale *(National Assessment Governing Board, NAGB),* qui élabore la politique du NAEP, choisit les domaines qui doivent être évalués, détermine les plans de sondage et les instruments d'évaluation, leur utilisation et les analyses de données.

Depuis la fin des années soixante, le NAEP a évalué les échantillons nationaux des élèves de 9, 13 et 17 ans et, depuis 1983, dans une logique de cohortes, des élèves selon leur niveau et leur âge. Plusieurs pôles d'intérêt comme la lecture, les mathématiques, les sciences et l'écriture sont évalués tous les 4 à 5 ans, d'autres disciplines, tous les 6 à 8 ans. Le plan de sondage permet une représentativité des États, des types d'établissements et des principales minorités. D'après Gipps et Murphy (1994), le modèle du NAEP est recommandé à l'OCDE, dans les années 1970, pour établir des outils de comparaison internationale plus systématiques, en plus des évaluations nationales de l'IEA. Ce n'est que quinze années plus tard que sera initié le programme PISA, plus dans l'optique d'une évaluation des compétences acquises que de validation de l'assimilation du programme.

Trois autres exemples

L'Australie est perçue comme l'un des pays en pointe dans les méthodes évaluatives des apprentissages. Toutefois, comme ceci vient d'être évoqué, même si ce pays est d'organisation décentralisée, les tendances actuelles du système d'évaluation vont dans le sens d'une mesure homogène au niveau national afin de mieux percevoir les écarts existant sur le territoire. L'objectif est ici de se servir de l'évaluation comme d'un outil d'intervention pour entamer des réformes ou allouer des moyens aux écoles ou, plus globalement, aux districts scolaires. Ce pays fortement déconcentré se caractérise par des évaluations en littératie et en numératie initiées il y a près de 30 ans. Le test national de lecture, d'expression écrite, de conventions linguistiques (orthographe, grammaire et ponctuation) et de numératie est passé par les élèves des grades 3, 5, 7 et 9 ; il est organisé, chaque année, au cours de la deuxième semaine de mai. Des évaluations en instruction civique et en technologies de l'information et de la communication (TIC) sont par ailleurs réalisées de manière sporadique en grade 6. En 2008, le Programme national d'évaluation en littératie et numéracie (*National Assessment Programme – Literacy and Numeracy,* NAPLAN) tend à faire converger plus encore les diverses initiatives évaluatives des provinces en un programme national : une commission nationale valide pour tout le territoire les tests uniques, qui constituent

un dénominateur commun du programme de chaque État. Ce test est largement exploité pour donner aux parents, d'une part, et à l'équipe pédagogique, d'autre part, une idée du niveau individuel d'acquisition et de celui de la classe. Ces résultats font l'objet d'une comparaison systématique aux niveaux provincial et national en termes d'écart par rapport à la moyenne nationale. Les élèves ou les écoles enregistrant de faibles performances à ces tests sont ainsi signalés pour déclencher des actions de remédiation au niveau du district scolaire (en général sur les questions de gestion ou de trop grande dispersion des résultats dans une école) ou de la province (lorsqu'une grande partie des écoles d'un district est concernée par des questions de fortes inégalités de résultats interécoles).

La Nouvelle Zélande se place, quant à elle, dans une logique beaucoup plus de périphérie vers le centre. L'évaluation des acquisitions comporte une série de tests dont l'objet premier est, pour l'enseignant, de mieux apprécier la diversité de sa classe et, éventuellement, de signaler des difficultés. Les tests du primaire y sont à la fois fréquents et spécialisés :

- *School Entry Assessment :* il s'agit plus d'un outil d'évaluation pour identifier les capacités en littératie de base des enfants qui entrent en primaire ;
- *Running Records :* ce test standardisé évalue les progrès en lecture de l'élève de grade 3 ; là aussi il s'agit plus d'un outil destiné à suivre la tendance des progrès individuels et des éventuels blocages ;
- *6-Year Net Observation Survey :* ce test, passé par des élèves âgés de 6 ans, (en général en seconde année) combine des exercices de reconnaissance de sons, mots et de syllabes, auxquels s'ajoutent des éléments de position spatiale et d'orientation entre texte et illustration ;
- *Progressive Achievement Tests* (PAT) évalue, en 4e année, l'acquisition du programme ; ce test est largement exploité pour donner aux parents et à l'équipe pédagogique une idée du niveau individuel d'acquisition et de celui de la classe, ainsi que ses dispersions. Ces résultats font l'objet de comparaisons systématiques aux niveaux local et national dans une logique proche du NAPLAN australien ;
- *Supplementary Tests of Achievement in Reading* (STAR) : essentiellement conçu pour évaluer l'acquisition de la lecture en 3e année, ce test permet d'identifier les difficultés individuelles.

Si ces cinq tests font l'objet de construction d'échantillons représentatifs et de traitements longitudinaux pour mesurer les tendances et leurs dispersions, ainsi que de fichiers de recherche et de traitements secondaires, seul le PAT correspond pleinement

à une évaluation nationale, les quatre autres tests consistant plus en des outils d'aide pédagogique et de diagnostic.

Nous verrons dans cette étude que les enquêtes nationales dérivent d'un besoin d'évaluation qualitative des apprentissages des élèves. Dans de nombreux cas, l'évaluation a été un produit conjoint de la certification. En effet, surtout dans les systèmes anglophones, il existait au niveau national un organisme de certification du primaire pour accéder au secondaire. Les premières évaluations ont donc été une exploitation secondaire de cette mesure incluant l'évaluation des tendances et des segmentations des résultats par strates (garçons/filles, régions, zones urbaines/rurales). Dans d'autres cas, la mise en place de l'évaluation est venue de la participation à une opération internationale (comme le MLA) ou de demandes externes, à l'exemple d'opérations d'évaluation liées à la mise en place d'un projet éducatif. On doit reconnaître que les « bilans pays » demandés par l'UNESCO en 2000 pour la conférence EPT ont influencé ce schéma. Les enquêtes comportent (ou pas) le recueil d'informations de contexte (moyens de l'école, contexte socioéconomique des élèves). À l'inverse, rares sont les pays où l'évaluation au niveau primaire a découlé de la mise en place d'un outil diagnostique sur le degré d'assimilation du programme. Le PAT néozélandais ou le système français (*cf.* encadré 2) sont des exemples assez limités.

Encadré 2 *France : une évaluation de tous les élèves à différents niveaux du système*

En France, l'ensemble des élèves de CE1 (2e année du primaire) et de CM2 (5e et dernière année du primaire) sont évalués en français et en mathématiques depuis près de 15 ans, *via* des tests nationaux effectués dans chaque école. Ces évaluations situent les acquis de chaque élève par rapport aux objectifs définis dans les programmes. À partir de ce constat, les enseignants apportent une aide personnalisée aux élèves qui en ont besoin. Les parents sont informés de ces résultats. L'évaluation est donc, dans ce cas, un diagnostic à partir d'un test sur la mesure des compétences attendues. Une synthèse nationale est effectuée et contribue au pilotage du système éducatif. Les résultats globaux et anonymes de la France entière, des académies et des départements sont rendus publics. Les résultats publiés sont calculés sur la base d'une analyse statistique réalisée par le ministère à partir d'un échantillon représentatif. Ils permettent d'effectuer une comparaison avec les résultats de l'année précédente ; ces bases sont utilisées à des fins d'analyses détaillées et chronologiques.

1.2.3. *Les enquêtes nationales sur les acquis des élèves*

Le développement des enquêtes nationales a permis aux ministres de l'Éducation de décrire le niveau national de la performance scolaire, notamment dans les domaines clés, et de comparer les niveaux de différents sous-groupes (garçons/filles, groupes ethniques, populations rurales/urbaines, écoles publiques/privées…). Elles ont également pour objectif de voir dans quelle mesure le niveau d'acquisition des élèves augmente ou diminue dans le temps. Elles ont donc un aspect diagnostique pouvant permettre, dans l'absolu, l'évaluation d'une modification du programme. L'exemple le plus caractéristique est lié à la mise en œuvre de la loi NCLB aux États-Unis (2000), où ces enquêtes servent à identifier les groupes vulnérables et à donner un indicateur de réalisation sur les objectifs et normes d'assimilation des programmes scolaires.

Les enquêtes régionales et internationales ont pour principal objectif de trouver des instruments de tests acceptables pour tous les pays participants. Or, elles ne parviennent pas toujours à expliquer la variabilité des performances à l'intérieur d'un pays. Dans ce sens, il est intéressant de noter que deux pays de structure fédérale, le Canada et la Suisse (*cf.* Encadré 4), utilisent le test PISA en pratiquant un échantillonnage afin de permettre une exploitation infranationale des résultats, en vue de piloter l'application aux territoires de la politique éducative. Les enquêtes nationales permettent de mieux appréhender les différences de performance qui existent au sein des pays, notamment du fait d'un échantillonnage plus large et du contenu du test plus ancré sur le système éducatif national. L'objectif principal de l'enquête nationale est de décrire le niveau d'acquisition des élèves du curriculum préalablement défini par le ministère de l'Éducation (ou plus généralement l'administration responsable de le définir). L'enquête nationale peut soit concerner l'ensemble de la population, soit une partie seulement, choisie selon des critères de représentativité. Elle est souvent accompagnée de questionnaires distribués aux enseignants, aux parents, ou aux directeurs d'écoles.

Un travail de synthèse des différentes évaluations nationales dans quatre régions du monde a été débuté par Encinas-Martin (2006) pour le compte de l'UNESCO. Prolongé par Benavot et Tanner (2007), ce travail représente aujourd'hui la plus grande base de données disponible concernant les enquêtes nationales.

Les informations des enquêtes nationales proviennent de différentes sources. Les sites Internet des gouvernements, des agences internationales et les sites dédiés aux évaluations ont constitué les sources principales pour les évaluations les plus célèbres. Benavot et Tanner (2007) ont ensuite demandé aux experts de plusieurs pays de

confirmer la présence d'évaluations nationales et de les commenter, afin d'obtenir des informations supplémentaires sur celles-ci. Les différents rapports et documents présents sur Internet ont également été utilisés dans une moindre mesure. Enfin, l'équipe de l'*Éducation for All Global Monitoring Report* (EFA GMR, *Rapport de suivi de l'Education pour tous*) crée, depuis plusieurs années, des commissions chargées de produire des études sur les évaluations nationales. Celles-ci ont été utilisées pour analyser quelques résultats saillants. Un rapport de la Banque mondiale a également présenté l'ensemble des évaluations nationales au sein des pays d'Amérique latine (Vegas et Petrow, 2008). Enfin, le Bureau international de l'éducation tente, dans le cadre de la 7e actualisation (depuis 2010), d'intégrer la présence (ou l'absence) d'évaluations nationales dans les questionnaires distribués aux différents pays. à ce jour, les informations ne sont cependant pas disponibles pour l'ensemble des pays.

Dans notre étude, nous tentons d'agréger l'ensemble des informations disponibles à partir de ces sources afin de compléter le travail initial effectué par le *Global Monitoring Report 2008* de l'UNESCO. Les résultats sont présentés dans les tableaux 10 à14. À notre connaissance, il s'agit de la base la plus complète existant à ce jour sur les évaluations nationales. Cependant, il convient de préciser que des incohérences ont été détectées entre les différentes sources utilisées. Par conséquent, des erreurs peuvent apparaître dans les données présentées. Ce travail demande à être actualisé dans le futur. Par ailleurs, les informations concernant la plupart des pays d'Europe orientale et d'Asie sont manquantes à ce jour. Les données présentées dans ce chapitre sont par conséquent susceptibles de changer à l'avenir.

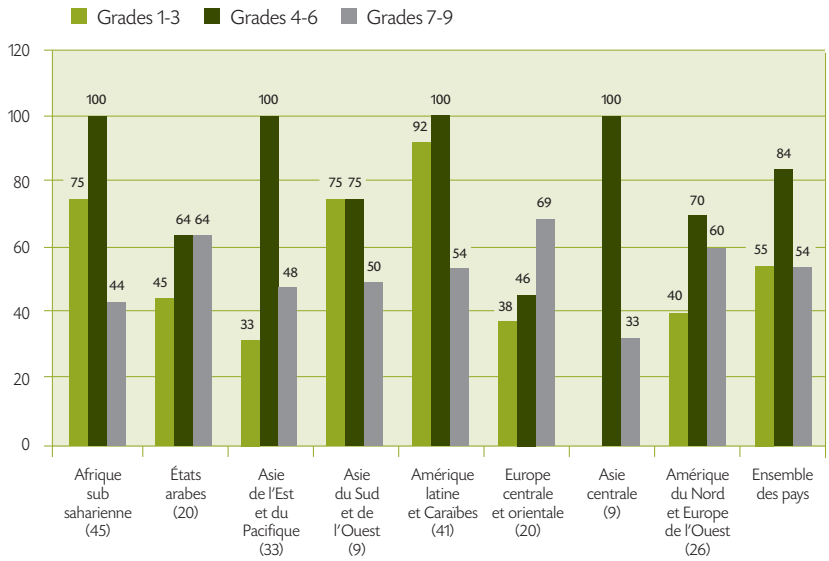
Résultats du recensement des enquêtes nationales

Quelques résultats intéressants de Benavot et Tanner (2007) sont à souligner :

- le nombre de pays ayant mené une évaluation nationale a augmenté depuis le début des années 1990 : sur la période 1995-1999, seulement 65 pays avaient mené une telle évaluation, tandis que ce chiffre est passé à 111 pays sur la période 2000-2006 ;
- ce sont essentiellement les pays d'Amérique du Nord et d'Europe occidentale qui ont effectué le plus grand nombre d'évaluations nationales (77 % d'entre eux sur la période 2000-2006), contre près de 36 % pour les pays d'Afrique subsaharienne ;
- près de 8 pays développés sur 10 ont mené une telle évaluation sur la période 2000-2006, contre environ la moitié des pays en développement. Cependant, on observe un essor récent de la participation des pays en développement à des enquêtes nationales ;

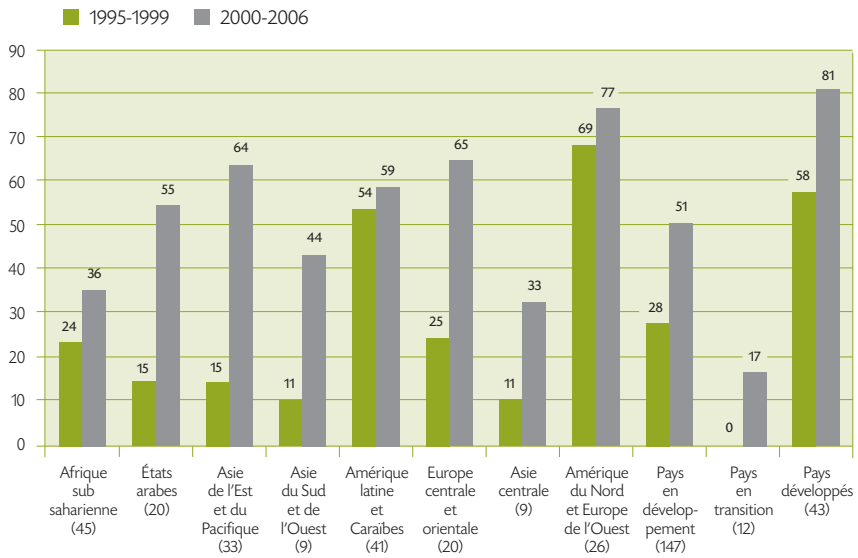
- la proportion des évaluations nationales a augmenté dans la totalité des régions, en particulier dans les pays arabes, où elle passe de 15 à 55 % des pays ;
- les domaines testés ont été le plus souvent les mathématiques (93 % des cas) et le langage (92 % des cas). Près du tiers des pays a également évalué les sciences, deux cinquièmes les sciences sociales (38 %), et un cinquième les langues étrangères (21 %) ;
- les évaluations nationales se focalisent davantage sur les grades 4 à 6 que sur les grades 1 à 3 ou 7 à 9. Entre 2000 et 2006, 84 pays ont conduit au moins une enquête dans les grades 4 à 6, contre seulement 55 pays pour les grades 1 à 3 et 54 pour les grades 7 à 9.

Graphique 1 *Proportion de pays ayant effectué au moins une enquête nationale, par grade et par région du monde (2000-2006)*



Source : Benavot et Tanner (2007).

Graphique 2 *Proportion de pays ayant conduit au moins une évaluation nationale entre 1995-1999 et 2000-2006, par région et niveau de développement*



Source : Benavot et Tanner (2007).

On retrouve, au sein des ENAE, des caractéristiques similaires :

- les domaines de savoirs/compétences testés sont le plus souvent la lecture et les mathématiques, bien que le domaine des sciences tende à se généraliser depuis quelques années ; d'autres domaines sont également testés, mais moins systématiquement (langue étrangère, éducation civique, etc.) ;
- le niveau testé est le plus souvent la scolarité obligatoire, dont l'école primaire. Certaines évaluations concernent parfois le niveau secondaire. Cependant, il n'existe pas de grade spécifique qui soit évalué de manière récurrente. Ceci est dû à deux facteurs :

  – le premier est lié à la différence de durée de la scolarité du primaire et du secondaire selon les pays ;

  – le second concerne davantage le manque de concertation entre pays afin de partager leur expérience dans le domaine de l'évaluation. Comme celles-ci débutent le plus souvent sous la pression des bailleurs de fonds internationaux ou régionaux,

les choix des grades sont le plus souvent effectués sans critères scientifiques avérés. Néanmoins, on peut retenir quelques convergences, comme dans le primaire, où l'on note une certaine concentration des évaluations autour du 3e ou du 4e grade (année), qui concernent les apprentissages fondamentaux des compétences de bases en lecture/compréhension (littératie) ou calcul (numéracie). Dans les évaluations des acquisitions en fin de primaire, une source de données privilégiée vient de l'exploitation du résultat des examens nationaux, lorsque ceux-ci existent, d'admission dans l'enseignement secondaire. À l'évidence, il existe sur ceci un biais venant du fait que les élèves qui ne souhaitent pas poursuivre en secondaire ne passent pas cet examen.

Les différences existent aussi au sein des enquêtes nationales entre les pays :

- la fréquence des tests est très variable : dans certains pays, les ENAE ont lieu tous les ans, tandis que la fréquence est irrégulière dans d'autres pays (souvent dépendants de l'aide internationale pour financer ces tests). Souvent, pour ne pas perturber les rythmes scolaires par trop de tests, seul un grand domaine de compétences (lecture ou calcul) est testé sur une année. Ainsi, chaque domaine de compétences étant testé sur une seule année, se créent des cycles de deux à quatre ans, où toutes les compétences sont évaluées. Malheureusement, les évaluations qui permettent de suivre l'évolution de la performance des élèves dans le temps sont actuellement très rares. Mis à part quelques pays tels que le Chili ou le Mexique, la majorité des pays ayant recours aux évaluations ne recherche pas systématiquement une telle approche comparative. À l'inverse, les enquêtes d'évaluation basées sur une exploitation secondaire de l'examen national de fin de primaire ont assez vite fait l'objet d'exploitation tendancielle. Souvent, de telles expériences ont été menées dans des pays à forte tradition de déconcentration du secteur de l'enseignement. Cet examen à l'échelle du pays est un critère pour observer l'homogénéité des résultats ;
- l'institution qui coordonne le test peut différer selon les pays, ce qui peut avoir des conséquences importantes sur la portée de l'évaluation. Souvent, c'est le ministère de l'Éducation qui gère le test, ou un organisme d'État indépendant, mais ce peut être aussi des centres de recherche ou encore des bailleurs de fonds. En effet, avec l'accroissement récent de l'intérêt des bailleurs de fonds pour la qualité de l'éducation, les pays sont incités à recourir aux évaluations nationales et internationales. La délégation de l'évaluation peut réduire l'impact de celle-ci sur la politique éducative ;
- le caractère volontaire ou non de la participation d'une école au test : lorsque le test est facultatif, le refus de certaines écoles d'y prendre part peut renforcer les biais d'estimation et remettre en cause la nature même de l'évaluation. Il faut aussi

s'interroger sur le caractère décisif d'une évaluation *(high takes versus low stakes)* : si elle n'a pas de finalité décisionnelle, l'évaluation sera le plus souvent acceptée par les enseignants ainsi que les élèves. Cependant, dès lors que la finalité de l'évaluation est de piloter efficacement le système éducatif, les enseignants, ou encore les familles d'élèves, peuvent tenter de bloquer le processus d'évaluation.

Bien que la plupart des pays développés effectuent des évaluations nationales depuis plusieurs décennies, c'est surtout dans les années 1990 que d'autres pays ont disposé des compétences requises pour le lancement de telles évaluations dans un grand nombre de pays. En Amérique latine, par exemple, le développement rapide des ENAE qui a eu lieu durant les années 1990 a permis de fournir des analyses intéressantes d'évaluations de réformes éducatives (Rojas et Esquivel, 1998). C'est d'ailleurs aujourd'hui ce continent qui est la région la plus avancée en termes d'évaluation des élèves dans les pays en développement. Au contraire, les pays d'Asie en sont encore au stade expérimental ; ceci s'explique notamment par l'absence d'une enquête régionale sur les acquis des élèves au sein des pays de cette région. Les pays arabes occupent, quant à eux, une position intermédiaire, notamment du fait de leur participation récente aux évaluations internationales qui leur permettent de souligner la nécessité de recourir également aux évaluations nationales. Enfin, les pays d'Afrique subsaharienne ont eu peu recours aux évaluations nationales au cours des dix dernières années et c'est surtout sous la pression de la Banque mondiale et des autres bailleurs de fonds qu'ils se sont tournés vers ce type d'évaluation. Comme on peut le constater dans le tableau 1, l'hétérogénéité qui règne sur le public concerné souligne le caractère expérimental de telles évaluations.

Nous présentons, dans ce même tableau, une classification des pays selon leur état d'avancement concernant leurs pratiques en matière d'évaluation nationale. Les pays sont classés en quatre catégories :

- *les pays en retard* sont ceux pour lesquels on ne dispose d'aucune évaluation nationale ou d'aucune information. Ces pays sont le plus souvent situés en Asie ou au Moyen-Orient et appartiennent généralement à la catégorie des pays dits « fragiles ». Plusieurs pays africains figurent également parmi cette catégorie tels que l'Angola, la Somalie ou encore la Tanzanie ;
- *les pays au stade préliminaire* sont ceux pour lesquels seules une ou deux évaluations ont été réalisées jusqu'à aujourd'hui, sans analyse particulière. Ces pays sont au nombre de 42. De nombreux pays africains figurent dans cette catégorie, dont la Gambie, le Lesotho ou encore le Mali. Ces pays ne sont pas encore entrés dans une phase systématique d'évaluation ;

- *les pays au stade intermédiaire* ont une certaine culture de l'évaluation, mais sans analyse approfondie des résultats et sans possibilité de fixer des objectifs précis. Cette catégorie regroupe 24 pays, dont un certain nombre se trouvent en Amérique latine, tels que la Bolivie ou encore l'Uruguay. Mis à part le Bénin et Madagascar, qui affichent depuis plusieurs années une volonté d'évaluer le niveau d'acquisition de leurs élèves, peu de pays africains sont présents ;
- *les pays au stade avancé* sont ceux pour lesquels plusieurs évaluations ont été déjà effectuées et un travail de suivi du niveau des acquis a été mis en place. Ces pays sont peu nombreux et la plupart se trouvent en Amérique latine. Parmi les pays africains, on peut citer l'Afrique du Sud, le Burkina Faso, le Malawi ou encore la Zambie (les pays anglophones d'Afrique subsaharienne apparaissent, dans ce domaine, plus en avance que les pays francophones).

Tableau 1 *Évaluation nationale : classification des pays selon leur état d'avancement*

Pays en retard	Pays au stade préliminaire	Pays au stade intermédiaire	Pays au stade avancé
Afghanistan ; Angola ; Arménie ; Azerbaïdjan ; Belize ; Cambodge ; Cap Vert ; Chine ; Congo ; Corée du Nord ; Guinée équatoriale ; Gabon ; Géorgie ; Guinée-Bissau ; Haïti ; Hong Kong ; Iran ; Irak ; Kenya ; Koweït ; Libye ; Maurice ; République centrafricaine (RCA) ; Rwanda ; Sierra Leone ; Somalie ; Soudan ; Suriname ; Swaziland ; Syrie ; Tadjikistan ; Tanzanie ; Tchad ; Thaïlande ; Timor-Leste ; Togo ; Tunisie ; Turkménistan ; Ouzbékistan ; Zimbabwe.	Algérie ; Arabie Saoudite ; Bahreïn ; Barbade ; Burundi ; Cameroun ; Comores ; RDC ; Costa Rica ; Côte d'Ivoire; Djibouti ; Emirats Arabes Unis ; Erythrée ; Éthiopie ; Gambie ; Grenada ; Guinée ; Honduras ; Inde ; Îles Cook et Fidji ; Liban ; Lesotho ; Liberia ; Malaisie ; Maldives ; Mali ; Mauritanie ; Maroc ; Mozambique ; Myanmar ; Namibie ; Népal ; Nicaragua ; Niger ; Nigeria ; Nouvelle Zélande ; Ouganda ; Papouasie Nouvelle Guinée ; Paraguay ; Sénégal ; Singapour ; Sri Lanka ; Venezuela ; Vietnam ; Yémen.	Bangladesh ; Bénin ; Bhoutan ; Bolivie ; Botswana ; Colombie ; Cuba ; Égypte ; Ghana ; Guatemala ; Jamaïque ; Kirghizistan ; Laos ; Madagascar ; Mongolie ; Oman ; Pakistan ; Panama ; Philippines ; République dominicaine ; Samoa ; Ste Lucie ; Trinidad et Tobago ; Uruguay.	Afrique du Sud ; Argentine ; Brésil ; Burkina Faso ; Chili ; Équateur ; El Salvador ; Guyana ; Inde ; Indonésie ; Jordanie ; Kazakhstan ; Malawi ; Mexique ; Pérou ; Qatar ; Zambie.

Source : auteurs, à partir des études citées.

1.3. Les enquêtes internationales sur les acquis des élèves

1.3.1. *L'enquête TIMSS*

La première mesure des acquis au niveau individuel permettant une comparabilité internationale a été initiée au début des années 1960 par l'Association internationale pour l'évaluation du rendement scolaire (*International Association for the Evaluation of Educational Achievement,* IEA). L'IEA a réalisé plusieurs enquêtes pluriannuelles dans des domaines très variés : les mathématiques, les sciences et la lecture, mais aussi auprès des écoles préprimaires (dans 14 pays, entre 1988 et 1995), ou encore sur le sujet de l'informatique à l'école (dans 20 pays, entre 1988 et 1992).

Encadré 3 *Enquête sur l'évaluation des apprentissages des élèves du primaire en Côte d'Ivoire (2001-2002)*

Une enquête sur l'évaluation des apprentissages des élèves du primaire en Côte d'Ivoire a été réalisée en 2001-2002 par le ministère de l'Éducation nationale, en collaboration avec le Service de développement et d'évaluation de programmes de formation (SEDEP, Liège) et l'Institut national d'étude et d'action pour le développement de l'éducation (INEADE, Dakar).

Deux objectifs principaux ont été assignés à cette enquête : tout d'abord, la collecte d'indications concernant la réussite des élèves aux trois niveaux de l'enseignement primaire (CP2, CE2 et CM2), mais aussi l'identification des facteurs de réussite des élèves. Les tests ont ainsi concerné le français écrit et les mathématiques dans les trois niveaux. Trois types de questionnaires ont été distribués et concernaient les directeurs d'école, les maîtres et les élèves des trois classes. La méthode d'échantillonnage reprenait la méthodologie utilisée par le PASEC, à savoir une stratification à deux étapes (d'abord les régions, puis les écoles au sein de celles-ci). Au total, cent écoles dans onze directions régionales du ministère de l'Éducation nationale ont été choisies de manière aléatoire. L'échantillon final, qui comprenait un peu plus de 7 000 élèves, était constitué de 88 classes de 2e année d'enseignement primaire (2 445 élèves), 88 classes de 4e année (2 353 élèves) et 88 classes de 6e année (2 368 élèves). Dans chaque matière testée, deux questionnaires ont été administrés.

Ainsi, à la différence du PASEC, les élèves sélectionnés ont été soumis à un seul test. Un élève testé en français ne l'était plus en mathématiques et une distinction était faite selon le questionnaire (A et B). Ainsi, au sein de chaque classe, quatre groupes disjoints d'élèves ont été constitués : le premier groupe a été testé uniquement en mathématiques pour la forme A, le second testé en mathématiques mais pour la

•••

•••

forme B, le troisième l'était en français pour la forme A et enfin le dernier, en français, pour la forme B. Au total, 28 élèves choisis dans chaque classe étaient testés uniquement dans l'une des épreuves et l'une des formes.

Les principaux résultats sont présentés dans le tableau 2. Ceux-ci sont, dans l'ensemble, très bas. Bien qu'une légère progression puisse être observée du CP2 au CM2, elle n'est pas significative : les pourcentages moyens de réussite sont le plus souvent inférieurs à 50 %. Par ailleurs, les pourcentages d'élèves atteignant ou dépassant le seuil de maîtrise fixé à 70 % demeurent très faibles dans pratiquement tous les cas.

Tableau 2 *Degré de maîtrise d'élèves (en %) et résultats moyens par domaine selon les épreuves*

Niveau	Non-maîtrise (score < 50 %)	En voie de maitrise (50 % <= score < 70 %	Maîtrise score>=70 %	Résultats moyens (Écart-type)
Français				
CP2	71,5	22,1	6,3	40,9 (17,3)
CE2	50,1	32,6	17,3	48,6 (17,8)
CM2	35,8	36,3	27,9	56,8 (17,8)
Mathématiques				
CP2	55,0	33,0	12,0	45,6 (19,2)
CE2	57,0	38,0	5,0	46,4 (14,7)
CM2	56,0	38,0	6,0	56,4 (15,3)

Source : Enquête sur l'évaluation du rendement scolaire, Côte d'Ivoire 2002.

Source : Sika (2011).

Si aujourd'hui les enquêtes internationales portent presque exclusivement sur les mathématiques, les sciences et la lecture, ce n'était pas le cas par le passé. Citons, par exemple, des enquêtes sur la civilité et la citoyenneté, menées entre 1968 et 1973 (*The Study of Civic Education,* 11 systèmes éducatifs participants), entre 1994 et 2002 (*Civic Education Study,* 31 systèmes éducatifs participants) et une enquête qui s'est déroulée en 2009 (*International Civic and Citizenship Education Study,* au moins 39 systèmes éducatifs s'étaient alors déclarés volontaires pour participer à l'enquête). L'évaluation de l'apprentissage des langues étrangères a également été l'une des premières missions de l'IEA avec, dans la période 1968-1973, la thématique de l'ap-

prentissage de l'anglais dans dix pays européens, pour les enfants de 14 ans et les élèves de la classe terminale du secondaire (Lewis et Massad, 1975). Une seconde vague d'enquêtes réalisées entre 1993 et 1995 devait analyser l'enseignement de quatre langues européennes (allemand, anglais, français et espagnol) dans 25 pays comme seconde langue. Faute de financement, l'évaluation des élèves n'a pas été réalisée. Ce fut également le cas pour d'autres thématiques (par exemple les NTIC, ou encore la pratique des langues étrangères).

La première enquête évaluant les mathématiques *(The First International Mathematics Study)* s'est déroulée entre 1963 et 1967 dans 12 pays développés (Angleterre, Australie, Belgique, Ecosse, États-Unis, Finlande, France, Israël, Japon, Pays-Bas, République fédérale d'Allemagne et Suède). Les élèves évalués avaient alors 13 ans ou étaient en dernière année du secondaire [11]. L'enquête qui évaluait les sciences s'est déroulée sur une plus longue période (1968-1972) et s'est focalisée sur la biologie, la chimie et la physique. La population évaluée avait 13/14 ans ou était en dernière année du secondaire. Au total, 19 systèmes éducatifs ont été évalués (Angleterre, Australie, Belgique francophone, Belgique flamande, Chili, Ecosse, États-Unis, Finlande, France, Hongrie, Inde, Iran, Italie, Nouvelle-Zélande, Pays-Bas, République Fédérale d'Allemagne, Suède, Thaïlande). Une synthèse des principaux résultats de cette enquête peut être trouvée dans Walker (1976).

La deuxième enquête sur les mathématiques *(The Second International Mathematics Study)* s'est déroulée entre 1977 et 1981 et a concerné les élèves de 13 ans ainsi que ceux étudiant en dernière année du secondaire (Burstein, 1992). Dix-neuf systèmes éducatifs ont été évalués, parmi lesquels figurent deux pays africains (Nigeria et Swaziland). Au début des années 1980 s'est déroulée la deuxième enquête sur les sciences *(The Second International Science Study),* sur 23 systèmes éducatifs, dont 3 pays africains (Ghana, Nigeria et Zimbabwe) ainsi que 5 autres pays en développement (Chine, Papouasie Nouvelle Guinée, Philippines, Pologne et Thaïlande).

[11] Pour plus d'informations, voir notamment Husén (1967).

Encadré 4 *Exemple d'élargissement d'une enquête internationale vers une évaluation nationale*

La Suisse a élargi l'échantillon requis dans sa participation à l'enquête PISA 2006 (OFS, 2007). Normalement, l'OCDE préconise de sélectionner au minimum 4 500 élèves provenant de 150 écoles au moins. La sélection s'effectue en deux phases :

- dans une première phase, les écoles sont tirées au sort. La probabilité qu'une école soit choisie dépend du nombre d'élèves de 15 ans qui la fréquentent ;
- dans une seconde phase, ce sont les élèves des écoles retenues qui sont choisis aléatoirement.

Les élèves de 15 ans qui participent à l'enquête en Suisse viennent des écoles et des filières suivantes :

- élèves du degré secondaire I (élèves de 7e, 8e, 9e et 10e années),
- écoles de formation générale du degré secondaire II (gymnases/lycées, école du degré diplôme, etc.),
- écoles professionnelle du degré secondaire II.

En plus de cet échantillon, la Suisse a sélectionné un autre échantillon d'élèves de 9e dans le but de procéder à des analyses approfondies au niveau des trois régions linguistiques ainsi que des cantons qui en ont fait la demande. Par conséquent, l'échantillon complet passe de 4 500 élèves et 150 écoles à 12 192 élèves et 510 écoles.

Plus spécifiquement, en Suisse romande, le nombre d'élèves de 9e année interrogés a été augmenté dès 2000 afin de disposer d'informations suffisantes pour permettre l'analyse au niveau de chaque canton et de la région. Alors qu'il existe 26 systèmes éducatifs différents et trois langues officielles dans le pays, les responsables ont pris la décision de saisir l'opportunité d'élargir la population interrogée par l'enquête PISA 2006.

Quelques résultats sont présentés dans le tableau 3. Les responsables de l'enquête PISA 2006 en Suisse ont préféré se baser sur l'ensemble des résultats des élèves de 9e année, indépendamment de leur âge. La comparaison des trois régions linguistiques fait ressortir systématiquement de meilleures performances pour les élèves alémaniques. Les scores enregistrés en Suisse romande et en Suisse italienne ne diffèrent pas sensiblement : les élèves francophones devancent de quelques points les élèves italophones en mathématiques et sur deux des trois échelles de compétence scientifique.

•••

...

Tableau 3 *Analyse régionale : scores moyens obtenus par les élèves de 9e année dans les divers domaines et dans les trois échelles de compétences scientifiques*

Région	Domaine			Echelles de compétences scientifiques		
	Sciences	Lecture	Mathématiques	Identification des questions d'ordre scientifique	Explication scientifique des phénomènes	Utilisation de faits scientifiques
Suisse alémanique	518	503	535	518	516	523
Suisse romande	502	497	528	513	491	511
Suisse italienne	501	496	523	499	497	508
Total	513	501	533	516	509	519

Source : Nidegger (2008).

C'est essentiellement le cycle d'enquête TIMSS qui va se révéler le plus efficace dans l'évaluation des mathématiques et des sciences. Son objectif est d'évaluer le niveau des élèves en mathématiques et en sciences ainsi que de décrire le contexte dans lequel se fait l'apprentissage. Par ce second objectif, les fondateurs de l'enquête TIMSS ont résolument adopté une approche en termes de finalité politique puisque les résultats des élèves sont associés aux différents facteurs utilisés dans le cadre de l'enseignement.

Le contenu des questionnaires est assez varié et des pondérations spécifiques sont attribuées à chaque thématique (on peut citer notamment les nombres, l'algèbre, la géométrie pour les mathématiques ; les sciences de la vie, les sciences physiques, ou encore l'histoire des sciences pour les sciences). L'évaluation des élèves se base essentiellement sur un référentiel commun de connaissances entre les pays. Plusieurs centaines d'items ont été évalués pour déterminer s'ils sont enseignés dans la plupart des pays participants avant d'être insérés dans les questionnaires. Une volonté de

maximiser le nombre d'items standardisés pour tous les pays a été recherchée, ce qui n'exclut toutefois pas la possibilité que, dans certains systèmes éducatifs, certains ne figurent pas réellement au programme.

Les enquêtes ne concernent pas seulement le niveau des élèves en mathématiques et en sciences : outre le questionnaire d'évaluation, d'autres ont été distribués aux différents acteurs du système :

- un questionnaire portant sur les caractéristiques individuelles et familiales de l'élève, qui regroupe des informations spécifiques à l'élève (sa motivation, son degré de fréquentation des bibliothèques), mais aussi relatives à sa famille en général (type d'emploi occupé par ses parents, taille de la ville de résidence, etc.) ;
- un questionnaire destiné aux enseignants qui recense des éléments généraux sur sa classe (taille, disponibilité d'une bibliothèque, d'ordinateurs, etc.) mais aussi sur ses pratiques pédagogiques (temps consacré à la correction des exercices, interactions avec les élèves, etc.) et sur sa formation initiale et continue (diplômes obtenus, formation spécifique à l'enseignement, expérience professionnelle, etc.) ;
- un questionnaire destiné au directeur de l'école, qui regroupe des informations générales sur l'école mais également sur les élèves évalués dans l'enquête (taille de l'école, modes de regroupement des élèves, indisponibilités éventuelles de certaines ressources, etc.) ;
- un questionnaire destiné aux personnels des ministères de l'Éducation afin d'obtenir des informations sur les curricula, leur degré d'application par les enseignants, les types d'évaluations effectuées, etc.

Trois populations différentes ont participé à la première enquête TIMSS, en 1994-1995 :

- population 1 = élèves de grades adjacents regroupant la plupart des élèves de 9 ans (en général grades 3 et 4) ;
- population 2 = élèves de grades adjacents regroupant la plupart des élèves de 13 ans (en général grades 7 et 8) ;
- population 3 = élèves en dernière année du secondaire avec une distinction entre deux sous-populations : (a) élèves ayant passé un test en mathématiques et en lecture, et (b) élèves spécialisés en mathématiques ou en physique ayant passé un test spécialisé.

Encadré 5 *La participation des pays en développement aux enquêtes TIMSS*

En 1994/1995, 45 systèmes éducatifs ont participé à l'enquête TIMSS et ont été évalués dans les populations 1, 2 et 3 (c'est-à-dire aux grades 3-4, 7-8 et en dernière année du secondaire). Certains pays n'ont été évalués que pour une partie des trois types de populations. Parmi les pays participants, on peut notamment recenser un pays africain (Afrique du Sud) et 9 autres pays en développement ou en transition (Bulgarie, Colombie, Iran, Lettonie, Lituanie, Roumanie, Fédération de Russie, République Slovaque et Thaïlande).

En 1999, 38 systèmes éducatifs ont été évalués ; seule la population 2 a été concernée. Parmi les pays participants, trois étaient africains (Afrique du Sud, Maroc et Tunisie) et 16 autres concernaient des pays en développement ou en transition (Bulgarie, Chili, Indonésie, Iran, Jordanie, Lettonie, Lituanie, Macédoine, Malaisie, Moldavie, Philippines, Roumanie, Fédération de Russie, République Slovaque, Thaïlande, Turquie).

Durant l'année 2003, 50 systèmes éducatifs ont été évalués pour les populations 1 et 2. Parmi ces pays, 6 étaient africains (Afrique du Sud, Botswana, Égypte, Ghana, Maroc et Tunisie). Par ailleurs, 25 autres pays en développement ont été évalués (Arménie, Brésil, Bulgarie, Chili, Estonie, Indonésie, Iran, Jordanie, Lettonie, Liban, Lituanie, Macédoine, Malaisie, Mexique, Moldavie, Palestine, Philippines, Pologne, Roumanie, Fédération de Russie, Serbie, République slovaque, Thaïlande, Turquie et Uruguay).

En 2007, la quatrième vague d'enquête TIMSS a inclus 59 pays ou régions du monde. Cette enquête concernait également les populations 1 et 2. Parmi les pays inclus, cinq étaient africains (Botswana, Égypte, Ghana, Maroc et Tunisie). Par ailleurs, certains pays ont participé pour la première fois à une évaluation internationale (Égypte, Kazakhstan, Mongolie, Syrie et Yémen).

Bien que son écho médiatique soit plutôt faible, l'IEA continue de proposer une évaluation en mathématiques et en sciences en dernière année du secondaire. La première vague comparable avait eu lieu en 1995 avec la participation d'environ 22 pays [12]. La deuxième édition de cette enquête n'a concerné que dix pays (Arménie, Iran, Italie, Liban, Norvège, Pays-Bas, Philippines, Fédération de Russie, Slovénie et Suède), dont cinq seulement avaient déjà participé à l'évaluation de 1995 (Italie, Norvège, Fédération de Russie, Slovénie et Suède). La faible participation des pays s'explique sûrement par le fait que, contrairement aux précédentes vagues d'enquêtes, celle-ci s'est déroulée de façon indépendante aux autres grades.

•••

[12] Le nombre de pays varie selon le domaine testé. On compte 22 pays pour le domaine des mathématiques et des sciences, 17 pour le domaine avancé en mathématiques et 18 pour le domaine des sciences physiques.

La dernière série d'enquêtes TIMSS a eu lieu en 2011 avec l'enquête TIMSS 2011 qui a réuni 63 pays et régions du monde[13], dont 7 en Afrique (Botswana, Ghana, Lybie, Maroc, Afrique du Sud, Tunisie et Yémen). Certains pays y participaient pour la première fois (Azerbaïdjan, Lybie), tandis que d'autres étaient présents depuis la première vague (Danemark, États-Unis, Iran, etc.).

À ce jour, quatre enquêtes TIMSS ont été effectuées :

- la première (1994-1995) portait sur 45 systèmes éducatifs et trois populations[14] (grades 3 et 4 ; 7 et 8 ; dernière année du secondaire) ;
- la seconde vague (1999) portait sur 38 systèmes éducatifs, pour le grade 8 seulement ;
- la troisième vague (2003) s'est déroulée dans 50 systèmes éducatifs et pour les grades 4 et 8 ;
- en février 2009, l'enquête TIMSS 2007 a été également finalisée. Celle-ci concernait les grades 4 et 8 et plus de 59 systèmes éducatifs.

La prochaine vague est prévue pour 2011 (les résultats seront publiés en février 2013 ; *cf.* encadrés 5 et 6).

Encadré 6 *Résultats de l'enquête TIMSS 2007*

L'enquête TIMSS 2007 est l'évaluation internationale qui a réuni le plus grand nombre de pays en développement, en comparaison des vagues précédentes et de l'enquête PISA. Les scores TIMSS ont été standardisés de manière à obtenir une moyenne internationale de 500 et un écart type international de 100. Ainsi, les scores doivent être interprétés non pas comme des valeurs absolues, mais plutôt comme des valeurs relatives.

On constate que les pays les plus performants sont le plus souvent asiatiques : 5 affichent les scores les plus élevés en mathématiques. Parmi eux, Taiwan, la Corée du Sud et Singapour ont les scores les plus élevés, avec une performance généralement supérieure à celle de tous les autres pays. Ces trois pays sont suivis par Hong Kong SAR et le Japon, dont les niveaux de performance sont proches ou supérieurs à l'ensemble des autres pays.

[13] Par ailleurs, 14 zones – le plus souvent des États fédéraux – participent à l'enquête TIMSS 2011. Il s'agit par exemple de l'Alabama, et de la Floride aux États-Unis, ou encore d'Abou Dhabi aux Émirats Arabes Unis.

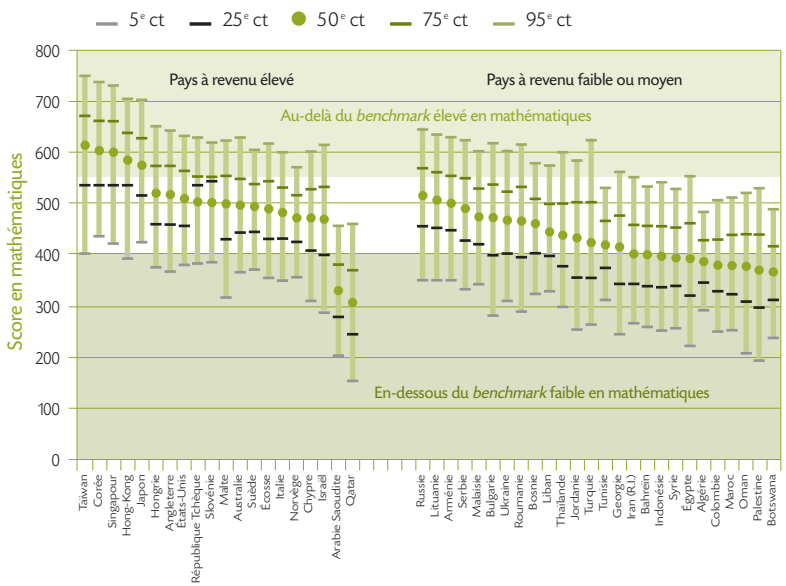
[14] Parfois, certaines provinces canadiennes ou certains États fédéraux des États-Unis ont participé aux enquêtes de l'IEA ; par souci de simplicité, nous n'incluons pas ces régions dans le calcul du nombre de pays participant aux enquêtes.

On note une grande différence dans la performance en mathématiques entre les cinq pays asiatiques et le groupe des quatre pays positionnés derrière les premiers pays (Hongrie, Angleterre, Fédération de Russie et États-Unis). Cette différence s'élève à 53 points entre le Japon (570) et la Hongrie (517).

Au niveau du grade 8, certains pays, tels que l'Égypte, l'Algérie, la Palestine, l'Arabie Saoudite ou encore le Maroc, obtiennent des scores relativement proches et affichent une meilleure performance que le Salvador (387), qui obtient de meilleurs résultats que le Botswana (355), qui dépasse le Qatar (319) et le Ghana (303), ces deux pays obtenant les scores les plus faibles au grade 8.

Graphique 3

Scores et déciles au grade 8 en mathématiques (TIMSS 2007)



« ct » = centiles.
Le benchmark *élevé représente un score supérieur à 550 points, tandis que le* benchmark *faible un score inférieur à 400 points.*
La moyenne internationale est de 500 points

Source : Mullis et al. *(2008).*

1.3.2. *L'enquête PIRLS*

Dans le domaine de la lecture, les enquêtes de l'IEA ont débuté en 1968-1972 avec l'enquête *The Study of Reading Comprehension* qui portait sur 15 systèmes éducatifs, dont 3 pays en développement (Chili, Inde et Iran). Axée sur les élèves de 10 ans, 14 ans et en dernière année du secondaire, cette première enquête sur la lecture a permis d'apprécier la difficulté de mesurer d'une façon standardisée les compétences des élèves dans un domaine fortement lié à la langue du pays considéré. On pourra trouver dans Thorndike (1973) et Walker (1976) des éléments de synthèse de cette première enquête. Par la suite, entre 1985 et 1994, fut lancée l'enquête *The Reading Literacy Study* (RLS) qui a servi d'exemple pour les enquêtes suivantes. L'objectif principal était ici de produire des tests internationaux valides et des questionnaires qui pouvaient être généralisés à l'ensemble des pays participants (voir notamment Elley, 1992 ; Postlethwaite et Ross, 1992). Les données ont été collectées entre 1990 et 1991. Deux types de populations étaient visés : les élèves de 9 et 14 ans. Au total 32 systèmes éducatifs ont participé à l'enquête, parmi lesquels figuraient trois pays africains (Botswana, Nigeria et Zimbabwe) ainsi que cinq autres pays en développement (Indonésie, Philippines, Thaïlande, Trinidad et Tobago et Venezuela ; *cf.* encadré 7).

Encadré 7 *La participation des pays en développement à l'enquête PIRLS*

Dans l'enquête de 2001, 35 systèmes éducatifs ont été évalués. Parmi les pays participants, un seul est africain (Maroc), tandis que l'on recense 13 pays à revenu moyen (Argentine, Bulgarie, Colombie, Iran, Lettonie, Lituanie, Macédoine, Moldavie, Maroc, Roumanie, Fédération de Russie, République Slovaque et Turquie).

Au total, 41 pays ou systèmes éducatifs ont pris part à l'enquête PIRLS 2006. Parmi eux, seuls deux sont africains (Afrique du Sud et Maroc). Au total, 15 des pays en développement ou en transition ont participé à l'enquête PIRLS 2006 (Bulgarie, Géorgie, Indonésie, Iran, Lettonie, Lituanie, Macédoine, Moldavie, Maroc, Pologne, Roumanie, Fédération de Russie, Répu-blique slovaque, Afrique du Sud et Trinidad et Tobago).

De façon conjointe à l'enquête TIMSS 2011, l'évaluation PIRLS a eu lieu la même année. Suivant les pays, trois niveaux d'intégration des deux enquêtes ont lieu : *(i)* le premier teste les mêmes élèves pour les deux enquêtes ; *(ii)* le second teste des classes des mêmes écoles en lecture ou en mathématiques ; *(iii)* dans un troisième cas, les échantillons des deux enquêtes sont indépendants entre les établissements. Regroupant 49 pays, dont trois pays africains (Afrique du Sud, Botswana et Maroc), l'enquête PIRLS reste la seule évaluation internationale de la lecture à l'école primaire.

•••

•••

Une innovation majeure de PIRLS 2011 est de s'adapter aux besoins des pays participants. Ainsi, dans certains pays, les élèves du grade 4 étant encore dans un processus d'apprentissage de la langue – une situation qui pourrait remettre en cause la légitimité d'une telle évaluation –, il est possible d'évaluer les grades supérieurs (5 et 6).

Par ailleurs l'évaluation pré PIRLS est une forme d'évaluation supplémentaire dans la dernière année du primaire (grades 4, 5 ou 6 selon les pays), mais avec un niveau moins difficile. En effet, dans certains pays, comme l'Afrique du Sud par exemple, la quasi totalité des élèves affichant une performance dans le benchmark le moins élevé, on notait une désaffection des pays en développement pour les évaluations internationales comme PIRLS. L'évaluation pré PIRLS consiste donc à vérifier les compétences suivantes : les élèves peuvent-ils reconnaître des mots et des groupes de mots ? Lire des phrases ? De simples paragraphes ? Retrouver de l'information précise dans un texte ? Effectuer des inférences sur des histoires ? Pré PIRLS est ainsi davantage une évaluation diagnostique de l'état du système éducatif primaire dans le domaine des connaissances générales en lecture et doit être considérée comme une introduction à l'évaluation PIRLS à proprement parler. Trois pays ont participé à l'évaluation pré PIRLS (Afrique du Sud, Botswana et Colombie).

L'enquête PIRLS constituera le cycle majeur de l'évaluation de la lecture au niveau primaire. Elle évalue le niveau des élèves en compréhension de l'écrit. Basée sur l'enquête RLS de 1990-1991, PIRLS a été effectuée à deux reprises jusqu'à aujourd'hui (2001 et 2006 – *cf.* encadré 8). Seuls les élèves du grade 4 ont été évalués (*i.e.* d'un âge moyen de 9 ans). L'évaluation concerne deux objectifs de lecture :

- l'objectif de littératie (en anglais *literacy*) : il renvoie renvoie à la lecture qui implique d'imaginer des événements et des objets et de les mettre en mouvement dans un texte ;
- l'objectif d'information (en anglais *informational*), où la lecture sert à acquérir et à utiliser de l'information organisée chronologiquement (par exemple dans des biographies) et/ou de façon logique (par exemple dans des textes de réflexion).

Au total, ont été évalués quatre processus de compréhension de la lecture qui concernent les capacités à : *(i)* effectuer un ciblage et une explicitation d'informations spécifiques ; *(ii)* effectuer des inférences à partir de suites logiques ou chronologiques et à relier des événements entre eux ; *(iii)* interpréter et intégrer des idées et des informations ; *(iv)* examiner et évaluer le contenu, le langage ainsi que les éléments textuels. Comme pour l'enquête TIMSS, afin de réduire le temps de l'évaluation des

élèves, l'examen type a été « découpé » en dix livrets, chaque élève n'en ayant reçu qu'un seul (soit un dixième de l'examen complet). En ayant recours à des méthodes psychométriques, les spécialistes ont ensuite recomposé des scores comparables pour chaque élève. Deux types de questions ont été utilisées : à choix multiples et ouvertes. Par ailleurs, un questionnaire spécifique a été distribué à chaque élève, enseignant et directeur d'école. Les élèves évalués étant assez jeunes (9 ans environ), le questionnaire portant sur leurs caractéristiques individuelles et familiales a été distribué aux parents.

Encadré 8 *Résultats du PIRLS 2006*

L'enquête PIRLS 2006 est la seule évaluation internationale qui teste un grand nombre de pays en lecture au niveau primaire. En comparaison, PISA n'évalue que les élèves de 15 ans. Les scores PIRLS ont été standardisés de manière à obtenir une moyenne internationale de 500 et un écart type international de 100. Ainsi, il faut interpréter les scores non comme des valeurs absolues, mais plutôt comme des valeurs relatives.

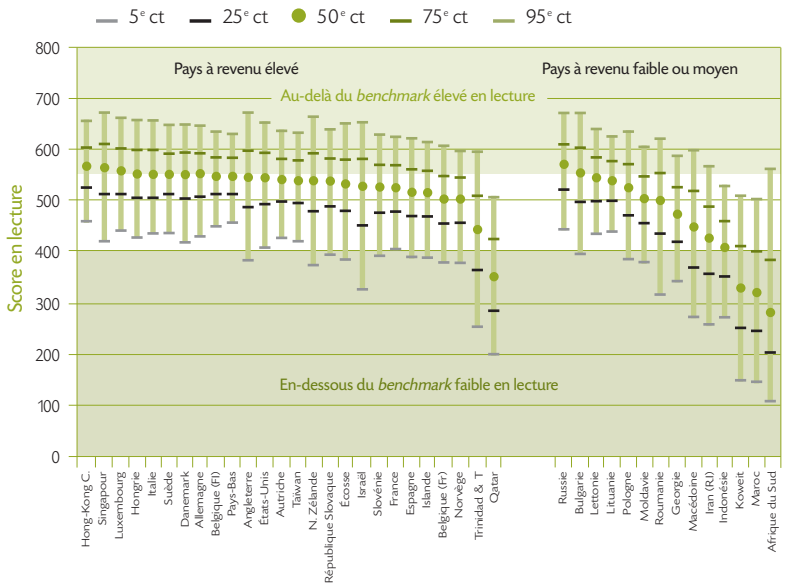
Dans le graphique 4, nous présentons les principaux résultats de l'enquête PIRLS 2006. On constate que les pays les plus performants sont deux pays/régions asiatiques (Hong-Kong et Singapour), et que les pays européens affichent également des performances importantes. Celle du Luxembourg est en effet d'environ 30 points supérieure à celle des États-Unis. Parmi les pays à revenu faible ou moyen, une performance notable vient de la Fédération de Russie, suivie de la Bulgarie : ces deux pays affichent une performance en lecture au moins aussi élevée que celle des pays européens les mieux classés.

Parmi les pays les moins performants, on constate la présence de pays arabes tels que le Qatar, le Koweït ou encore le Maroc. Le résultat le plus faible est obtenu par l'Afrique du Sud avec un score moyen inférieur à 300 points, soit près de 200 points en dessous de la moyenne internationale et un chiffre presque égal à la moitié de celui de la Fédération de Russie ou de Hong-Kong. Pour l'Afrique du Sud, on constate même que plus de 98 % des élèves affichent une performance inférieure à 500 points (en Fédération de Russie, cette proportion n'est que d'environ 20 %) et qu'aucun élève n'a pu obtenir un score supérieur à 600 points (ce qui est le cas de près d'un quart des élèves de Bulgarie). Les scores très bas obtenus par l'Afrique du Sud expliquent en partie pourquoi l'équipe de l'IEA a lancé l'initiative pré PIRLS : mieux expliquer les raisons de performances aussi faibles.

•••

Graphique 4

Principaux résultats de l'enquête PIRLS 2006



« ct » = centiles.
Le benchmark *élevé représente un score supérieur à 550 points, tandis que le benchmark faible un score inférieur à 400 points.*
La moyenne internationale est de 500 points.

Source : Mullis et al. *(2008).*

1.3.3. *L'enquête PISA*

L'OCDE a lancé l'enquête PISA en 1997 pour répondre à la nécessité de disposer de données sur la performance des élèves qui soient comparables au niveau international. Le principe fondamental de base de l'enquête PISA repose sur la notion de « littératie », soit à la capacité des élèves de faire des extrapolations à partir de ce qu'ils ont appris et d'appliquer leurs connaissances dans des situations nouvelles. L'enquête est également intéressante par l'attention qu'elle porte à l'apprentissage tout au long de la vie, et sa périodicité. Elle évalue, tous les 3 ans (depuis 2000 ; *cf.* encadré 9), les compétences des élèves de 15 ans dans des pays représentant, en tout, près de 90 % de l'économie

mondiale. Néanmoins, tout comme pour les enquêtes de l'IEA, ni la Chine ni l'Inde ne participent aux enquêtes de l'OCDE, ce qui représente une carence significative pour les enquêtes internationales. L'enquête PISA se base essentiellement sur la notion de « compétence » plus que sur celle de « savoir » (comme c'est le cas pour les enquêtes de l'IEA) : elle évalue dans quelle mesure les élèves en fin d'obligation scolaire ont acquis certaines des connaissances et compétences qui sont essentielles pour participer pleinement à la vie de la société des adultes.

L'enquête PISA se concentre sur trois domaines fondamentaux : les mathématiques, les sciences et la compréhension de l'écrit. Chaque cycle tend à privilégier un domaine de compétences et ainsi à fournir davantage d'informations dans l'évaluation de ce domaine. En 2000, c'est la lecture qui figurait comme priorité, supplantée par les mathématiques en 2003 puis, en 2006, par les sciences. Le cycle étant « bouclé », PISA 2009 se concentrait de nouveau sur la lecture. À la différence des enquêtes de l'IEA, l'enquête PISA évalue des élèves qui ont tous 15 ans, quel que soit le grade qu'ils fréquentent. (Pour les enquêtes de l'IEA, c'est davantage le grade de l'élève qui prime dans sa sélection – ou non – dans l'échantillon à évaluer).

Encadré 9 *Les différentes vagues de PISA*

Tout comme les enquêtes récentes de l'IEA, l'enquête PISA est avant tout un instrument de suivi : elle évalue tous les trois ans les connaissances et compétences des élèves en mathématiques, en sciences et en lecture. Le modèle fondamental de l'évaluation reste constant pour préserver la comparabilité d'un cycle à l'autre. À long terme, cette approche permet ainsi la comparabilité sur le moyen terme du niveau des élèves dans les trois domaines de compétences.

En 2000, 32 pays avaient participé à l'enquête PISA ; 11 pays supplémentaires sont venus compléter la liste deux années plus tard. Au total, 43 pays ont donc participé au premier cycle de PISA. Aucun pays africain n'est inclus dans cette enquête et l'on compte 13 pays en développement ou émergents (Albanie, Argentine, Brésil, Bulgarie, Chili, Fédération de Russie, Indonésie, Lettonie, Macédoine, Mexique, Pérou, Pologne et Thaïlande).

Le second cycle s'est déroulé en 2003 et a regroupé 41 pays. Parmi ceux-ci, un seul était africain (Tunisie) et 11 étaient en développement ou émergents (Brésil, Fédération de Russie, Indonésie, Lettonie, Mexique, Pologne, République Slovaque, Serbie, Thaïlande, Turquie et Uruguay).

Le troisième cycle a eu lieu en 2006 et a concerné 57 pays. Là encore, seule la Tunisie représente le continent africain ; le nombre de pays en développement ou émergents

•••

•••

ayant participé à PISA 2006 s'élève à 21 (Argentine, Azerbaïdjan, Brésil, Bulgarie, Chili, Colombie, Croatie, Estonie, Indonésie, Jordanie, République Kirghize, Lettonie, Lituanie, Mexique, Pologne, République Slovaque, Serbie, Thaïlande, Tunisie, Turquie et Uruguay).

Le quatrième cycle PISA s'est achevé le 16 décembre 2011 et s'est centré pour la deuxième fois sur les compétences en littératie. Réunissant 67 pays/régions, PISA 2009 a inclus un nombre plus important de pays en développement (tels que l'Azerbaïdjan ou le Kirghizistan). La lecture ayant été le thème principal de la première évaluation PISA, il est possible de retracer l'évolution de la performance des pays ayant participé aux vagues d'enquêtes (2000 et 2009 ; *cf.* encadré 10). Une dizaine de pays/régions supplémentaires ont décidé de rejoindre PISA 2009 avec une année de retard. Ainsi, leur évaluation a eu lieu durant l'année 2010. Parmi ces régions/pays, on trouve deux États indiens (Himachal Pradesh et Tamil Nadu) ainsi que le Costa Rica.

Basée sur le principe de « compétence », l'enquête PISA cherche à évaluer la capacité des jeunes à utiliser leurs connaissances et compétences pour relever les défis du monde réel. La priorité est accordée à ce que les élèves savent faire avec ce qu'ils ont appris à l'école et non seulement à la mesure dans laquelle ils sont simplement capables de les reproduire.

Cinq grands principes servent de base à l'enquête PISA (OCDE, 2007) :

- l'orientation de sa politique : l'enquête est résolument tournée vers la possibilité de définir des politiques éducatives à partir de comparaisons internationales ;
- son approche basée sur la notion de « littératie » qui renvoie davantage à la notion de « compétence » qu'à celle de « savoir » ;
- sa pertinence par rapport à l'apprentissage tout au long de la vie (en demandant notamment aux élèves leurs projets d'avenir, leurs perceptions ou encore leurs stratégies d'apprentissage) ;
- sa périodicité, qui permet aux pays de suivre l'évolution de la performance des élèves à l'école ;
- sa grande couverture géographique (57 pays avaient participé à l'enquête PISA 2006).

Près de 350 000 élèves, représentatifs des millions de jeunes de 15 ans scolarisés dans les pays participants, sont sélectionnés de manière aléatoire pour participer à chaque cycle PISA. On notera que si PISA identifie clairement l'établissement pour

le raccrocher au contexte socio-économique, l'enquête n'identifie pas la classe, et l'appariement du maître et de l'élève n'est donc pas possible. Seules les caractéristiques moyennes de l'équipe pédagogique de l'établissement sont décrites. Les élèves répondent à des épreuves sur papier d'une durée de deux heures. Les épreuves PISA sont constituées de questions demandant aux élèves d'élaborer leurs propres réponses ainsi que de QCM. Les questions sont regroupées par unités, qui s'articulent autour de textes ou de graphiques que les élèves sont susceptibles de rencontrer dans la vie courante. Il est par ailleurs distribué à chaque élève un questionnaire sur son milieu familial, ses habitudes d'apprentissage et ses attitudes à l'égard de chaque domaine de compétences, ainsi que son engagement et sa motivation. Les chefs d'établissement sont également tenus de remplir un questionnaire concernant leur établissement, notamment ses caractéristiques démographiques et la qualité de son environnement d'apprentissage. Par ailleurs, 16 pays ayant participé à l'enquête PISA 2006 ont remis un questionnaire aux parents des élèves sélectionnés pour participer aux épreuves PISA[15]. Ce questionnaire a permis de recueillir des informations sur l'investissement des parents dans l'éducation de leurs enfants, leur point de vue sur des questions et sur des professions scientifiques. 39 pays[16] ont également choisi de mettre en œuvre une option du questionnaire élève dans PISA 2006 : leurs élèves ont répondu à des questions portant sur les endroits où ils peuvent se servir d'un ordinateur, la fréquence à laquelle ils les utilisent et les usages qu'ils en font.

Les compétences requises sont au nombre de trois dans chaque domaine : pour les sciences, l'élève doit être capable d'identifier les questions d'ordre scientifique, expliquer des phénomènes de manière scientifique et utiliser des faits scientifiques ; dans le domaine de la lecture (ou encore « compréhension de l'écrit »), les compétences évaluées sont la capacité à localiser des informations, interpréter des textes et réfléchir sur des textes et les évaluer ; enfin, en ce qui concerne les mathématiques, les compétences attendues sont celles de reproduction (réalisation d'opérations mathématiques simples), de connexion (l'établissement de liens entre des idées pour résoudre des problèmes directs) et de réflexion (la pensée mathématique au sens large).

[15] Ces pays sont : l'Allemagne, la Bulgarie, la Colombie, la Corée du Sud, la Croatie, le Danemark, Hong-Kong-la Chine, l'Islande, l'Italie, le Luxembourg, Macao-Chine, la Nouvelle-Zélande, la Pologne, le Portugal, le Qatar et la Turquie.

[16] Ces pays sont : l'Australie, l'Autriche, la Belgique, la Bulgarie, le Canada, le Chili, la Colombie, la Croatie, la Corée, le Danemark, l'Espagne, la Finlande, la Fédération de Russie, la Grèce, la Hongrie, l'Irlande, l'Islande, l'Italie, le Japon, la Jordanie, la Lettonie, la Lituanie, Macao-Chine, le Monténégro, la Norvège, la Nouvelle-Zélande, les Pays-Bas, la Pologne, le Portugal, le Qatar, la République slovaque, la République tchèque, la Serbie, la Slovénie, la Suède, la Suisse, la Thaïlande, la Turquie et l'Uruguay. Une composante similaire a été administrée lors du cycle PISA 2003 (voir notamment OCDE, 2007).

Encadré 10 *Enquêtes PISA 2000-2009 : évolution des scores en lecture*

L'enquête PISA a été réalisée pour la première fois en 2000. Comme elle a lieu tous les trois ans, la dernière évaluation disponible (en 2010) est l'enquête de 2009. La lecture ayant été évaluée en 2000, elle l'a de nouveau été 9 ans plus tard, ce qui permet d'analyser l'évolution de la performance en lecture des pays sur la période.

Le graphique 5 montre la relation entre le score de 2009 et son évolution (en nombre de points) entre 2000 et 2009. Cette analyse est possible pour 38 pays : d'autres pays, comme la Tunisie ou la Turquie, ont commencé à participer à PISA en 2003 seulement et certains, tels que la Slovénie ou l'Estonie, n'ont pris part qu'aux évaluations de 2006 et de 2009.

Les pays qui ont connu la plus grande croissance de leur performance en lecture sont des pays en développement : Pérou (+43 points), Chili (+40 points), Albanie (+36 points) et Indonésie (+31 points). Parmi les pays développés qui ont connu une évolution significativement positive, figurent la Pologne (+21 points), le Portugal (+19 points) ou encore la Corée du Sud (+15 points).

Certains pays ont vu leur score diminuer entre 2000 et 2009. C'est notamment le cas de l'Irlande (-31 points), de l'Argentine (-20 points), de la Suède (-19 points) et de la République tchèque (-13 points).

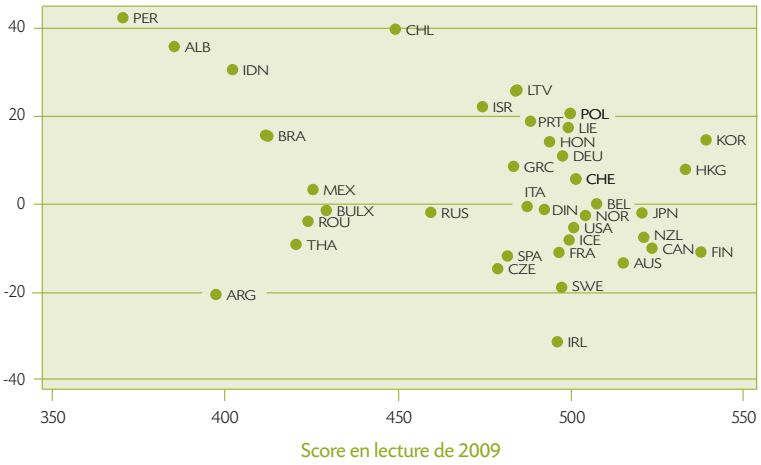
Au contraire, pour certains pays l'évolution est restée plutôt stable (ou n'a connu qu'une variation très faible) ; c'est, par exemple, le cas du Mexique (+3 points), de la Belgique (-1 point), de la Fédération de Russie (-2 points) ou encore du Japon (-2 points). Notons que la France a connu une baisse de sa performance assez importante, puisque le score en lecture passe de 505 points à 496 points.

Plus généralement, les pays les plus performants en 2009 sont la Corée du Sud et la Finlande. La région de Shanghai obtient le score le plus élevé pour PISA en lecture depuis sa création, ce qui témoigne de la forte concentration du « stock » de capital humain en Chine, dans ces villes économiques. Les pays les moins performants sont le Pérou (370 points), l'Azerbaïdjan (362 points) mais aussi et surtout le Kirghizstan (314).

•••

Graphique 5

Score en lecture de PISA 2009 et évolution du score depuis 2000



Source : OCDE (2010).

1.3.4. *Les enquêtes passées et prévues*

Un certain nombre d'évaluations ne sont actuellement plus conduites, bien qu'elles aient eu un impact important par le passé. Il s'agit des enquêtes du MLA, dans les pays en développement, et de l'*International Assessment of Education Progress* (IAEP) dans une optique internationale. Par ailleurs, des enquêtes sont en cours d'élaboration et pourraient être menées dans les années à venir : il s'agit ici de l'enquête *Assessment of Higher Education Learning Outcomes* (AHELO).

Basé sur l'expérience du NAEP, l'IAEP est une série de deux enquêtes qui a débuté en 1988. La méthodologie statistique et les procédures d'évaluation de l'IAEP sont basées sur celle du NAEP, qui est devenu, à partir de 1970, le principal instrument de mesure de la qualité des acquis scolaires aux États-Unis. L'enquête IAEP est donc fortement influencée par le curriculum américain. La première enquête de l'IAEP a été menée en 1988 pour six pays, deux domaines de compétence (mathématiques

et sciences) et auprès des élèves âgés de 13 ans. La deuxième enquête de l'IEAP s'est déroulée en 1991 pour deux groupes d'âges, 9 et 13 ans, et 19 pays. Ici encore, seules les mathématiques et les sciences étaient testées. Ces deux séries d'enquêtes ont fait l'objet de certaines critiques certaines critiques de la part des experts en évaluation, qui soulignaient qu'elles ne mesuraient que les savoirs et compétences des élèves en relation avec les curricula américains. En ne s'adaptant pas suffisamment aux spécificités des systèmes éducatifs des différents pays participants, cette enquête a donc été arrêtée et a renforcé la crédibilité des enquêtes de l'IEA.

Dans le cadre d'un projet conjoint de l'UNESCO et de l'UNICEF, le programme MLA mène des études sur les acquis de l'apprentissage à une vaste échelle géographique : elles sont effectuées dans plus de 40 pays avec la volonté de transférer la capacité d'analyse au niveau national (Chinapah, 2003). Sur la base de données recueillies une fois les apprentissages fondamentaux réalisés (soit, suivant les pays, entre la 3e et 6e année), les pays doivent être en mesure d'identifier les facteurs qui favorisent ou freinent les apprentissages à l'école primaire, d'analyser les problèmes, de formuler des adaptations des politiques éducatives et de suggérer de nouvelles pratiques pour améliorer la qualité de l'enseignement. Plus récemment, le projet MLA II a élargi les enquêtes au début du secondaire (grade 8). Contrairement aux enquêtes du PASEC et du SACMEQ, où les élèves sont testés uniquement sur des connaissances scolaires, le MLA porte également sur des questions de connaissances pratiques et de prévention. Il offre, en outre, la possibilité d'ajouter des items à la demande du pays étudié. Au total, 72 pays ont pris part à l'évaluation du niveau des élèves par le biais de l'enquête MLA. Toutefois, les données n'ont pas toutes été publiées. En complément des rapports nationaux, un rapport séparé sur MLA I a été rédigé pour 11 pays d'Afrique (Botswana, Madagascar, Malawi, Mali, Maroc, Maurice, Niger, Ouganda, Sénégal, Tunisie et Zambie ; voir UNESCO, 2000). Notons que l'enquête MLA a été l'objet de nombreuses critiques, concernant notamment la qualité de l'administration des questionnaires, les procédures d'échantillonnage effectuées dans certains pays et la difficulté d'obtenir des items comparables entre les différents pays. À l'inverse, il faut reconnaître que le programme MLA, avec un certain recul dans l'observation, a déclenché dans certains pays (en particulier en Asie du Sud) des actions d'évaluation des acquisitions.

L'évaluation internationale des performances des étudiants et des universités (AHELO) est un projet d'étude de l'OCDE qui vise à évaluer le niveau de compétences d'étudiants de l'enseignement supérieur. Ce dernier faisant l'objet d'un investissement stratégique et croissant de la part de pays et d'individus, il apparaît nécessaire d'analyser le niveau d'acquisition et de compétences des étudiants suivant ces formations. Environ 135 millions d'étudiants sont actuellement inscrits dans les 17 000

systèmes d'enseignement supérieur existants dans le monde. Cependant, devant l'absence d'évaluation d'instruments capables de comparer la qualité de l'enseignement et de l'acquisition dans une optique internationale, l'OCDE prévoit de lancer le projet AHELO. Celui-ci, actuellement en analyse de faisabilité, vise à évaluer la qualité et l'intérêt des apprentissages des étudiants dans le monde entier. L'objectif d'AHELO est de déterminer, d'ici fin 2012, dans quelle mesure une évaluation internationale des acquisitions dans l'enseignement supérieur serait scientifiquement et concrètement possible.

L'étude de faisabilité s'appuie sur trois thèmes (étudiés séparément mais avec une certaine cohérence) :

- le premier concerne le spectre des compétences génériques (*Generic Skills Strand).* L'instrument utilisé sera adapté du *Collegiate Learning Assessment* (CLA) élaboré par le *Council for Aid to Education* (CAE), un consortium international géré par l'*Australian Council for Educational Research* (ACER), l'OCDE et les équipes nationales dans les pays participants. Le CLA, largement administré aux États-Unis, est l'objet d'un projet visant son adaptation à différents contextes culturels et linguistiques. L'objectif du test visera à évaluer un ensemble de compétences intégrées incluant la réflexion critique, le raisonnement analytique, la résolution de problèmes et la communication écrite. Parmi les pays participants, on peut citer la Colombie, la Corée du Sud, l'Égypte, les États-Unis, la Finlande, le Koweït, le Mexique et la Norvège ;
- l'étude de faisabilité se concentre également sur les disciplines qui sont le plus proches, indépendamment du contexte culturel ; elle évalue ainsi actuellement l'économie et l'ingénierie. L'axe de travail sur les compétences en économie sera supervisé par l'*Educational Testing Service* (ETS), en collaboration avec l'ACER. Les pays participants à cet axe incluent la Belgique, l'Égypte, la Fédération de Russie, l'Italie, le Mexique et les Pays-Bas ;
- le dernier projet de test vise les compétences en ingénierie ; cet axe évaluera les compétences disciplinaires en génie civil. Les pays participants à cet axe incluent l'Australie, l'Égypte, le Japon et la Suède.

Au total, 15 pays ont accepté de participer à cette étude de faisabilité qui devrait s'achever en décembre 2012 : l'Australie, la Belgique flamande, la Colombie, la Corée du Sud, l'Égypte, les États-Unis, la Fédération de Russie, la Finlande, l'Italie, le Japon, Koweït, le Mexique, la Norvège, les Pays-Bas et la Suède.

1.4. Les enquêtes régionales sur les acquis des élèves

Trois grandes enquêtes régionales concernent les continents latino-américain et africain : SAQMEQ, PASEC et LLECE.

1.4.1. *L'enquête SACMEQ*

Le consortium connu sous le nom de SACMEQ trouve ses origines dans une enquête nationale de grande envergure menée en 1991 au Zimbabwe sur la qualité de son éducation primaire, avec le soutien de l'Institut international pour la planification de l'éducation (*International Institute for Educational Planning,* IIEP ; *cf.* Ross et Postlethwaite, 1991). Désireux de poursuivre cette réussite inspirée de l'IEA, un certain nombre de ministres de l'Éducation des pays d'Afrique de l'Est et du Sud ont exprimé leur intérêt pour cette étude et leur volonté de participer à une telle évaluation. Les planificateurs de sept pays se sont donc réunis à Paris en juillet 2004 et ont créé le SACMEQ. Les quinze ministères de l'éducation qui en sont membres sont ceux du Botswana, du Kenya, du Lesotho, du Malawi, de Maurice, du Mozambique, de Namibie, des Seychelles, d'Afrique du Sud, du Swaziland, de Tanzanie (et Tanzanie pour Zanzibar), d'Ouganda, de Zambie et du Zimbabwe.

Le premier volet de l'enquête SACMEQ (SACMEQ I) s'est déroulé entre 1995 et 1999 ; il regroupait alors sept pays (Kenya, Malawi, Maurice, Namibie, Tanzanie – Zanzibar, Zambie et Zimbabwe) et a évalué le niveau en lecture des élèves du grade 6. Bien qu'il s'agissait principalement d'études nationales, celles-ci avaient une dimension internationale, ayant en commun de nombreux éléments (questions de recherche, instruments, populations cibles, procédures d'échantillonnage et analyses). Un rapport séparé a été préparé pour chaque pays.

Le volet SACQMEQ II s'est déroulé de 2000 à 2002 et a regroupé 14 pays (Botswana, Kenya, Lesotho, Malawi, Maurice, Mozambique, Namibie, Seychelles, Afrique du Sud, Swaziland, Tanzanie, Ouganda, Zambie et Zanzibar). L'évaluation concernait cette fois-ci les mathématiques et la lecture. La population cible restait la même que celle du SACQMEQ I, à savoir les élèves du grade 6. Il est important de noter qu'un certain nombre d'items du SACMEQ II ont été repris de l'enquête TIMSS dans le but de produire des résultats comparables. Les questionnaires ont été utilisés pour obtenir des informations relatives aux *inputs* de l'éducation, aux conditions scolaires, et aux questions d'équité dans l'allocation des ressources humaines et matérielles. Les informations relatives au contexte socioéconomique ont été obtenues par le biais de questionnaires adressés aux élèves. Plus généralement, l'enquête SACMEQ II inclut

des items sélectionnés dans quatre enquêtes antérieures : l'enquête sur les indicateurs de la qualité de l'éducation du Zimbabwe *(Indicators of the Quality of Education Study)*, SACMEQ I, TIMSS et l'enquête RLS de l'IEA, conduite de 1990 à 1999.

Le troisième volet de l'enquête SACMEQ (SACMEQ III) s'est déroulé en septembre 2007, dans les mêmes pays que ceux du SACQMEQ II. La principale originalité de ce troisième volet a consisté à évaluer l'éducation concernant le VIH-Sida. En effet, une partie des questionnaires adressés aux élèves, aux enseignants ainsi qu'aux directeurs d'école intégrait des questions relatives à cette pandémie. Cette originalité est signe d'une réelle adaptation de l'enquête SACMEQ aux problèmes importants que connaît cette zone géographique. Notons que si le protocole SACMEQ est assez exigeant sur l'échantillonnage, pour des raisons de simplification de ce dernier, les écoles de petite taille sont exclues (écoles comportant en général 15 élèves et moins dans le grade concerné) afin de ne pas biaiser les résultats de certains pays.

Encadré 11 *Principaux résultats de l'évaluation SACMEQ (1995-2007)*

Le tableau 4 présente les principaux résultats des évaluations SACMEQ depuis 1995 par pays (classés par ordre alphabétique). Le suivi des scores en lecture est possible depuis 1995, tandis que, pour les mathématiques, seule la période 2000-2007 est disponible. De façon générale, on peut regrouper les performances en trois groupes :

- les pays qui ont connu une évolution positive dans les deux domaines : ils sont au nombre de six (Lesotho, Maurice, Namibie, Swaziland, Tanzanie et Zanzibar). La hausse la plus prononcée concerne la Namibie avec environ 45 points de plus dans les deux domaines entre 2000 et 2007 ; notons cependant que le niveau relatif de ce pays était plutôt bas au début des années 2000 ;
- les pays qui ont connu une évolution stable ou contrastée : quatre pays sont inclus dans cette catégorie (Afrique du Sud, Botswana, Kenya et Seychelles). Dans les deux premiers, la performance augmente peu tandis que dans les deux derniers, elle baisse légèrement. Ces évolutions ne sont cependant pas significatives ;
- les pays qui ont connu une baisse de leur performance dans les deux domaines : quatre pays sont concernés (Malawi, Mozambique, Ouganda et Zambie). La baisse la plus prononcée se situe au Mozambique avec près de 40 points en moins en lecture et plus de 46 points de moins en mathématiques.

•••

Tableau 4 *Principaux résultats des trois évaluations SACMEQ (1995, 2000 et 2007)*

Pays	Score			Différence		Score		Différence
	1995	2000	2007	1995-2007	1995-2007	2000	2007	2000-2007
Afrique du Sud		492,3	494,9		2,6	486,1	494,8	8,7
Botswana		521,1	534,6		13,5	512,9	520,5	7,6
Kenya	543,3	546,5	543,1	-0,2	-3,4	563,3	557	-6,3
Lesotho		451,2	467,9		16,7	447,2	476,9	29,7
Malawi	462,6	428,9	433,5	-29,1	4,6	432,9	447	14,1
Maurice	550,2	536,4	573,5	23,3	37,1	584,6	623,3	38,7
Mozambique		516,7	476		-40,7	530	483,8	-46,2
Namibie	472,9	448,8	496,9	24	48,1	430,9	471	40,1
Ouganda		482,4	478,7		-3,7	506,3	481,9	-24,4
Seychelles		582	575,1		-6,9	554,3	550,7	-3,6
Swaziland		529,6	549,4		19,8	516,5	540,8	24,3
Tanzanie		545,9	577,8		31,9	522,4	552,7	30,3
Zambie	477,5	440,1	434,4	-43,1	-5,7	435,2	435,2	0
Zanzibar	489,2	478,2	533,9	44,7	55,7	478,1	486,2	8,1
Zimbabwe	504,7		507,7	3			519,8	
SACMEQ		500	511,8		11,8	500	509,5	9,5

Sources : UNESCO IIEP (2010) et site Internet du SACMEQ : http://www.sacmeq.org/indicators.htm

1.4.2. *L'enquête PASEC*

Les enquêtes PASEC de la CONFEMEN concernent les pays francophones d'Afrique subsaharienne. En 1990, la 42e conférence ministérielle de la CONFEMEN (à Bamako) a constitué une réponse concrète de l'Afrique francophone au défi de l'EPT, lancé à Jomtien la même année. Les ministres ont alors décidé d'entreprendre en commun un Programme d'évaluation pour aider à la réflexion et à leur action : le PASEC a vu le jour l'année suivante, lors de la 43e conférence, à Djibouti en 1991. La CONFEMEN a fixé au PASEC quatre objectifs (CIEP, 2007) :

- identifier des modèles d'écoles efficaces et peu coûteux en comparant les performances des élèves, les méthodes d'enseignement et les moyens mis en œuvre ;
- développer une capacité interne et permanente d'évaluation du système éducatif dans chacun des pays participants ;
- diffuser librement les résultats obtenus, de même que les méthodes et les instruments d'évaluation préconisés ;
- renforcer le rôle d'observatoire permanent des systèmes éducatifs du Secrétariat technique permanent de la CONFEMEN.

Trois types d'évaluations sont menés par le PASEC :

- *l'évaluation diagnostique,* qui consiste à mesurer les acquisitions des élèves au cours d'une année scolaire, puis à identifier les facteurs qui influent positivement ou négativement sur les apprentissages (à ce jour, près d'une quinzaine d'évaluations de ce type ont été réalisées ou sont en cours de réalisation) ;
- *l'évaluation thématique,* qui se base sur les principes de l'évaluation diagnostique, mais tout en s'intéressant à un thème précis tel que la formation professionnelle des enseignants, ou encore le recrutement d'enseignants contractuels. Dans ce cas, l'échantillonnage n'est plus basé sur l'élève mais sur les catégories d'enseignants que l'on cherche à analyser (à cette date, quatre évaluations thématiques ont été réalisées) ;
- *le suivi de cohorte,* qui permet de suivre l'évolution d'un même groupe d'élèves pendant cinq années consécutives, en évaluant chaque année leurs acquisitions scolaires (à ce jour, deux suivis de cohorte ont pu être réalisés de façon complète).

L'enquête PASEC vise à évaluer, en début et en fin d'année, les élèves des grades 2 et 5. Par exemple, le test de mathématiques au grade 5 inclut des items qui évaluent les connaissances des élèves dans les propriétés des nombres et leur habilité à effectuer des calculs simples (addition et soustraction). Les tests incluent également des items qui demandent aux élèves d'utiliser l'addition, la soustraction, la multiplication et la division dans la résolution de problèmes. D'autres items évaluent les connaissances acquises dans les décimales, les fractions et les concepts de base en géométrie. Cette base de données comprend les performances scolaires au primaire en mathématiques et en français. Aux niveaux du CP2 (deuxième classe du primaire) et du CM1 (cinquième classe du primaire), entre 2 000 et 2 500 élèves dans une centaine d'écoles, ainsi que leurs professeurs et les directeurs ont été interrogés, dans chacun des onze pays. Le protocole de l'enquête présente la particularité de deux évaluations des acquis scolaires (en début et en fin d'année) ; il s'agit à ce jour de la seule enquête internationale menée en termes de « valeur ajoutée » (CONFEMEN, 2004).

Certains pays ont participé plusieurs fois à l'enquête PASEC. On recense, par ordre chronologique, les participations suivantes :

- PASEC I : Djibouti (1993/1994), Congo (1993/1994), Mali (1994/1995)[17] ;
- PASEC II : RCA (1994/1995), Sénégal (1995) ;
- PASEC III : Burkina Faso (1995/1996), Cameroun (1995/1996), Côte d'Ivoire (1995/1996) ;
- PASEC IV : Burkina Faso (1996/1998), Côte d'Ivoire (1996/1998), Sénégal (1996/2000), Madagascar (1997/1998), Tchad (2000/2001) ;
- PASEC V : Togo (2000/2001) ;
- PASEC VI : Guinée (2003/2004), Mali (2001/2002), Niger (2001/2002) ;
- PASEC VII : Bénin (2004/2005), Cameroun (2004/2005), Mauritanie (2003/2004), Madagascar (2005/2006), Maurice (2006), Tchad (2003/2004) ;
- PASEC VIII : Congo (2006/2007), Sénégal (2006/2007) et Burkina Faso (2006/2007) ;
- PASEC IX : Burundi (2009/2010), Côte d'Ivoire (2008/2009), Comores (2009-2010), Liban (2009-2010) ;
- PASEC X : Tchad, Togo, RDC (2009/2010) ;
- PASEC XI[18] : Vietnam (2011/2012), Cambodge (2011/2012), Laos (2011/2012), Mali (2011/2012).

Les enquêtes des phases PASEC V et VI étaient des enquêtes thématiques. La première enquête de Guinée du PASEC VI a été suivie d'une réplication en 2005/2006 afin d'évaluer un changement dans le cursus de formation des maîtres contractuels. Les enquêtes du Sénégal (1995/2000) et de Côte d'Ivoire (1995/1998), et plus partiellement du Burkina Faso (1995/1998) étaient des suivis de cohorte, tandis que les autres étaient des enquêtes diagnostiques. Les résultats de quatre premières évaluations sont difficilement accessibles (PASEC I et RCA en PASEC II) car les enquêtes n'ont pas été réalisées dans des conditions permettant la comparaison. En 2005, pour alimenter une étude sectorielle en RCA, le Pôle de Dakar (Bureau régional pour l'éducation en Afrique [BREDA]-UNESCO) a appliqué les outils PASEC en RCA, limitant l'analyse à un test unique de fin d'année, et en 5e année uniquement. Enfin, même si la diffusion des résultats est plus rapide que pour le SACMEQ, les résultats d'enquêtes récentes

[17] Dans cette première vague, réalisée par des équipes indépendantes dans chaque pays, des hétérogénéités de traitement ont rendu difficile la comparaison ; aussi les vagues suivantes ont été coordonnées par une équipe technique centrale.

[18] Le PASEC X 2e phase + Mali est devenu le PASEC XI.

ne sont pas encore publiés par la CONFEMEN, comme la très récente enquête sur la Côte d'Ivoire ou des enquêtes sur des pays non francophones (Guinée-Bissau, Vietnam, Liban). Les données constituent une base excellente pour une analyse des déterminants de la qualité des acquis scolaires au primaire.

Encadré 12 *Quelques résultats de l'évaluation PASEC : Sénégal et autres pays (2007)*

L'étude du PASEC au Sénégal a été effectuée au cours de l'année scolaire 2006-2007. Les élèves de deuxième année et cinquième année ont été soumis à des tests de français et de mathématiques, en début et en fin d'année. L'échantillon a été élaboré sur la base de données du ministère de l'Éducation nationale et fait référence au statut des écoles (privé, public et arabophone) et aux zones géographiques. Au total, 1 979 élèves de deuxième année et 1 910 de cinquième année ont participé à l'enquête. Le taux de déperdition était d'environ 14 %, ce qui reste assez élevé pour une école primaire.

Dans le tableau 5, nous présentons les résultats de l'évaluation PASEC pour le Sénégal mais également pour huit autres pays. La performance du Sénégal est faible en deuxième année, en comparaison de celle d'autres pays tels que le Cameroun ou Madagascar. Cependant, les scores augmentent de 19 % entre le début et la fin de l'année, ce qui marque une hausse conséquente, surtout lorsqu'on la compare aux évolutions des autres pays. Il est en effet assez alarmant d'observer une baisse de la performance des élèves dans des pays tels que la Mauritanie ou le Tchad. Cependant, notons que les questionnaires de fin d'année sont plus compliqués que ceux de début d'année, ce qui peut expliquer ces baisses quasiment généralisées. Tout comme l'ensemble des pays du tableau, la performance du Sénégal diminue entre le début et la fin de la cinquième année. Les pays les plus performants parmi les pays du PASEC sont le Cameroun et Madagascar, auxquels il conviendrait d'inclure Maurice, qui obtient des résultats plutôt élevés.

•••

Tableau 5 *Résultats pour 9 pays au test du PASEC (2004-2007)*

	Deuxième année			Cinquième année		
Pays	Début	Fin	Variation	Début	Fin	Variation
Burkina Faso	33,5	33,5	0	40,1	38,2	-5
Cameroun	53,2	54,7	3	54,6	47,2	-14
Congo	47,1	44,8	-5	44,9	36	-20
Madagascar	59,3	53,4	-10	64,0	52	-19
Sénégal	37,9	45,2	19	45,8	40,9	-11
Tchad	48,0	42,5	-11	39,5	34	-14
Bénin	39,9	33,1	-17	44,5	31,9	-28
Gambie	45,6	48,3	6	53,1	42,4	-20
Mauritanie	42,1	31,7	-25	24,9	22,2	-11

L'analyse de quelques résultats du PASEC Sénégal montre tout l'intérêt de la stratification utilisée. On constate ainsi de fortes différences de performance entre le secteur privé, le secteur public de la capitale et le public hors Dakar. Le score final moyen en deuxième année est d'environ 75 points, contre moins de 50 points pour Dakar et environ 41 points pour le public hors Dakar. Ainsi, le score moyen du privé est environ le double de celui du public hors Dakar. Une étude de l'évolution des scores entre 1996 et 2007 est également possible : on constate globalement que la performance est restée plutôt stable en deuxième année et qu'elle a augmenté d'environ 3 points de pourcentage en cinquième année.

Source : CONFEMEN (2007b).

1.4.3. *L'enquête LLECE*

Le réseau LLECE a été créé en 1994 et est coordonné par l'Office régional de l'UNESCO en Amérique latine et aux Caraïbes. L'objectif principal de cette enquête est de fournir des informations sur les niveaux des élèves et les facteurs associés à ces performances, susceptibles de guider les politiciens dans leur choix en termes de politique éducative pour les pays d'Amérique latine. Pour ce faire, le LLECE vise à répondre aux questions suivantes : Qu'apprennent les élèves ? À quel niveau l'apprentissage se réalise-t-il ? Quelles sont les compétences que les élèves développent ? Quand l'apprentissage se réalise-t-il ? Sous quelles conditions se réalise-t-il ? Quel écart constate-t-on entre les zones rurales et urbaines ? (Casassus *et al.,* 1998).

Dans chaque pays, des échantillons d'environ 4 000 élèves du grade 3 (âgés de 8 et 9 ans) et du grade 4 (âgés de 9 et 10 ans) ont été élaborés. Ces enquêtes ont concerné plus de 50 000 enfants, soit au moins 100 classes par pays.

L'enquête LLECE II (ou encore appelée SERCE pour la seconde vague) évalue le niveau des élèves en mathématiques en découpant ce domaine en cinq parties (UNESCO-OREALC, 2008, p.14) :

- domaine numérique,
- domaine de la géométrie,
- domaine de la mesure,
- domaine des compétences basées sur les informations,
- domaine de la variation.

On constate des similitudes avec l'enquête TIMSS. Par ailleurs, une partie des items des tests LLECE sont inspirés de ceux de TIMSS au grade 4 [19]. Les items sont basés sur la concordance entre les programmes scolaires et les acquis des élèves. Il est ainsi question de mesurer le niveau des acquis scolaires, plus que les compétences des élèves, qui leurs sont utiles dans la vie active (comme le fait PISA). Au questionnaire remis à l'élève, visant à cerner son environnement, a été couplé le test sur les acquis. Par ailleurs, un questionnaire remis à l'enseignant et un autre au directeur de l'école, ont été utiles pour obtenir des informations sur l'environnement de l'école. De manière générale, ces enquêtes privilégient l'opposition entre les zones urbaines et les zones rurales. L'enquête LLECE I évalue deux grades adjacents (les grades 3 et 4) mais à la même période, ce qui avait également été le cas de l'enquête TIMSS 1995. Il est pour autant impossible de suivre les élèves sur une période définie car ce sont des élèves différents qui sont évalués sur ces deux grades. Les élèves étaient âgés de 8 à 9 ans, selon les pays. Pour autant, la deuxième enquête LLECE (SERCE) a consisté à évaluer deux grades différents : 3 et 6 (UNESCO-OREALC, 2008 ; *cf.* encadré 13).

Les enquêtes LLECE se sont également intéressées aux acquisitions en lecture et en mathématiques aux grades 3 et 4 dans 13 pays du sous-continent (Casassus *et al.,* 1998) : l'Argentine, la Bolivie, le Brésil, le Chili, la Colombie, le Costa Rica, Cuba, la République dominicaine, le Honduras, le Mexique, le Paraguay, le Pérou et le Venezuela. Des données pour 11 pays seulement ont été incluses dans le rapport officiel (Casassus *et al.,* 1998). En 2006, le deuxième volet de l'enquête LLECE a été lancé dans les mêmes

[19] À la différence de TIMSS, l'enquête LLECE II (SERCE) évalue les élèves des grades 3 et 6. En effet, dans TIMSS, seuls les élèves du grade 4 au niveau primaire sont évalués (les élèves du grade 8 sont au niveau secondaire).

pays que ceux du LLECE I Cependant, à la différence du LLECE I, ce second volet a intégré le domaine des sciences en plus des mathématiques et de la lecture pour six pays volontaires. Les grades testés étaient les grades 3 et 6 (UNESCO-OREALC, 2008).

Encadré 13 *Principaux résultats de l'enquête SERCE (LLECE 2006)*

Au grade 3 et en mathématiques, les performances des pays dans l'enquête SERCE peuvent être classées en trois groupes de pays :

- les pays qui obtiennent une performance significativement supérieure à la moyenne régionale (Chili, Cuba, Costa Rica, Mexique, Uruguay et l'État du Nuevo León au Mexique) ;
- les pays qui affichent un niveau proche de la moyenne régionale (Argentine, Brésil et Colombie) ;
- les pays qui ont un score significativement inférieur à la moyenne régionale (Guatemala, Équateur, El Salvador, Nicaragua, Panama, Paraguay, Pérou et République dominicaine).

Tableau 6 *Scores moyens issus de l'enquête SERCE*

	Mathématiques		Lecture		Sciences
Pays	Grade 3	Grade 6	Grade 3	Grade 6	Grade 6
Argentine	505	513	510	504	489
Brésil	505	499	504	511	-
Chili	530	517	562	562	-
Colombie	499	493	511	510	503
Costa Rica	538	-	563	563	533
Cuba	648	638	627	627	662
Équateur	473	460	452	447	-
Guatemala	457	456	447	452	-
Mexique	532	542	530	523	-
Nicaragua	473	458	470	470	-
Nuevo León	563	554	558	558	511
Panama	463	452	467	469	473
Paraguay	486	468	469	467	469
Pérou	474	490	474	474	465
Rép. dominicaine	396	416	395	395	426
Salvador	483	472	496	496	479
Uruguay	539	578	523	530	-

•••

D'autres résultats intéressants peuvent être soulignés. Cette enquête a ainsi montré que les filles obtenaient des résultats sensiblement meilleurs que les garçons en lecture en grades 3 et 6 dans la moitié des 16 pays participants (Argentine, Brésil, Cuba, Mexique, Panama, Paraguay, République dominicaine et Uruguay). Par ailleurs, les scores obtenus en lecture par les élèves de grade 3 étaient très variables, indépendamment du niveau de performance des pays. À Cuba, pour la lecture, l'écart entre les 10 % d'élèves les meilleurs et les 10 % d'élèves les moins performants était de 295 points sur l'échelle de compétences (779-484). Dans la plupart des autres pays de la région, les écarts étaient plus faibles, comme en Argentine (236 points), au Costa Rica (231), au Salvador (219) et au Paraguay (241).

Source : UNESCO-OREALC (2008).

1.5. Les enquêtes hybrides

1.5.1. *L'enquête EGRA*

Les ministères de l'Éducation et les professionnels du développement de la Banque mondiale, de l'Agence américaine pour le développement international (USAID) et d'autres institutions ont demandé la création d'analyses simples, efficaces et peu coûteuses des performances d'apprentissage des élèves[20]. C'est pour répondre à cette demande qu'a été élaborée l'enquête EGRA. L'objectif fondamental de cette évaluation a été de tester les premiers pas des élèves dans l'apprentissage de la lecture : la reconnaissance des lettres de l'alphabet, la lecture de mots simples et la compréhension des phrases et des paragraphes.

EGRA est née en 2006, lorsque l'USAID a fait appel à RTI International pour élaborer un instrument d'évaluation des compétences fondamentales en lecture, dans le cadre du projet EdData II. L'objectif était d'analyser dans quelle mesure les élèves des premiers niveaux de primaire acquièrent les compétences en lecture et de susciter, par la suite, des politiques efficaces afin d'améliorer la performance de cette compétence essentielle à l'apprentissage. RTI International a développé un protocole pour une évaluation orale individuelle des compétences fondamentales des élèves en lecture. Un atelier a été organisé en novembre 2006 par l'USAID, la Banque mondiale et RTI International en vue de tester les différents outils qui pouvaient être utilisés. Les premiers tests ont eu lieu en 2007 au Sénégal (en français et en wolof) et en Gambie (en anglais).

[20] *Cf.* Abadzi (2006), Center for Global Development (2006) et Chabbott (2006).

Par ailleurs, l'USAID soutenait une application au Nicaragua (en espagnol). D'autres évaluations pilotées par les gouvernements nationaux, les missions d'USAID et des organisations non gouvernementales (ONG) ont eu lieu en Afrique du Sud, au Kenya, en Haïti, en Afghanistan et au Bangladesh.

L'évaluation EGRA a été développée pour plusieurs raisons : tout d'abord, son lien direct avec les recherches sur l'acquisition de la lecture et le développement cognitif. L'importance du caractère fondamental des compétences a été soulignée par la littérature dans ce domaine. Stanovich (1986) souligne en particulier l'existence d'un *« effet Matthieu »* dans l'acquisition de la lecture[21]. En d'autres termes : si des compétences fondamentales solides ne sont pas acquises dès le plus jeune âge, le fossé des performances d'apprentissage se creuse entre les élèves. Une autre raison renvoie à la simplicité de l'analyse pour les parties prenantes (ministères, enseignants, parents). L'évaluation réalisée au Pérou souligne ainsi l'importance des résultats d'une simple évaluation de la lecture et des mathématiques, faite auprès de 330 000 enfants dans 10 000 villages en utilisant du personnel entièrement bénévole. Cette initiative a été réalisée sous la coordination de l'ONG indienne Pratham, du ministère britannique du Développement international (*Department for International Development,* DFID) et de la Banque mondiale. Ce test a suscité un dialogue important sur la qualité de l'éducation et l'apprentissage des élèves.

L'instrument EGRA est conçu pour être une analyse « diagnostique du système » basée sur des exemples. Son objectif est de documenter la performance des élèves en matière de compétences fondamentales en lecture afin d'informer les ministères et les donateurs sur les besoins systémiques d'amélioration de l'instruction (USAID, 2009, p.9). EGRA n'est donc pas destiné à être directement utilisé par les enseignants et n'est pas non plus conçu pour être une analyse de haute responsabilisation, utilisée pour prendre des décisions en matière d'investissement ou pour déterminer le passage des élèves d'une année primaire à l'autre.

Plusieurs caractéristiques d'EGRA soulignent la faible portée de cette évaluation, en comparaison des tests PASEC, SACMEQ ou encore LLECE. Comme nous venons de le souligner, EGRA ne peut pas être utilisé comme un outil de haute responsabilisation : il s'agit uniquement d'un outil de diagnostic destiné au personnel ministériel. Par ailleurs, il n'est pas adapté aux comparaisons internationales : les différences entre

[21] L'expression « effet Matthieu » provient d'une situation qui apparaît dans une parabole biblique dans l'évangile de Saint Matthieu : *« Car à celui qui a, on donnera encore, et il sera dans l'abondance. Mais à celui qui n'a pas, on ôtera même ce qu'il a »* (25:29). Cet effet renvoie à l'idée que *« les riches s'enrichissent et les pauvres s'appauvrissent »*.

structures linguistiques et taux d'acquisition, notamment, réduisent les chances de comparaisons directes. Quelle que soit la langue, tous les enfants qui apprennent à lire évoluent du niveau de non-lecture (incapable de lire des mots), passent par celui de lecteur partiel (peut lire quelques éléments mais pas d'autres) et atteignent le niveau de lecteur compétent (peut lire tous les éléments, ou la majorité d'entre eux). Or, cette évolution change en fonction des langues : dans les langues qui possèdent une orthographe plus complexe, ce processus peut prendre plusieurs années (deux ans ou plus), comme c'est le cas de l'anglais. Au contraire, des langues régulières et transparentes telles que l'italien, le finnois et le grec ne nécessitent qu'une année d'instruction pour que les élèves atteignent un niveau comparable (Seymour *et al.,* 2003). Par ailleurs, au sein d'une même langue, la comparaison demeure difficile : les différences entre dialectes, ou encore et les vocabulaires différenciés selon les localités sont parfois importants. À ce jour, près de 41 pays sont en cours d'évaluation EGRA ou l'ont finalisée.

Encadré 14 *L'évaluation EGRA au Sénégal (2009)*

Le test EGRA s'est déroulé au Sénégal entre mai et juin 2009 dans les classes de CE1 (troisième année de scolarisation), dans 50 écoles de onze régions. L'échantillon ainsi obtenu est de 687 élèves dont 48,5 % de filles et 51,5 % de garçons. Quatre épreuves ont été soumises aux élèves : identification des lettres et de leurs sons, capacité à déchiffrer des mots inconnus (inventés) ; compréhension d'un texte lu à l'oral par l'élève ; questions de compréhension. Ces exercices ont été complétés par un entretien avec l'élève sur son environnement extrascolaire. L'étude incluait également une observation de l'enseignement tel qu'il est dispensé dans les classes des élèves évalués, ainsi que des entretiens avec les directeurs et les enseignants du CE1.

Selon le ministère, les élèves du CE1 devraient être capables de lire et écrire dès le début de l'année scolaire. On constate cependant que le nombre d'élèves ne sachant pas du tout lire est très important : au niveau du nombre de « mots inventés lus correctement en une minute » (MIM), plus de 26 % des élèves n'ont pas pu lire un seul mot ; pour le niveau « mots/texte lus correctement en une minute » (MTM), 123 élèves n'ont pas pu lire un seul mot, soit plus de 17 % de l'échantillon. Par ailleurs, 52,1 % des élèves ayant été interrogés sur au moins une question de compréhension n'ont pu répondre correctement à aucune question.

•••

•••

Tableau 7 *Résultats généraux d'EGRA au Sénégal (2009)*

	N (ayant complété l'épreuve)	Nombre de « 0 »	% de « 0 » par rapport à l'échantillon total
Nombre de GCM	683	4	< 1 %
Nombre de MIM	506	181	26,7 %
Nombre de MTM	566	121	17,8 %

GCM : graphèmes lus correctement en une minute
MIM : mots inventés lus correctement en une minute
MTM : mots/texte lus correctement en une minute

D'autres résultats peuvent être énoncés : les garçons ont enregistré des résultats légèrement meilleurs que ceux des filles, et de grandes différences ont été observées au sein d'une même classe (l'écart-type de la classe représente plus de la moitié de l'écart-type total). Enfin, l'enquête a révélé un lien fort entre certains facteurs environnementaux et la performance des élèves : possession d'un livre de français au programme (plus de 10 MTM de différence), fréquentation préalable de la maternelle (plus de 6 MTM de différence) ou encore présence, dans la famille, d'une personne sachant lire (près de 5 MTM de différence).

Source : Pouezevara et al.*(2010).*

1.5.2. *Les enquêtes EGMA et SSME*

Basée sur l'enquête EGRA et financée par l'USAID, l'évaluation EGMA vise à tester les élèves sur leurs compétences en mathématiques. Les grades évalués (1 à 3) sont les premières années du cycle scolaire obligatoire[22]. Après une étude sur les programmes scolaires, les experts de l'USAID ont confirmé l'idée selon laquelle les compétences et savoirs attendus en mathématiques convergeaient, dans la plupart des pays, au sein des premiers grades de la scolarisation obligatoire. Les objectifs tels que la connaissance et l'utilisation des nombres arabes et la comparaison et la classification d'un certain nombre d'objets sont des savoirs essentiels pour les mathématiques, présents dans la plupart des programmes scolaires du primaire. EGMA est ainsi un test oral qui

[22] La plupart des évaluations nationales et internationales concernent les grades 3 ou supérieurs. L'évaluation EGMA vise donc à combler le manque d'évaluation des grades antérieurs et à fournir un premier test diagnostique sur l'état des compétences en mathématiques.

permet de détecter les éventuelles déficiences et ainsi d'aider les acteurs politiques et les autorités éducatives à évaluer le niveau en mathématiques de leur système primaire.

L'étude pilote d'EGMA a été effectuée en juillet 2009 au Kenya (dans la région du Malindi) et a souligné la capacité de cette évaluation à montrer les forces et faiblesses des élèves en mathématiques. Le test requiert environ 15 minutes par élève et consiste en sept différentes tâches avec, chacune, plusieurs problèmes : identification des nombres, discrimination quantitative, nombres manquants, problèmes de mots, problèmes d'addition/soustraction, reconnaissance de formes et extension de schémas. Un enquêteur soumet le test à l'élève et le note. Un certain nombre d'items devraient être comparables entre les pays, étant donné la nature plus universelle des mathématiques (en comparaison avec la lecture). L'enquête EGMA étant encore à son stade expérimental, le nombre de pays qui comptent y participer reste encore indéterminé à ce jour. On sait seulement que les premiers participants devraient être l'Afrique du Sud, la Jamaïque et le Kenya.

Parallèlement aux évaluations EGRA et EGMA, l'USAID a lancé l'évaluation sur l'efficacité du management scolaire (SSME), dont l'objectif est de fournir une photographie multidimensionnelle et rigoureuse des pratiques de gestion scolaires au sein d'un pays ou d'une région. Les données obtenues sont construites afin de fournir aux administrateurs (d'écoles, de districts, de provinces ou d'États) des informations sur ce qui se passe dans leurs écoles et classes et de les aider ainsi à en améliorer l'efficacité. L'évaluation SSME collecte les données par le biais de l'observation directe des classes et des écoles, de questionnaires distribués aux élèves, d'interviews de parents, d'enseignants, ou encore de responsables d'écoles. L'avantage principal de ce type d'analyse est de fournir des informations à moindre coût. À ce jour, l'évaluation SSME a été appliquée en Jamaïque et au Pérou, en 2007. Ces études pilotes ont permis de montrer la faisabilité de l'étude et de différencier les écoles efficaces de celles qui avaient besoin de changements spécifiques, et de définir des groupes d'écoles selon leur efficacité. Depuis 2007, certaines parties du SSME ont été appliquées dans plus de dix pays présents d'Afrique, d'Amérique latine et d'Asie, conjointement aux évaluations EGRA et EGMA. Les pays ayant finalisé une évaluation SSME sont le Pérou, la Jamaïque, le Liberia, de Sénégal et l'Éthiopie. Les pays qui sont en cours d'évaluation sont le Kenya, l'Ouganda, le Mali, le Honduras, le Guatemala et le Népal[23].

[23] La liste des pays participants est disponible sur le site Internet de l'USAID : https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&ID=269

1.6. Les enquêtes sur les populations adultes

Pour rappel, l'objectif 6 de l'EPT est le suivant : *« Améliorer sous tous ses aspects la qualité de l'éducation et garantir son excellence de façon à obtenir pour tous des résultats d'apprentissage reconnus et quantifiables – notamment en ce qui concerne la lecture, l'écriture, le calcul et les compétences indispensables dans la vie courante. »* Au-delà de l'objectif d'éducation de qualité est mentionnée ici la nécessité d'apprentissages durables qui permettent l'autonomie et l'initiative dans la vie courante. Ces dernières années, particulièrement pour les pays d'Afrique subsaharienne, la question des compétences de bases acquises, et plus particulièrement de leur mesure, a été réexaminée. Un indicateur indirect de la qualité de l'éducation se base sur la mesure de la non-obsolescence des apprentissages, c'est-à-dire la probabilité pour un individu scolarisé dans l'éducation de base de conserver ces acquis à l'âge adulte.

- Traditionnellement, l'UNESCO publiait une statistique de population alphabétisée à l'âge adulte[24]. Celle-ci, basée principalement sur l'autodéclaration des individus aux recensements, a souvent été critiquée tant ces mesures paraissaient fragiles. Les enquêtes MICS, EDS-DHS, initiées par l'USAID, l'UNICEF et l'Organisation mondiale de la santé (OMS) et disponibles sur la presque totalité des pays en développement (souvent avec plusieurs vagues depuis 2000) mesurent, *via* des tests simples, le degré d'alphabétisme des populations enquêtées. Ceci permet de comparer ces niveaux provenant des enquêtes, les statistiques d'achèvement scolaire, et la source UNESCO des populations alphabétisées. Un travail du Pôle de Dakar (2007) permet d'apprécier l'ampleur des divergences entre ces trois mesures, celles-ci peuvent constituer une mesure alternative de la qualité de l'éducation reçue.
- Une large reconsidération du lien des savoirs acquis durant la formation initiale et leur opérationnalité dans la vie d'adulte a été impulsée par les enquêtes sur la littératie de l'OCDE. L'EIAA (IALS en anglais) a été la première évaluation sur les compétences des adultes. Elle est le fruit d'une collaboration entre l'OCDE, Statistiques Canada, le *National Center for Education Statistics* (NCES) et L'*Educational Testing Service* (ETS) aux États-Unis. L'EIAA a été administrée à des échantillons nationaux représentatifs, composés d'adultes âgés de 16 à 65 ans. L'enquête a été effectuée en trois cycles entre 1994 et 1998. Le cycle de 1994 incluait neuf pays : le Canada (populations anglophones et francophones), la France[25], l'Allemagne,

[24] Largement repris dans les anciens annuaires de l'UNESCO, le résultat de cet indicateur soulevait des interrogations, car des pays montraient des taux de progression significatifs sur la moyenne période alors que l'accès de la population jeune à l'éducation stagnait.

[25] La participation de la France a été annulée par la suite, juste avant la diffusion des résultats (Guerin-Pace et Blum, 1999, p.274).

l'Irlande, les Pays-Bas, la Pologne, la Suède, la Suisse (germanique et francophone), et les États-Unis. Cinq nouveaux pays ou territoires ont administré les instruments IALS en 1996 : l'Australie, la Communauté flamande de Belgique, la Grande Bretagne, la Nouvelle-Zélande et l'Irlande du Nord. Neuf autres pays ou régions ont participé à la troisième vague de 1998 : le Chili, la République tchèque, le Danemark, la Finlande, la Hongrie, l'Italie, la Norvège, la Slovénie et la région italophone de Suisse. Un rapport final a été publié en 2000, reprenant les principaux résultats de l'enquête EIAA (OCDE et Statistique Canada, 2000). L'EIAA s'est basée sur la notion de littératie. Celle-ci est définie comme une capacité et un mode de comportement particuliers : *« aptitude à comprendre et à utiliser l'information écrite dans la vie courante, à la maison, au travail et dans la collectivité en vue d'atteindre des buts personnels et d'étendre ses connaissances et ses capacités »* (OCDE et Statistique Canada, 2000). Plus spécifiquement, trois aspects de la littératie ont été testés :

– la compréhension de textes suivis, *i.e.* les connaissances et compétences nécessaires pour comprendre et utiliser l'information contenue dans des textes suivis, tels des éditoriaux, des nouvelles, des brochures et des modes d'emploi ;

– la compréhension de textes schématiques, *i.e.* les connaissances et compétences requises pour repérer et utiliser l'information présentée sous diverses formes, entre autres, les demandes d'emploi, les fiches de paie, les horaires de transports, etc. ;

– la compréhension de textes au contenu quantitatif, *i.e.* les connaissances et compétences nécessaires à l'application des opérations arithmétiques, séparément ou successivement, à des nombres contenus dans des imprimés, par exemple pour établir le solde d'un compte-chèques, calculer un pourboire, ou remplir un bon de commande.

L'enquête a employé par ailleurs une méthodologie mise au point par l'ETS pour mesurer la capacité de littératie sur une échelle allant de 0 à 500 points. Cinq niveaux ont été définis afin de permettre des analyses détaillées sur les performances. Depuis la fin de l'EIAA, la Chine, le Japon, la Malaisie, le Portugal, le Vanuatu et la province de l'Ontario ont aussi recueilli des données au moyen d'instruments inspirés de ceux de l'EIAA. À l'expérience, les analyses de l'EIAA sont restées plus difficiles à analyser comparativement dans la mesure où, à la différence du résultat de l'éducation initiale, il faut prendre en compte les contextes d'utilisation de ces compétences dans la vie d'adulte, avec les biais culturels qu'ils comportent (Guerin-Pace et Blum, 1999).

Encadré 15 *Principaux résultats de l'enquête EIAA (1994-1998)*

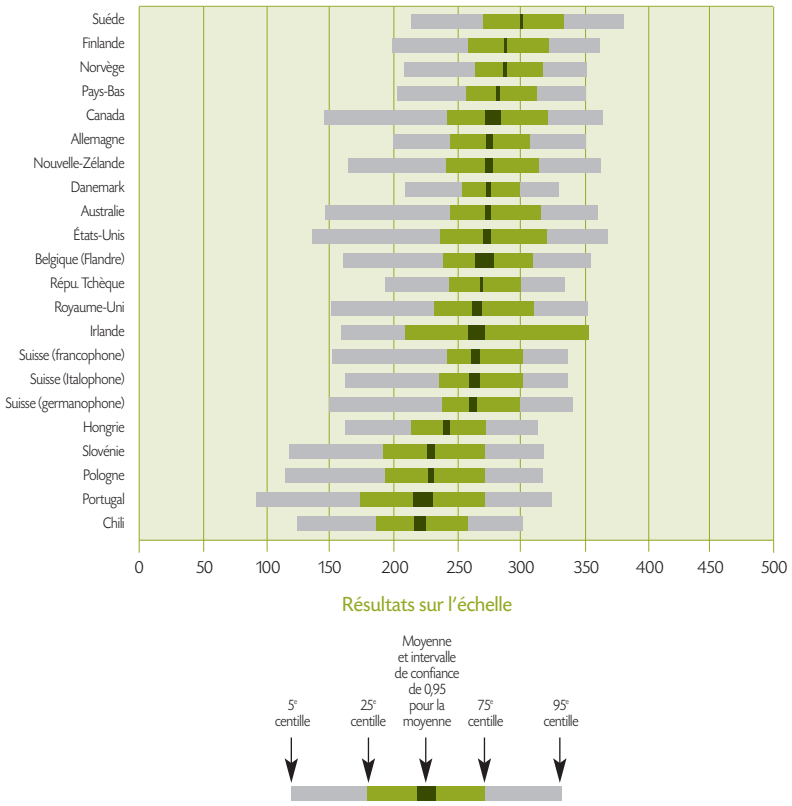
Le graphique 6 montre le résultat moyen et les résultats à divers centiles sur l'échelle des textes suivis, pour la population âgée de 16 à 65 ans, entre 1994 et 1998. Le résultat moyen révèle une variation considérable d'un pays à l'autre. La Suède enregistre la moyenne la plus élevée sur les trois échelles (compréhension de textes suivis, compréhension de textes schématiques, compréhension de textes au contenu quantitatif), tandis que le Chili obtient la note la plus faible.

La répartition des niveaux de littératie à l'intérieur d'un pays varie sensiblement, elle aussi, sur chacune des trois échelles allant de 0 à 500 points : au Danemark, par exemple, la fourchette des résultats du 5e au 95e centiles de l'échelle de compréhension de textes suivis est d'environ 120 points. Certains pays obtiennent des résultats moyens semblables sur les trois échelles : la Suède et la Norvège figurent parmi les quatre pays les plus performants sur les trois échelles, tandis que le Chili, la Pologne, le Portugal et la Slovénie y occupent un rang modeste. D'autres pays se classent différemment d'une échelle à l'autre : la République tchèque, par exemple, se situe au milieu du classement sur l'échelle de compréhension de textes suivis, mais au sommet de celle des textes au contenu quantitatif ; inversement, le Canada se classe parmi les meilleurs sur l'échelle de compréhension de textes suivis, mais au milieu de celle des textes au contenu quantitatif.

•••

Graphique 6

Résultats moyens avec un intervalle de confiance de 0,95 et résultats aux 5e, 25e, 75e et 95e centiles sur l'échelle des textes suivis ; population âgée de 16 à 65 ans ; 1994-1998



Source : OCDE et Statistique Canada (2009).

- Prolongeant l'enquête EIAA, ELCA s'est déroulée en 2003. Il s'agit d'une « [...] *étude conjointe à grande échelle menée par des gouvernements, des organismes statistiques nationaux, des établissements de recherche et des organismes multilatéraux* » qui fournit des mesures comparables à l'échelle internationale dans quatre domaines : la compréhension de textes suivis et de textes schématiques (qui représentent les deux domaines de la littératie), la numératie et les résolutions de problèmes (OCDE et Statistique Canada, 2005). Plus spécifiquement, l'élaboration et la gestion de l'ELCA ont été coordonnées par Statistique Canada et l'ETS en collaboration avec le NCES du ministère de l'Éducation des États-Unis, l'OCDE et l'Institut de statistique (IS) de l'UNESCO. Les pays participants sont les Bermudes, le Canada, l'Italie, la Suisse, les États-Unis et l'État mexicain de Nuevo León. Tous devaient recueillir des données d'un échantillon national représentatif d'au moins 3 000 répondants âgés de 16 à 65 ans pour chaque langue sondée. Comme cela avait été le cas avec l'EIAA, l'ELCA de 2003 conceptualise les compétences selon un continuum qui indique dans quelle mesure les adultes utilisent l'information afin de « fonctionner » dans la société et dans l'économie. Les compétences dans chaque domaine sont mesurées selon une échelle continue. Chaque échelle de compétence va de zéro jusqu'à un maximum théorique de 500 points. Une personne dont les connaissances la situent à un échelon donné de l'échelle a 80 % de probabilité de réussir une tâche qui comporte ce même niveau de difficulté.

Encadré 16 *Principaux résultats de l'enquête ELCA*

Les résultats de l'enquête ELCA sont présentés dans le tableau 8 (pour des raisons de place, nous ne présentons que les résultats en compréhension des textes suivis et en numératie). L'échelle des scores est comprise entre 0 et 500 points. Les pays les plus performants sont la Norvège, les Bermudes et le Canada. Les pays les moins performants sont l'Italie et les États-Unis. La performance dans les deux domaines est fortement corrélée : l'Italie obtient plus de 60 points de moins que la Norvège. Selon les pays, entre un tiers et deux tiers des adultes n'atteignent pas le niveau 3 des compétences, niveau considéré par les experts comme un socle minimum de la société du savoir et de l'information.

L'écart de performance des adultes entre les centiles les moins performants et les centiles les plus performants est significativement plus faible dans certains pays (Norvège et Suisse) et plus élevé dans d'autres (Italie et États-Unis). Certains pays ont de meilleurs scores dans certains domaines. Ainsi, la Suisse a une performance relativement meilleure dans la numératie tandis que les Bermudes affichent une performance meilleure dans la compréhension de textes suivis. •••

Tableau 8 *Principaux résultats de l'enquête ELCA*

Pays	5e centile	25e centile	Moyenne	75e centile	95e centile
Compréhension des textes suivis					
Bermudes	192,0	255,6	289,8	328,4	374,1
Canada	178,1	250,6	280,8	318,0	358,7
États-Unis	175,9	235,5	268,6	306,1	346,9
Italie	135,8	192,3	229,1	267,2	318,7
Norvège	211,5	263,5	290,1	320,5	355,8
Nuevo León (Mex.)	143,3	206,1	228,3	255,8	292,0
Suisse	193,8	242,1	272,1	303,7	346,0
Numératie					
Bermudes	176,8	233,3	269,7	308,5	359,4
Canada	170,4	237,2	272,3	311,9	357,7
États-Unis	162,8	222,4	260,9	302,2	351,5
Italie	148,8	200,4	233,3	267,1	313,9
Norvège	204,9	255,2	284,9	316,2	357,8
Suisse	212,4	257,8	289,8	322,2	368,9

Source : OCDE et Statistique Canada (2005).

- Le PIAAC est un projet en cours de préparation. Faisant suite aux évaluations ELCA et EIAA, son objectif est de mesurer le niveau et la répartition des compétences des adultes dans une optique internationale. Il porte sur les facultés cognitives et les compétences professionnelles essentiellement requises pour participer avec succès à l'économie et à la société du 21e siècle. La principale compétence mesurée sera la littératie à l'ère de l'information. Quatre domaines de compétences seront ainsi évalués : la résolution de problèmes dans un environnement à forte composante technologique, la littératie, l'aptitude au calcul, et la maîtrise des savoirs fondamentaux. L'innovation majeure du programme PIAAC est le recueil d'informations relatives à l'utilisation que font les enquêtés des compétences professionnelles de base dans l'exercice de leur emploi. Il sera ainsi possible d'utiliser les données obtenues pour étudier les différences entre les pays dans l'utilisation de ces compétences et de repérer l'existence et la nature des déficits de qualifications.

L'enquête a été administrée en 2011 et 2012 et les résultats seront publiés en octobre 2013. Un test « terrain » a eu lieu en 2010, suite aux travaux préparatoires qui se sont déroulés en 2008 et 2009. Pour chaque pays participant, environ 5 000 adultes âgés de 16 à 65 ans ont été interrogés. Ainsi, pour certains pays, il a été possible de comparer l'évolution des compétences des adultes en littératie sur une période de 13 à 17 ans selon les pays. Au total, 23 pays ont participé à la première vague de PIAAC, la quasi-totalité étant des pays développés. On compte seulement la Fédération de Russie comme pays à revenus intermédiaires parmi les pays membres.

- L'évaluation PIIAC étant organisée par l'OCDE, seuls les pays développés ou quelques pays à revenus moyens seulement peuvent y participer. Afin de permettre à des pays en développement d'évaluer les compétences en littératie, l'IS de l'UNESCO a élaboré le Programme LAMP afin de fournir aux décideurs politiques, aux bailleurs et aux autres parties les renseignements requis pour planifier et mettre en œuvre efficacement les programmes d'alphabétisation. Les compétences mesurées par LAMP sont les aptitudes requises pour composer avec les situations de la vie quotidienne ; parmi les tests proposés, on trouve : *(i)* des textes de nature continue structurés en phrases et paragraphes ; *(ii)* des textes de nature discontinue organisés en tableaux, diagrammes ou graphiques ; *(iii)* l'usage des nombres et de concepts mathématiques de base tels que la forme, la dimension et le volume. Les avantages de LAMP sont de *(i)* répondre aux besoins des pays à tous les niveaux de développement, *(ii)* d'avoir été validé depuis 2009 en neuf langues (appartenant à cinq familles linguistiques) et de *(iii)* pouvoir produire des données qui peuvent être comparées sans tenir compte de la période, au pays ni à la culture. À ce jour, le programme LAMP a été validé dans huit pays pilotes (El Salvador, la Jordanie, le Maroc, la Mongolie, le Niger, le Paraguay, les Territoires autonomes palestiniens, le Vietnam)[26]. Les tests ont eu lieu entre octobre 2010 et avril 2011 dans la plupart des pays. D'autres pays montrent également leur intérêt à cette évaluation (Anguilla, l'Inde, la Jamaïque, le Laos et la Namibie). Etant donné que certains domaines du test sont similaires à l'enquête ELCA, il est possible de comparer les résultats de LAMP avec les évaluations effectuées par Statistique Canada et l'OCDE (ELCA, EIAA, PIAAC).

Le tableau 9 présente de façon synthétique l'ensemble des évaluations existantes à ce jour avec les organismes qui les supervisent, les abréviations couramment utilisées, les domaines évalués, le nombre de pays participants ainsi que les grades testés. Parfois, le nombre de pays varie au sein de la même évaluation, car certaines provinces

[26] Voir UIS-UNESCO (2010).

ont pris part à l'enquête ou parce que certains pays ont participé plusieurs fois à différentes étapes. Par ailleurs, les informations sur les grades des élèves doivent être prises avec précaution puisque, pour certaines enquêtes, le mode de sélection des élèves est basé sur leur âge et non sur leur grade uniquement.

Tableau 9 *Évaluations sur les acquis des élèves ou adultes depuis 1959*

Année	Organisme	Abréviation	Matière	Nombre de pays	Grade
1959-1960	IEA	Étude Pilote	Mathématiques Lecture Sciences	12 12 12	7,8 7,8 7,8
1964	IEA	FIMS	Mathématiques	12	7, FS
1970-1971	IEA	SRC	Lecture	15	4,8, FS.
1970-1972	IEA	FISS	Sciences	19	4,8, FS.
1980-82	IEA	SIMS	Mathématiques	19	8, FS
1983-1984	IEA	SISS	Sciences	23	4,8, FS
1988, 1990-91	NCES	IAEP	Mathématiques Sciences	6, 19 6, 19	7-8 / 4,7-8 7-8 / 4,7-8
1990-1991	IEA	RLS	Lecture	32	3-4,7-8
1995, 1999, 2003, 2007, 2011	IEA	TIMSS	Mathématiques Sciences	45 / 38 / 26,48 / 66 / 65 45 / 38 / 26,48 / 66 / 65	3-4,7-8, FS* / 8 / 4,8 / 4,8 / 4,8 3-4,7-8, FS / 8 / 4,8 / 4,8 / 4,8
1992-1997	UNESCO	MLA	Mathématiques Sciences Lecture	72 72 72	6,8 6,8 6,8
1997, 2006	UNESCO/ OREALC	LLECE/ SERCE	Mathématiques Lecture Sciences	11, 13 11, 13 0, 6	3,4 / 3,6 3,4 / 3,6 ... / 6
1999, 2002, 2007	UNESCO/ IIEP	SACMEQ	Lecture Mathématiques	7, 15 0, 15	6 6
1993-2011	CONFEMEN	PASEC	Mathématiques Lecture	22 22	2,5 2,5
2001, 2006, 2011	IEA	PIRLS	Lecture	35, 41, 55	4
2011 2000, 2003,	IEA OCDE	prePIRLS PISA	Lecture Mathématiques	6 43, 41, 57	2,3 8, 9, 10

Année	Organisme	Abréviation	Matière	Nombre de pays	Grade
2006, 2009			Sciences Lecture	43, 41, 57 43, 41, 57	8, 9, 10 8, 9, 10
2013	OCDE	AHELO	Économie Sciences Général	6 4 8	Université
2006 **	USAID	EGRA	Lecture	41	1-3
2009 **	USAID	EGMA	Mathématiques	-	1-3
2007 **	USAID	SSME	Management	11	Primaire
1994-1998	OCDE / Stat Canada / NCES / ETS	EIAA IALS	Littéracie	23	Population adulte
2003	OCDE / Stat Canada / NCES / ETS	ELCA ALL	Littéracie	6	Population adulte
2011 **	OCDE	PIAAC	Littéracie	26	Population adulte
2009 **	UIS UNESCO	LAMP	Littéracie	8	Population adulte

** FS :* Final Secondary *(dernière année du cycle secondaire).*
*** : évaluations toujours en cours.*
Source : auteurs, à partir des rapports cités.

Tableau 10 *Évaluations sur les acquis des élèves ou adultes depuis 1959 dans les États arabes*

Année	Nom	Organisation(s)	Niveau(x) évalué(s)	Domaine(s) évalué(s)	Année(s)
Algérie	Programme national d'évaluation du rendement	MdE	Grades 3, 6, 9, 10	Lecture (arabe, français), mathématiques	nd
Bahreïn	NEU	MdE	Grades 3, 6	Lecture (arabe, anglais), mathématiques, sciences	2009, 2010
Djibouti	Évaluation du niveau de qualité et du rendement cognitif	Centre de Recherche, d'Information et de Production de l'Éducation Nationale	Primaire et secondaire	Lecture, mathématiques	1991, 1992, 1997-2000
Égypte	Global Evaluation	MdE	Grades 1, 2, 3	Lecture, mathématiques, sciences	Annuel depuis 2005
Jordanie	National Assessment	National Centre for Human Resource Development	Grades 4, 6	Lecture, mathématiques, sciences	1993, 1995
	National Test	MdE, DFID	Grade 10	Lecture, mathématiques, sciences	Annuel depuis 2000
Liban	Mesure des acquis d'apprentissage	Centre de recherche et de développement pédagogiques	Grade 4	Lecture, mathématiques, sciences	1994, 1995, 1996
	Assessment	MdE	Grade 3, 4	Lecture,` mathématiques	1999
Mauritanie	Évaluation	MdE	Grades 3, 4, 5, 6	Lecture, mathématiques	1998
	Évaluation	MdE	Grade 5	Lecture, mathématiques	2003

Année	Nom	Organisation(s)	Niveau(x) évalué(s)	Domaine(s) évalué(s)	Année(s)
Maroc	Diagnostic et appui aux apprentissages	MdE	Grades 3, 5, 8	Lecture, mathématiques	2000
	Évaluation des prérequis	MdE, UNICEF	Grades 4, 6	Lecture, mathématiques, sciences	2001
	Évaluation des acquis des élèves	MdE, EU	Grade 6	Lecture, mathématiques, sciences	2006
Oman	MLA	NICEF-UNESCO	Grades 4, 6, 8, 9, 10	Mathématiques, lecture, Social Studies	1993-2001
	Assessment	MdE	Grade 4	Lecture, mathématiques, sciences	2004
	Assessment	MdE	Grades 4, 7, 10	Lecture (arabe, anglais), mathématiques, sciences	2007
Qatar	QCEA	MdE	Grades 1-12	Lecture (arabe, anglais), mathématiques, sciences	2004, 2006
Arabie Saoudite	Diagnostic Test in the Public Evaluation System	MdE	Grades 1, 2, 3	Lecture, mathématiques	n.d.
Émirats Arabes Unis	National Assessment of Student Achievement and Progress	ACER	Grades 5, 7	Lecture (Literacy), mathématiques	2005
Yémen	MLA	UNICEF-UNESCO	Grades 4, 6	Lecture, mathématiques, sciences	2002, 2005

Sources : données du BIE et rapports nationaux.

Notes : EU : *Union européenne*
MdE : *Ministère de l'Éducation*
NEU : *National Examinations Unit*
QCEA : *Qatari Comprehensive Educational Assessment*

Tableau 11 *Évaluations sur les acquis des élèves ou adultes depuis 1959 dans les pays d'Asie centrale et du Sud*

Année	Nom	Organisation(s)	Niveau(x) évalué(s)	Domaine(s) évalué(s)	Année(s)
Bhutan	National Education Assessment				
Inde	NCERT tests annuels normalisés	NCERT	Grades 3 et 5	Lecture, mathématiques, sciences (5)	Depuis 1993 (sporadique)
Kazakhstan	MLA	UNICEF-UNESCO	Grade 4	Lecture, mathématiques, sciences	1999
	Unified National Testing		Grade 4	Lecture, mathématiques	2004, 2007, 2009
Kirghizistan	MLA	UNICEF-UNESCO	Grade 5	Lecture, mathématiques, sciences	2000
	MLA	UNICEF-UNESCO	Grade 8	Lecture, mathématiques, sciences	2002
	NAEAS	MdE, Banque mondiale	Grade 4	Lecture, mathématiques	2006/2007
Maldives	PAST/EQUIS	MdE	Grades 4 et 7	Lecture, mathématiques	2008
Mongolie	National test	MdE	Grades 5, 9, 11	Lecture, mathématiques	Annuel depuis 1997
	Regional Test at aigmag (district) level	State Professional Assessment Agency	Grades 5, 9, 11 (variable)	Lecture, mathématiques, sciences	Tous les 5/6 ans depuis 1997
	NASA	MdE	Grade 8	Mathématiques, éducation civique	2005
	NAPEMR	MdE	Grade 5	Lecture, mathématiques	2008
	NAEP	MdE	Grades 3 et 5	Lecture, mathématiques	1992

Année	Nom	Organisation(s)	Niveau(x) évalué(s)	Domaine(s) évalué(s)	Année(s)
Népal	National Primary Assesments	MdE sur base MLA	Grade 4	Lecture, mathématiques, sciences, civisme	1999, non continu
Pakistan	National Assessments Evaluation System	MdE	Grade 4	Lecture, mathématiques, sciences Une compétence par an sur un cycle de 3 ans	2003
Sri Lanka	An Evaluation of Implementation of the New Curriculum	NIE	Grade 6 et 10	Mathématiques, anglais, cingalais, sciences et histoire	2007

Sources : données du BIE et rapports nationaux.

Notes : EQUIS : *European Quality Improvement System*
NAEAS : *National Assessment of Educational Achievements of Students*
NAPEMR : *Assessment of Primary Education Mathematics and Reading*
NASA : *National Assessment of Students' Achievement*
NCERT : *National Council of Educational Research and Training*
NIE : *National Institute for Educational Policy Research*
PAST : *Program for Achievement School Test*

Tableau 12 *Évaluations sur les acquis des élèves ou adultes depuis 1959 dans les pays d'Asie de l'Est et du Pacifique*

Année	Nom	Organisation(s)	Niveau(x) évalué(s)	Domaine(s) évalué(s)	Année(s)
Australie	National Assessment	MdE	Grades 3, 5	Lecture, mathématiques	1998, 2000
	NAPLAN	MdE	Grades 3, 5, 7, 9	Lecture, mathématiques	Annuel depuis 2008
Iles Fidji	FILNA	MdE	Grades 4-6	Lecture, mathématiques	Annuel depuis 2004
Iles Cook	Assessment	MdE	Grade 4	Lecture, mathématiques	1999
	Assessment	MdE	Grade 6	Lecture, mathématiques	2003
Indonésie	Assessment of Students Learning Achievement	Educational National Standard Board	Grade 3	Lecture, mathématiques	Annuel depuis 2005
Japon	NAAA	MdE, NIER	Grades 3, 6, 9	Lecture, mathématiques, sciences, langues étrangères	2007
	National Assessment of Learning Outcomes	NIER	Grades 5, 9, 12 (variable)	Lecture (japonais, anglais), mathématiques, sciences	2002, 2003, 2004
Corée du Sud	National Assessment	MdE	Grades 4-6	Lecture, mathématiques	1994-1997
	BASDA	MdE	Grade 3	Lecture, mathématiques, sciences	
	NAEA	Korean Institute of Curriculum and Evaluation	Grades 6, 9, 10	Lecture, mathématiques, sciences	Annuel depuis 2003

Année	Nom	Organisation(s)	Niveau(x) évalué(s)	Domaine(s) évalué(s)	Année(s)
Laos	Assessment	MdE, ESQAC	Grade 3	Lecture, mathématiques, sciences	2009
	Assessment	RIES	Grade 5	Lecture, mathématiques	2007
	National Literacy Survey	MdE, UNESCO, UNICEF	Grade 1 (âge 6)	Lecture, mathématiques	2000
Malaisie	Primary Assessment Test	MdE	Grade 6	Lecture, mathématiques, sciences	1994-1998
Myanmar	Learning Achievement Study	MdE, UNICEF	Grades 3, 5	Lecture, mathématiques, sciences	2005, 2006
N. Zélande	NEMP	MdE	Grades 4, 8	Lecture, mathématiques	1991
Papouasie N. Guinée	CBE	MdE	Grade 8	Lecture, mathématiques	2006
Philippines	NEAT	MdE		Lecture, mathématiques	1995, 1998
	NAT	MdE	Grade 6	Lecture, mathématiques	2009/2010
Samoa	SPELL I	MdE	Grade 4	Lecture, mathématiques	2006
	SPELL II	MdE	Grade 6	Lecture, mathématiques	2009
Singapour	Core Research Program	Centre for Research in Pedagogy and Practice	De la maternelle au secondaire	Lecture, mathématiques	2001
Vietnam	Lecture and Mathématiques Assessment Study	MdE, Banque mondiale	Grade 5	Lecture, mathématiques	2001

Source : auteurs, à partir des rapports cités.

Notes : BASDA : *Basic Academic Skills Diagnostic Assessment*
CBE : *Certification of Basic Education*
ESQAC : *Educational Standards and Quality Assurance Center*
FILNA : *Fiji Island literacy numeray assessment*
NAAA : *National Assessment of Academic Abiblity*
NAEA : *National Assessment of Educational Achievement*
NAT : *Nationa Achievement Test)*
NEAT : *National Elementary Achievement Test*
NEMP : *National Education Monitoring Project*
RIES : *Research Institute for the Educational Sciences*
SPELL I : *Samoa Primary Education Literacy Levels*

Tableau 13 *Évaluations sur les acquis des élèves ou adultes depuis 1959 dans les pays d'Amérique latine et des Caraïbes*

(Partie 1)

Année	Nom	Organisation(s)	Niveau(x) évalué(s)	Domaine(s) évalué(s)	Année(s)
Argentine	ONE	MdE	Grades 3, 6, 7, 9, 12	Lecture, Mathématiques, Sciences	1993-2005 (sauf 2001), 2007
	Sistema Nacional de Evaluacion de la Calidad Educativa	Instituto de Calidad Educativa	Grades 3, 7, 9, 12	Lecture, Mathématiques, Sciences	Annuel 1993-2001
Barbados	NCRA	Ministry of Education	Grade 2	Lecture, Mathématiques	2002
Bolivie	SIMECAL	MdE	Grades 3, 6, 8, 12 (variable)	Lecture, Mathématiques	Annuel 1993-2005
Brésil	SAEB	MdE	Grades 4, 8, 11	Lecture, Mathématiques	Tous les deux ans depuis 1990
	National System of Evaluation of Basic Education	MdE, INEP	Grades 1, 3, 4, 5, 7, 8, 11 (variable)	Lecture, Mathématiques, Sciences	1990-2005 (variable)
	Examen Nacional de Ensenanza Media	INEP	Grade 9	nd	nd
Chili	SIMCE	MdE	Grade 1, 4, 8, 10	Lecture, Mathématiques, Sciences	Annuel depuis 1988
	Prueba de Evaluacion del Rendimiento Escolar	MdE, Universidad Catolica	Grades 4, 8	Lecture, Mathématiques, Sciences	1982, 1983, 1984
Colombie	Medicion y Evaluacion de Aprendizajes	MdE, ICFES	Grades 3, 5, 7 ,9	Lecture, Mathématiques	Annuel 1991-1994
	Examenes de Estado	MdE, ICFES	Grade 11	Lecture, Mathématiques, Sciences	Annuel 1980-2005

Année	Nom	Organisation(s)	Niveau(x) évalué(s)	Domaine(s) évalué(s)	Année(s)
	SABER	MdE	Grades 3, 5, 7, 9	Lecture, Mathématiques, Sciences	1991, 1992, 1997, 1998, 2002, 2003, 2006
Costa Rica	Pruebas de Conocimientos	MdE, Universidad de Costa Rica	Grades 3, 6, 9, 11, 12 (variable)	Lecture, Mathématiques, Sciences	Annuel 1986-1997
	Pruebas Nacionales de Bachillerato	MdE	Secondary School	Lecture, Mathématiques, Sciences	Annuel 1988-2003
Cuba	Pruebas de Aprendizaje	MdE, SECE	Grades 3, 4, 6, 9, 12	Lecture, Mathématiques	1975, 1996, 1997, 1998, 2000, 2002
Dominique (Ile)	National assesment Test	Dept de l'Éducation	Grade 2 et 6	Lecture, Mathématiques	2006
R. Dominicaine	Sistema dePruebas Nacionales	MdE, IDB, Banque mondiale	Grades 8, 12	Lecture, Mathématiques, Sciences	Annuel 1991-2003
Équateur	APPRENDO	MdE, Banque mondiale, Uni. Catolica	Grades 3, 7, 10	Lecture, Mathématiques	1996,1997, 1998,2000, 2007
El Salvador	SINEA	MdE	Grades 3, 6, 9	Lecture, Mathématiques, Sciences	Tous les deux ans depuis 2001
	SABE	MdE	Grades 1, 2, 3, 4, 5, 6, 9, 11 (variable)	Lecture, Mathématiques, Sciences	Annuel 1993-2001
	PAES	MdE	Grades 10, 12	Lecture, Mathématiques, Sciences	Annuel 1997-2004
Grenade	MCT	Ministry of Education	Grades 2, 4, 6, 8, 9	Lecture, Mathématiques	2004
Guatemala	PRONERE	MdE, Banque mondiale, Valle de Guatemala University	Grades 1, 3, 6	Lecture, Mathématiques	1998,1999, 2000,2004
	Direccion General de Educacion Bilingüe Intercultural	MdE, IDB	Grades 1, 3	Lecture, Mathématiques	2003

Année	Nom	Organisation(s)	Niveau(x) évalué(s)	Domaine(s) évalué(s)	Année(s)
	Sistema Nacional de Medicion del Logro Académico	MdE, Banque mondiale, Valle de Guatemala University	Grades 3, 7, 9, 12	Lecture, Mathématiques, Sciences	Annuel 1992-1996
Guyana	National Assessment		Grades 2, 4, 6, 9	Lecture, Mathématiques	2007

Source : auteurs, à partir des rapports cités.

Notes : ICFES : *Instituto Colombiano para el Fomento de la Educacion Superior*
IDB : *Inter-American Development Bank*
INEP : *Instituto Nacional de Estudos e Pesquisas Educacionais Anisio Teixeira*
MCT : *Minimum Competency Testing*
NCRA : *National Criterion-Referenced Assessment*
ONE : *Operativo National de Evaluación*
PAES : *Prueba de Aptitudes y Aprendizaje para Estudiantes de Educacion Media*
PRONERE : *Programa Nacioal de Evaluacion del Rendimiento Escolar*
SABE : *Strengthening Achievement in Basic Education*
SAEB : *National System for the Evaluation of Basic Education*
SECE : *Sistema de Evaluacion de la Calidad de la Educacion*
SIMECAL : *Sistema de Medicion de la Calidad*
SINEA : *Sistema de Informacion, Monitoreo y Evaluacion de Aprendizajes*

(Partie 2)

Année	Nom	Organisation(s)	Niveau(x) évalué(s)	Domaine(s) évalué(s)	Année(s)
Honduras	UMCE	Banque mondiale, Universidad Pedagógica Nacional Francisco Morazán	Grades 1, 3, 6, 9	Lecture, mathématiques	Variable (1997, 2000, 2004), Annuel depuis 2004 (Grade 1,3), depuis 2005 (Grade 9)
Jamaïque	National Assessment Programme	MdE	Grades 1, 3, 4, 6	Lecture, mathématiques	1999, 2006, 2008
Mexique	Estandares Nacionales	MdE, INEE	Grades 2, 3, 4, 5, 6, 7, 8, 9	Lecture, mathématiques	Annuel 1997-2004
	Sistema Nacional de Evaluacion Educativa de la Educacion Primaria	MdE	Grades 3, 4, 5, 6	Lecture, mathématiques, sciences	Annuel 1996-2000

Année	Nom	Organisation(s)	Niveau(x) évalué(s)	Domaine(s) évalué(s)	Année(s)
	Aprovecha-miento Escolar – Carrera Magistral	MdE, INEE	Grades 3, 4, 5, 6, 7 8, 9	Lecture, mathématiques, Sciences	Annuel 1994-2005
	Instrumento para el Diagnostico de Alumnos de Nuevo Ingreso a Secondaria	MdE	Grade 6	Lecture, mathématiques	Annuel 1995-2005
	EXCALE	MdE	Grades 3, 5, 6,7, 8, 9	Lecture, mathématiques	Annuel depuis 2005 (différents grades)
	ENLACE	MdE	Grades 3, 4, 5, 6, 9	Lecture, mathématiques	Annuel depuis 2006
Nicaragua	Evaluacion del Curriculo Transformado	MdE	Grades 4, 5, 8	Lecture, mathématiques	1996, 1997
	SNE	USAID, UNESCO	Grades 3, 6	Lecture, mathématiques	2002
Panama	Programa de Pruebas de Diagnostico	MdE, various agencies	Grades 3, 6, 12	Lecture, mathématiques	1985, 1986, 1987, 1988, 1992
	CECE	MdE, various agencies	Grades 7, 8, 9, 10, 11, 12	Lecture, mathématiques	1995
	SINECA	MdE, Coordinacion Educativa Cultural Centroamericana	Grades 3, 6, 9, 12	Mathématiques, lecture, Social Studies	1999, 2000, 2001
Paraguay	SNEPE	MdE, IDB	Grades 3, 6, 9, 12	Lecture, mathématiques, sciences	Annuel 1996-2001
Pérou	EN	MdE	Grades 2, 4, 6, 9, 11	Lecture, mathématiques, sciences	1996, 1998, 2001, 2004, 2008

Année	Nom	Organisation(s)	Niveau(x) évalué(s)	Domaine(s) évalué(s)	Année(s)
St. Kitts et Nevis	National Assessment	MdE	Grades 3, 5	Lecture, mathématiques	1998
	MST	MdE	Grades 2, 4	Lecture, mathématiques	2002, 2007
	CEE	MdE	Grade 6	Lecture, mathématiques	2003
St. Lucie	National Assessment	MdE	Grade 5	Lecture, mathématiques	1998
	MST	MdE	Grades 2, 4	Lecture, mathématiques	2002, 2007
Trinidad et Tobago	National Test	MdE	Grades 1, 3	Lecture, mathématiques	2007
	National Test	MdE	Grades 2, 4	Sciences, Social Studies	2008
Uruguay	Unidad de Medicion de Resultados Educativos	Administracion Nacional de Educacion Publica, Banque mondiale	Grades 1, 2, 3, 4, 6 (variable)	Lecture, mathématiques, sciences	1996, 1998, 1999, 2001, 2002
	Programa de Evaluacio de Aprendizajes	MdE	Grade 6	Lecture, mathématiques	Tous les 3 ans depuis 1996
Venezuela	SINEA	MdE, Banque mondiale, Univ. Catolica, Centro Nacional para el Mejoramiento de la Ensenanza en Ciencia	Grade 6	Lecture, mathématiques	1998

Source : auteurs, à partir des rapports cités.

Notes : CECE : *Centro para el Estudio de la Calidad Educativa (Université du Salvador)*
CEE : *Common Entrance Examination*
EN : *Evaluaciones Nacionales*
ENLACE : *Evaluacion Nacional del Logro Académico en Centros Escolares*
EXCALE : *Examen de la Calidad y el Logro Educativos*
INEE : *Instituto Nacional para la Evaluacion de la Educacion*
MST : Minimum Standard Tests
SINECA : *Sistema Nacional de Evaluacion de la Calidad de los Aprendizajes*
SNE : *Sistema Nacional de Evaluacion*
SNEPE : *Sistema Nacional de Evaluacion del Proceso Educativo*
UMCE : *Unidad de Medicion de la Calidad de la Educacion*

Tableau 14 *Évaluations sur les acquis des élèves ou adultes depuis 1959 dans les pays d'Afrique subsaharienne*

(Partie 1)

Année	Nom	Organisation(s)	Niveau(x) évalué(s)	Domaine(s) évalué(s)	Année(s)
Bénin	Evaluation MdE	Ministry of Education, Benin	Grades 1, 4	Lecture, mathématiques	1995
	Evaluation DEP-PAGE	DEP-PAGE	Grade 6	Lecture, mathématiques	2005/2006
	Evaluation ABE LINK	MEPS, USAID	Grades 3, 4, 5, 6	Lecture, mathématiques, autres	2006/2007
Botswana	Standard 4 Attainment Test	BEC	Grade 4	Lecture, mathématiques	2007
	PSLE	BEC	Grade 7	Lecture, mathématiques, sciences	2009
	JCE	BEC	Grade 3	nd	Annuel depuis 2008
	BGCS (Botswana General Certificate of Secondary Education)	BEC	Grade 12	Lecture, mathématiques, sciences	Annuel depuis 2008
Burkina Faso	Processus d'évaluation Scolaire	MdE	Grades 3, 6	Lecture, mathématiques, sciences	2002, 2003, 2004, 2007
Burundi	Évaluation Coopération Française	Coopération Française	Last grade of primary education	Lecture, mathématiques	1989
	MLA	UNICEF-UNESCO	Grade 4	Lecture, mathématiques	2001
Cameroun	MLA	UNICEF-UNESCO	Grade 4	Lecture, mathématiques	1998/1999
Comores	Assessment	MdE	Grade 8	Mathématiques, sciences	2004

Année	Nom	Organisation(s)	Niveau(x) évalué(s)	Domaine(s) évalué(s)	Année(s)
RDC	Enquête sur l'évaluation des acquis scolaires des élèves de la 5e année primaire	MdE	Grade 5	Lecture, mathématiques, culture	1994
	Évaluation UNICEF	UNICEF	Last year of primary	Lecture	1999/2000
	Tests d'évaluation des acquis scolaires	Direction des études et de la planification du MEPSP	nd	nd	nd
	MLA	UNICEF-UNESCO	Grade 4	Lecture, mathématiques	nd
Côte d'Ivoire	Projet appui au secteur	INEADE, SEDEP (Liège)	Grade 2,4 et 6	Lecture, mathématiques	2002
Erythrée	Learning Achievement	Ministry of Education	Grades 1,5	Lecture, mathématiques en G5, compréhension langue locale en G1	1996
Ethiopie	Sample Baseline on Student Learning Assessment	National Organization for Examinations	Grades 4,8	Lecture, mathématiques, sciences	2000, 2004
Gambie	National Test	MdE	Grades 2, 4, 6	Lecture, mathématiques	1997, 1998, 1999, 2000
	MLA	UNICEF-UNESCO	Grades 3,5	Lecture, mathématiques, General Sciences, Social and Environmental Studies	2000
Ghana	CRT	Ghana Education Service	Grade 6	Lecture, mathématiques	1997
	Evaluation of Implementation of Ghana's School Language Policy	USAID and IEQ	Grades 1, 2, 3, 4	Lecture	1999, 2000, 2001

Année	Nom	Organisation(s)	Niveau(x) évalué(s)	Domaine(s) évalué(s)	Année(s)
	NEAT	Ghana Education Service	Grades 3,6	Lecture, mathématiques	2005, 2007
Guinée	Évaluation du niveau des élèves	Cellule Nationale de Coordination des Évaluations du Système Éducatif	Grades 2, 4, 6	Lecture, mathématiques	Annuel 1997- 2000
Lesotho	PEP	Ministry of Education, USAID	Grades 3, 6	Lecture, mathématiques	1993
	National Assessment	Ministry of Education – WB	Grades 3, 6	Lecture (anglais et sesotho), mathématiques	2005
Liberia	MLA	UNICEF-UNESCO	Grade 4	Lecture, mathématiques, sciences	2000

Source : auteurs, à partir des rapports cités.

Notes : ABE LINK : *Assistance to Basic Education / Linkages in Education aud Health*
BEC : *Botswana Examination Council*
CRT : *Criterion Referenced Testing*
DEP-PAGE : *Direction de l'enseignement primaire et projet d'appui à la gestion de l'éducation*
IEQ : *Improving Edicational Quality*
JCE : *Junior Certificate Examination*
MEPS : *Ministère des Enseignements primaire et secondaire*
PEP : *Primary Education Project*
PSLE : *Primary School Leaving Examination*

(Partie 2)

Année	Nom	Organisation(s)	Niveau(x) évalué(s)	Domaine(s) évalué(s)	Année(s)
Madagascar	Evaluation des acquis scolaires	MdE, Direction de la Planification de l'éducation	Grade 4	Lecture, mathématiques	1999
	MLA	UNICEF-UNESCO	Grade 4	Lecture, mathématiques, sciences	1998
	Étude sur la progression scolaire et la performance acacémique à Madagascar	MdE	Grades 2, 5	Lecture, mathématiques, Life Skills	2005
	Étude sur la progression scolaire et la performance acacémique à Madagascar	MdE, Cornell University	Ages 7, 14	Lecture	2004
Malawi	MLA	UNICEF-UNESCO	Grades 4, 6	Lecture, mathématiques, sciences	1999
	Primary Schools Learner Achievements Level	Annual Basic Education Statistics Census	Grades 3, 5, 7	Lecture, mathématiques	Annuel depuis 2004
	Quality of Learning and Teaching in Developing Countries	DFID	Grade 4	Lecture	1996, 1997, 1998
	Lecture in English in Primary Schools	DFID	Grades 3, 4, 6	Lecture	1993
	Lecture Levels and Bilingual Literacy in Primary Schools	DFID	Grades 3, 4, 5, 6	Lecture	1998

Année	Nom	Organisation(s)	Niveau(x) évalué(s)	Domaine(s) évalué(s)	Année(s)
	Literacy Development through a Local Language in a Multilingual Setting	USAID, IEQ	Grades 2, 3, 4	Lecture	1999, 2000
Mali	Évaluation	Centre National de l'éducation	Grades 2, 4, 6	Lecture, mathématiques	2007
Mozambique	Assessment	Ministry of Education	Grades 2, 3	Lecture, mathématiques, sciences	1999
	Assessment	Ministry of Education	Grade 10	Lecture, mathématiques, sciences	2005
Namibie	National Learner Baseline Assessment	Ministry of Education	Grades 4 et 7	Lecture, mathématiques	Annuel depuis 2001
Niger	MLA	UNICEF-UNESCO	Grade 4	Lecture, mathématiques, sciences	nd
Nigeria	MLA	UNICEF-UNESCO	Grade 4	Lecture, mathématiques, Sciences	1994
	Universal Basic Education Programme	UBEC	Grades 1,4, 5,6	Lecture, mathématiques, sciences, sciences sociales	2001, 2003
Sénégal	SNERS	nd	nd	nd	nd
	MLA	UNICEF-UNESCO	Grades 3, 6	Lecture, mathématiques, sciences	nd
Afrique du Sud	Systemic Evaluation	MdE, HSRC	Grade 3	Lecture, mathématiques, Life Skills	2001
	Systemic Evaluation	MdE, HSRC	Grade 6	Lecture, mathématiques, Life Skills	2005

Année	Nom	Organisation(s)	Niveau(x) évalué(s)	Domaine(s) évalué(s)	Année(s)
	Systemic Evaluation	MdE, HSRC mathématiques	Grade 3	Lecture,	2007
	MLA	UNICEF-UNESCO	Grade 4	Lecture, mathématiques, sciences	1999
	Annual National Assessments	MdE	Grades 3, 6	Lecture, mathématiques	2008-2011
	Learner Assessment Results	HSRC, District Development Support Programme, USAID	Grade 3	Lecture	2003
	Monitoring Education Quality	HSRC	Grade 9	Lecture, mathématiques, sciences	Annuel depuis 1996
Ouganda	NAPE	MdE	Grades 3, 6	Lecture, mathématiques	2008
	NST	MdE	Grade 8	Lecture, mathématiques, sciences	2008
Zambie	NAS	MdE	Grade 5	Lecture, mathématiques	1999, 2001, 2003, 2000
	Lecture Levels and Biligual Literacy in Primary Schools	DFID	Grades 3, 4, 5, 6	Lecture	1998
	Primary Lecture Programme	ADEA	Grades 1, 2, 3, 4 ,5, 6	Lecture	1999, 2002

Source : auteurs, à partir des rapports cités.

Notes : ADEA : *Association for the Development of Education in Africa*
HSRC : *Human Sciences Research Council*
UBEC : *Universal Basic Education Commission*

2. Approche comparative des évaluations sur les acquis des élèves

Introduction

Cette partie a pour objectif de mener une analyse technique et comparative des méthodes adoptées lors des enquêtes sur les acquis des élèves, tant au niveau national qu'au niveau international. Pour ce faire, nous proposons d'explorer plusieurs dimensions relatives aux enquêtes sur les acquis et la performance cognitive. Les personnes évaluées peuvent être tout autant des élèves, des adultes ou encore des jeunes non (ou peu) scolarisés. Cependant, nous nous concentrerons ici le plus souvent sur les évaluations des élèves.

Les différentes dimensions de notre analyse sont présentées en cinq points principaux.

- Le premier traite des caractéristiques statistiques et techniques des tests. Nous présentons les différentes procédures d'évaluation utilisées par les enquêtes. Les techniques d'échantillonnages et les problèmes qu'ils peuvent entraîner sont aussi analysés.
- Le second analyse la nature du contenu des tests en précisant les différentes approches utilisées pour évaluer les élèves (savoirs *versus* compétences), mais aussi en traitant du type de population évalué. Il est également question de l'angle de mesure d'indicateurs de marginalisation qui pourraient être utiles dans le cadre d'une politique de suivi des inégalités dans les acquis scolaires au sein de pays en développement, dans le cadre d'un projet tel que *Education for All - Fast Track Initiative* (EFA FTI).
- Le troisième se concentre sur la population cible en soulignant les différences de sélection des grades évalués mais aussi des critères d'exclusion des populations qui pourraient différer selon les tests.
- Le quatrième point traite de la question de la fiabilité des tests sous l'angle de la validité de ceux-ci ainsi que de leur crédibilité en tant qu'évaluation de haut niveau.

- Le cinquième enfin analyse la comparabilité des résultats sous plusieurs dimensions : au-delà des possibilités de comparaison temporelle, nous explorons les options disponibles pour des comparaisons internationale et intranationale.

À partir de ces différents points, nous présenterons sous forme de synthèse les avantages et inconvénients de chaque type d'évaluation et leurs différents apports aux ministères de l'Éducation. Par ailleurs, des recommandations pour l'amélioration du PASEC seront proposées.

2.1. Caractéristiques statistiques des tests

2.1.1. *Aspects techniques de la procédure d'évaluation*

Cette section tente d'expliquer comment les tests ont été élaborés d'un point de vue statistique, et plus spécifiquement psychométrique. Certaines méthodes telles que la méthode de Rasch ou plus généralement d'*Item Response Theory* (IRT) sont de plus en plus utilisées. De manière générale ces méthodes retracent, à partir de tests effectués par un échantillon d'individus, une échelle de mesure qui puisse décrire, au travers d'une fonction continue, la répartition des compétences de chacun. La méthode IRT, ou encore *Item Response Modeling* (IRM), est une méthode de mesure psychométrique (*cf.* Hambleton et Swaminathan, 1985 ; Bottani et Vrignaud, 2005 ; Laveault et Gregoire, 2002). Elle permet de vérifier la validité de la mesure en psychométrie. En effet, en psychométrie, toute mesure, étant une construction, ne reflète pas systématiquement et parfaitement ce qui est mesuré. Dans le programme PISA, par exemple, où sont mesurées les compétences en mathématiques des élèves, les items administrés sont une construction tendant à s'approcher de ces compétences. Elles ne sont pas une mesure directe de ces compétences. Il faut donc valider les items afin de s'assurer qu'ils mesurent bien la compétence désirée.

La validation de cette construction nécessite avant tout de pouvoir émettre des hypothèses sur sa nature et son fonctionnement. On parlera donc d'un modèle de mesure, et la démarche hypothético-déductive consiste à tester l'adéquation de ce modèle aux données. Plusieurs approches peuvent être mises en œuvre pour tester cette adéquation (entre les items du test et les compétences évaluées) ; parmi elles, trois sont plus généralement utilisées : l'approche classique (formalisée par Lord et Novick, 1968), l'IRT et les modèles structuraux.

Parmi ces méthodes, la plupart des enquêtes utilisent l'IRT. Créée il y a une quarantaine d'année par Georg Rasch et Allan Birnbaum, elle suit des modèles probabilistes : on suppose que la probabilité qu'un sujet *j* donne une réponse correcte à un item *i* est

fonction de la compétence du sujet et de la difficulté de l'item. Le modèle le plus général comprend trois paramètres pour modéliser le fonctionnement de l'item (la difficulté de l'item, la pente ou sélectivité de l'item, le paramètre de réponse au hasard). Le modèle de Rasch est un modèle particulier de la méthode IRT. Il s'agit le plus souvent d'un modèle IRT à un seul paramètre[27].

Le tableau 15 présente quelques caractéristiques propres aux évaluations. La plupart ont recours à la méthode d'estimation IRT, à l'exception du PASEC et du test EGRA. Le nombre de paramètres utilisés dans l'estimation varie de 1 pour PISA (méthode de Rasch) à 3 pour TIMSS. Cependant, comme le note Wu (2010), il n'y a à ce jour aucun travail permettant de voir quelle est la méthode la plus efficace pour estimer les moyennes et les erreurs type. Pour certaines évaluations, les élèves n'ont pas à répondre à l'intégralité du test. Dans PISA, par exemple, pour optimiser les temps de passation, certains élèves n'ont répondu qu'au test de mathématiques, sans avoir été évalués en sciences. La technique d'imputation des données basées sur l'échelle IRT permet de retrouver le score de l'élève avec un degré d'erreur minimisé puisque basé sur un plus large ensemble d'items. Celui-ci est possible avec les valeurs plausibles qui sont disponibles dans les bases internationales (TIMSS, PIRLS et PISA). De tels outils ne présentent aucune utilité dans SACMEQ, LLECE, PASEC ou EGRA, dans la mesure où les élèves sont testés sur l'intégralité des items disponibles[28]. La méthode de réplication permet d'avoir des erreurs standards non biaisées. Une divergence existe entre PISA et l'IEA : le premier utilise la méthode *Balances Repeated Replication* (BRR), tandis que l'IEA a recours à la méthode de Jacknife. La prochaine réforme de la méthodologie du PASEC devrait conduire à utiliser la méthode de Jacknife adoptée par l'IEA. La question de la stratification est importante car elle pourra nous donner des indications précieuses sur l'hypothèse de comparabilité intranationale. En règle générale, les évaluateurs choisissent d'abord les écoles, puis ensuite des classes. Cependant, dans le cadre de PISA, ce sont les élèves qui constituent le *cluster* numéro 2. Aucune classe entière n'est systématiquement sélectionnée, ce qui inhibe toute possibilité d'évaluer des effets « enseignants » dans cette évaluation. La stratification des enquêtes régionales SACMEQ et LLECE se base d'abord sur des régions puis sur des écoles. Cependant, les tests PASEC ont parfois des méthodes différentes de stratification afin de faire ressortir des caractéristiques spécifiques d'organisation de certains systèmes éducatifs. Ainsi, pour certaines évaluations, les établissements privés de l'ensemble

[27] Des extensions récentes ont permis d'effectuer des modèles de Rasch à plusieurs paramètres.

[28] Ceci peut d'ailleurs conduire à remettre en question l'hypothèse d'indépendance des tests. Dans SACMEQ, par exemple, les élèves pouvant être testés durant 4 heures, on peut se demander dans quelle mesure ils répondent jusqu'au bout, avec sérieux, au questionnaire contextuel.

national constituent une strate de premier niveau, alors que les établissements publics sont sélectionnés dans des strates régionales. À titre d'exemple, l'évaluation de Maurice, en 2006, a mis aussi l'accent sur les écoles situées en zones d'éducation prioritaires (ZEP[29]). La stratification a donc inclus ce critère au détriment des zones géographiques[30]. Ce trait d'hétérogénéité propre au PASEC le différencie fortement du SACMEQ, où les régions sont représentées de façon précise, puisqu'elles constituent toujours le premier critère de stratification. Par le biais de la méthode IRT, une standardisation des scores est possible, et c'est d'ailleurs ce qu'effectuent la plupart des enquêtes pour obtenir une moyenne globale de 500 points et un écart-type de 100 points. Seules les évaluations EGRA et PASEC utilisent d'autres méthodes.

Tableau 15 *Caractéristiques statistiques des tests*

Enquête	Modèle d'estimation	Nombre de paramètres	Méthode de *scoring*	Méthode de réplication	Stratification	Standardisation des scores
TIMSS	IRT	3 paramètres	5 valeurs plausibles	*Jacknife replication method*	2 étapes : écoles puis classes entières	Moyenne 500 Écart-type 100
PIRLS	IRT	3 paramètres	5 valeurs plausibles	*Jacknife replication method*	2 étapes : écoles puis classes entières	Moyenne 500 Écart-type 100
PISA	IRT	1 paramètre	5 valeurs plausibles	*Balanced repeated replication*	2 étapes : écoles puis élèves au hasard	Moyenne 500 Écart-type 100
SACMEQ	IRT	-	Simple score	-	2 étapes : régions puis écoles	Moyenne 500 Écart-type 100
LLECE	IRT	-	Simple score	-	2 étapes : régions puis écoles	Moyenne 500 Écart-type 100

[29] Une école classée ZEP est une école située dans une région plus pauvre que la moyenne et qui bénéficie de subventions et de soutiens divers de la part du gouvernement. Ce système s'inspire du modèle éducatif français de ZEP.

[30] Les écoles ZEP ont été surreprésentées par rapport aux autres écoles.

Enquête	Modèle d'estimation	Nombre de paramètres	Méthode de *scoring*	Méthode de réplication	Stratification	Standardisation des scores
PASEC	-	-	Simple score	-	2 étapes : strates puis écoles	Note maximale 100
EGRA	-	-	Simple score	-	2 étapes : régions puis écoles	Nombre de mots lus correctement par minute

Source : auteurs, à partir des rapports cités.

Les modèles IRT reposent sur de nombreuses conditions de validité : unidimensionnalité [31], indépendance conditionnelle des items [32] et égal pouvoir discriminant des différents items [33]. Dans la mesure où la plupart des enquêtes internationales et régionales [34] utilisent cette procédure, il est important de se demander quelles sont les limites de cette approche.

D'après Bottani et Vrignaud (2005), plusieurs critiques peuvent être adressées à la méthode IRT. Une première, liée au difficile jugement de l'adéquation de ces modèles aux données, est une limite importante (critique applicable à toute tentative de modélisation) : l'estimation des paramètres des modèles IRT repose sur de nombreuses conditions analytiques de validité qui sont parfois difficiles à tenir et à vérifier dans la variété des cas pratiques. Ainsi, Hambleton *et al.* (1991) recensent les procédures à mettre en œuvre pour s'assurer de la possibilité d'application du modèle des données ; plusieurs étapes sont nécessaires, et celles-ci sont coûteuses en termes de capacités techniques, de temps et donc d'argent (Flieller, 1989).

Une deuxième critique porte sur la réalité psychologique du modèle. Ainsi, Reuchlin (1997) remet en cause le caractère continu du modèle IRT, qui présuppose qu'un sujet

[31] L'unidimensionnalité de la variable latente présuppose que les différences interindividuelles ne sont que des différences de puissance, et que les différences de difficulté entre items ne sont que des différences quantitatives. Il est donc admis 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.

[32] L'indépendance conditionnelle présuppose que, pour un niveau de compétence donné, la réussite à un item quelconque est indépendante de la réussite aux autres items.

[33] 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'informations sur la compétence du sujet, un item peu discriminant renseigne peu sur la compétence du sujet.

[34] Seules quelques enquêtes telles PASEC et MLA n'utilisent pas cette procédure. Cependant, il faut rappeler que MLA est désormais stoppée et que l'enquête PASEC doit inclure la procédure IRT dans la mesure des savoirs des élèves.

peut toujours réussir un item et que la réponse à un item a donc un caractère discret. Or, la réussite à un item difficile n'est pas peu probable pour un sujet peu compétent, elle est fonctionnellement impossible. L'unidimensionnalité de la variable latente présuppose 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. Or, il est tout à fait possible que les stratégies des élèves/adultes diffèrent du fait même des caractéristiques personnelles[35]. L'hypothèse d'unidimensionnalité justifie l'utilisation de la variable mesurée pour classer les sujets sur un continuum. En particulier, Harvey Goldstein et d'autres chercheurs ont montré, en appliquant les modèles d'équations structurales aux données anglaises et françaises, que les données n'étaient pas unidimensionnelles, mais au minimum bidimensionnelles (voir notamment Goldstein, 2004 ; Goldstein *et al.,* 2007). L'écart à l'unidimensionnalité est révélateur de failles dans le dispositif de mesure et doit recevoir une interprétation psychologique (Bottani et Vrignaud, 2005, p.103). On peut aussi souligner que l'unidimensionnalité s'accorde plus à une vision dichotomique de la réponse ; à l'inverse, une réponse parmi des choix multiples dépend étroitement des stratégies individuelles de prise de risque face à l'erreur, ce qui individualise donc, dans ce cas, le seuil de réponse au hasard du modèle.

Une troisième limite concerne l'hypothèse d'indépendance locale. En effet, la plupart des modèles psychométriques nécessitent de faire l'hypothèse que la réponse d'un sujet à un item ne dépend pas de ses réponses aux autres items de l'instrument. Or, il est souvent difficile de tester l'hypothèse d'indépendance conditionnelle. Il est, en revanche, aisé d'identifier des situations où, par construction, cette hypothèse est violée : dans le cas, par exemple, des items en cascade lorsque l'on cote la réponse et sa justification. D'après Bonora et Vrignaud (1997), on cherche rarement à vérifier que les réponses des sujets sont bien indépendantes, ou plutôt qu'on ne peut pas mettre en évidence de dépendance entre les réponses aux items. Dans l'évaluation de la lecture dans PISA, PIRLS, SACMEQ et LLECE, il est souvent demandé de répondre à plusieurs questions posées sur le même texte. Si cette approche est justifiée pour des raisons cognitives et temporelles, il est rare que les psychométriciens à l'origine de la réalisation des enquêtes évoquent des biais induits par cette dépendance dans le traitement des résultats des enquêtes internationales sur la littératie (Dickes et Vrignaud, 1995 ; Vrignaud, 2005). Ces biais ont pourtant des effets non négligeables, comme l'ont montré les quelques recherches réalisées sur cette question (Wainer et Thissen, 1996).

[35] Mislevy a, par exemple, proposé des modèles prenant en compte des différences entre populations en particulier lorsque l'on fait l'hypothèse que les sujets emploient des stratégies différentes (Mislevy et Vergelst, 1990).

2.1.2. *Techniques d'échantillonnage des populations*

Les enquêtes internationales supposent que la performance des populations évaluées suit une distribution normale (d'un point de vue statistique), ce qui reste problématique dans le cas de pays où l'hétérogénéité apparaît importante. La procédure de normalisation selon certaines dimensions telles que la région géographique ou le groupe ethnolinguistique d'appartenance peut conduire à des biais à l'intérieur du test.

Par ailleurs, les niveaux d'exigence des évaluations changent fortement. À titre d'illustration, le tableau 16 souligne les principales caractéristiques des échantillons à travers les différentes enquêtes. C'est dans PISA que le nombre minimum d'élèves est le plus élevé (plus de 5 000). Au contraire, l'évaluation EGRA ne pose que la limite de 400 élèves. Or, plus le nombre d'élèves est important, plus la précision des outils de mesure sera forte. Il est alors assez aisé de comprendre pourquoi l'évaluation EGRA annonce rarement, dans les rapports publiés, pouvoir représenter fidèlement la population d'un pays. Le niveau minimum du PASEC peut sembler faible au regard des pays lusophones prenant part à l'évaluation LLECE : le seuil de 2 400 élèves est bas comparé aux 4 000 élèves de LLECE. Ceci s'explique notamment par le fait que le PASEC évalue les élèves à deux périodes dans la même année scolaire, réduisant d'autant le nombre d'élèves afin d'éviter d'augmenter les coûts de l'évaluation.

L'utilité de prendre des classes entières renvoie à l'opportunité de pouvoir analyser plusieurs niveaux simultanément : région, école, classe et élève. Or, dans le test PISA, les élèves sont choisis de manière aléatoire au sein des écoles, ce qui élimine le niveau « classe » des analyses. Au contraire, l'avantage du PASEC et des tests de l'IEA est de se centrer sur ce niveau[36]. Grâce au questionnaire dédié aux enseignants, il est possible de mieux appréhender ce qui se passe dans la classe. Le nombre minimum d'écoles renvoie souvent à une limite inférieure de 150 écoles, sauf pour le SACMEQ où il peut descendre à 25. Ceci est surtout dû au fait que certains pays/zones sont faiblement peuplé(e)s (les Seychelles et Zanzibar, essentiellement).

[36] Selon la taille de l'école, parfois plusieurs classes sont incluses dans les échantillons des pays participant aux évaluations de l'IEA. Toutefois, cela n'a que très peu d'impact sur la variabilité intra-école, qui reste souvent cantonnée à une variabilité intraclasse.

Tableau 16 *Critères d'échantillonnage des différentes enquêtes*

Enquête	Nombre minimum d'élèves	Nombre minimum de classes	Nombre minimum d'écoles	Nombre d'élèves par école
TIMSS	4 000	1 ou 2 par école	150	1 ou 2 classe(s) entière(s)
PIRLS	4 000	1 ou 2 par école	150	1 ou 2 classe(s) entière(s)
PISA	5 250	Pas de sélection	150	35 élèves par école au hasard
SACMEQ	Indéterminé	Pas de sélection	Pas de limite (de 25 à 275, avec une moyenne de 165)	20 élèves par école
LLECE	4 000 (grade 3) 3 500 (grade 6)	Pas de sélection	150	Classes entières
PASEC	2 400	2 classes par école (tirage aléatoire)	150	15-20 élèves/ classe soit 30 élèves minimum au pré test par école
EGRA	400	Pas de sélection	Pas de limite	Pas de limite (souvent 15)

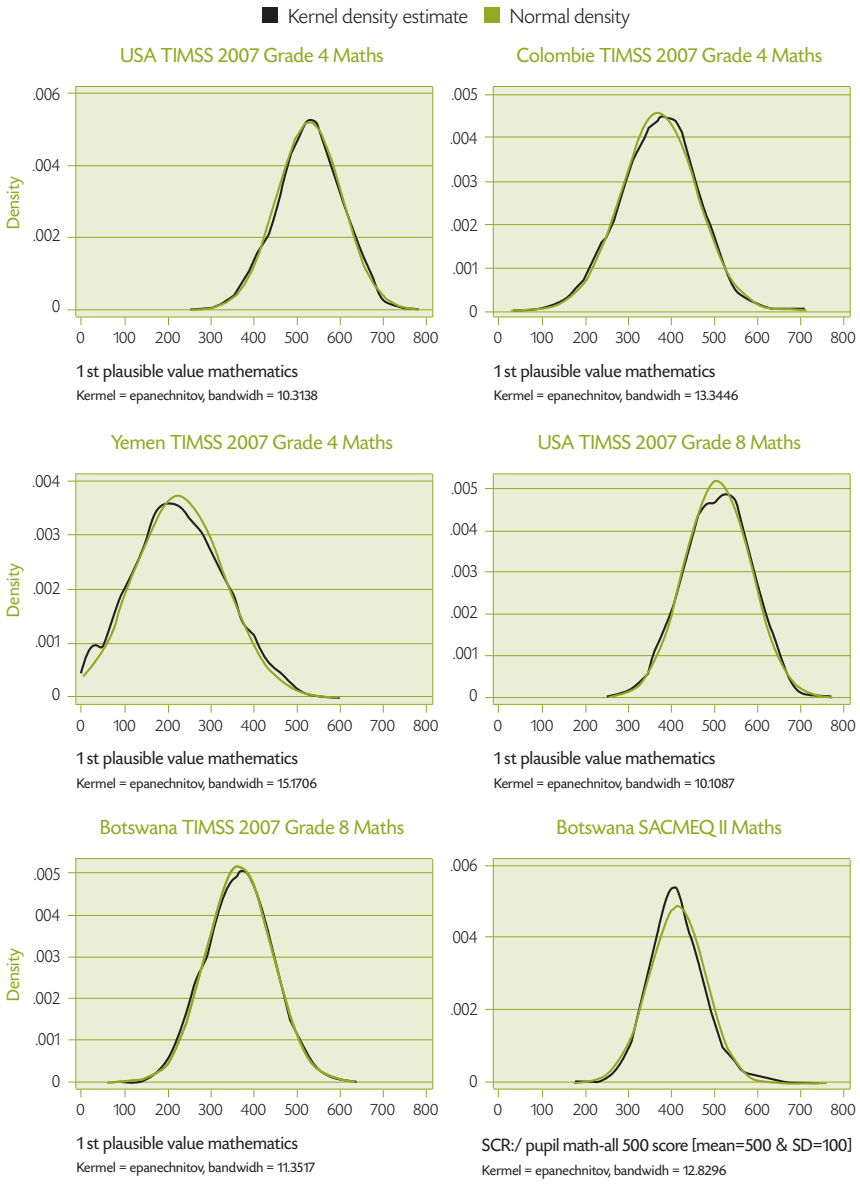
Source : auteurs, à partir des rapports cités.

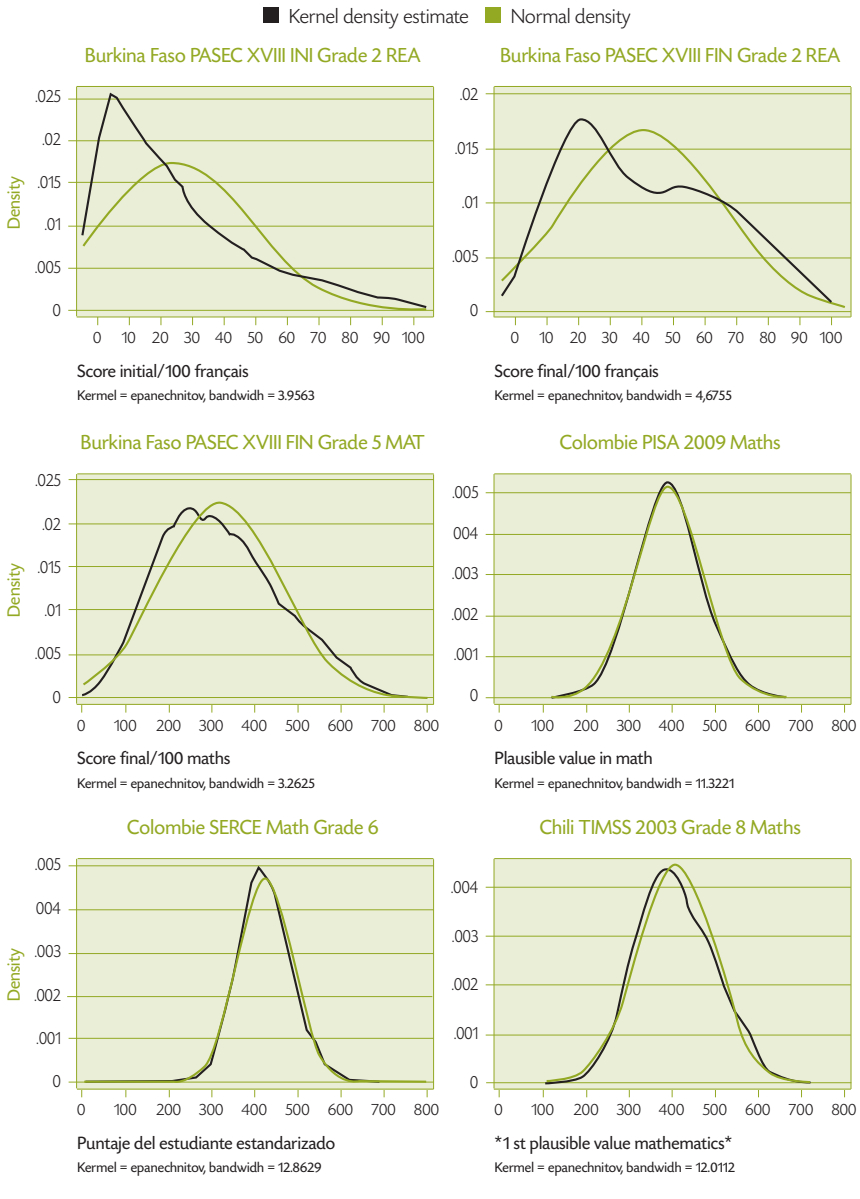
Afin d'analyser la distribution normale des échantillons d'élèves dans chaque évaluation, nous avons tracé les distributions kernell des scores en mathématiques. Celles-ci sont présentées dans le graphique 7. En toute logique, la distribution doit être normale pour accepter l'idée selon laquelle l'évaluation cible bien le niveau moyen de la population. Lorsque la courbe est trop tirée vers la gauche, cela signifie que beaucoup d'élèves ont obtenu des scores faibles. De façon symétrique, pour les courbes trop étirées vers la droite, le test semble avoir été trop simple pour la population évaluée. Un exemple assez parlant est le cas de la participation de trois pays (États-Unis, Yémen et Colombie) à l'enquête TIMSS 2007 (première ligne). L'observation de la courbe des États-Unis montre une distribution quasi normale. Par ailleurs, comme le niveau minimum d'acquis est de 400 points, on voit qu'une grande partie de la population

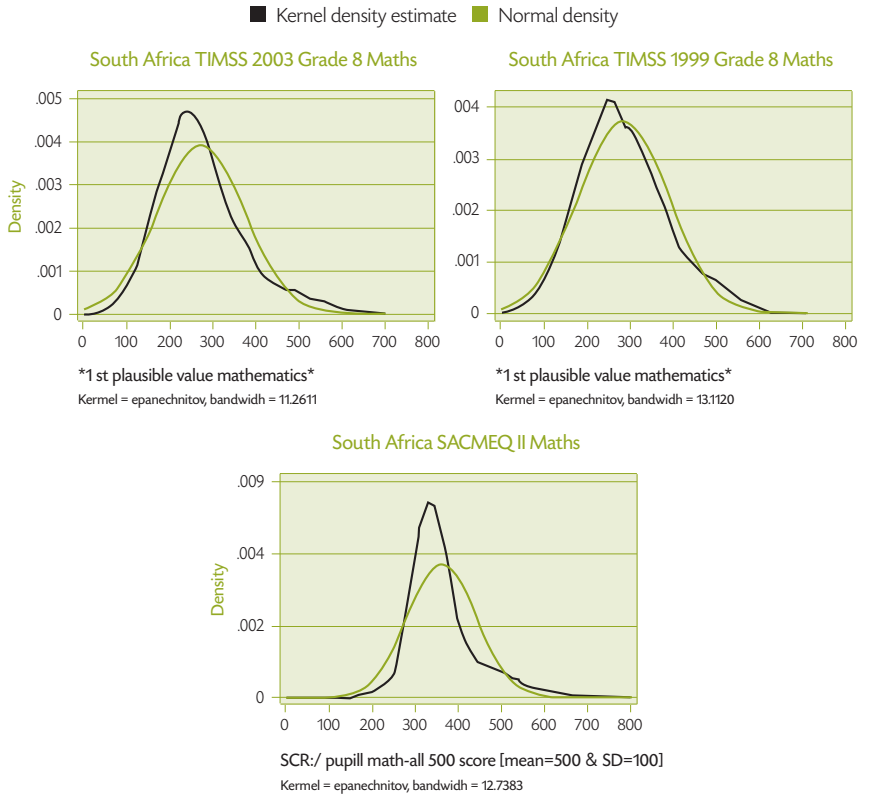
a atteint ce niveau minimum. Pour la Colombie, on observe une légère déviation vis-à-vis de la loi normale. Cependant, plus de la moitié de la population obtient un score inférieur au seuil minimum, ce qui témoigne d'une difficulté trop importante du test soumis aux élèves de ce pays. Le cas du Yémen (troisième colonne) est encore plus parlant : très peu d'élèves dépassent le seuil minimum fixé par l'IEA. Or, le test a été élaboré de telle façon à discriminer la performance des élèves à travers des seuils de performance allant de 400 à plus de 625 points. On constate ici que très peu d'élèves yéménites parviennent à atteindre la moyenne internationale de 500 points. Il devient alors pertinent de se demander quel est l'apport d'une participation à l'évaluation TIMSS pour un pays ayant un niveau de performance aussi faible ? Comment peut-on mesurer efficacement le niveau d'acquis des élèves si ceux-ci échouent presque tous à atteindre le seuil minimum fixé par les évaluateurs ?

D'autres exemples sont présentés dans ce graphique, dont le cas de l'évaluation PASEC. La distribution des scores du pré-test du Burkina Faso en 2006 est présentée dans la première colonne de la quatrième ligne. On observe nettement une proportion très élevée d'élèves qui ne parviennent pas à atteindre le score de 10 sur une échelle allant de 0 à 100 points. La distribution ne suit pas du tout une évolution normale, ce qui souligne le manque de pertinence du test par rapport à la population évaluée. Lorsque l'on analyse le test de fin d'année, l'évaluation semble plus normale, même si une déviation forte vers la moindre performance est toujours visible. Le pic des 10 points se déplace vers le seuil de 20 points. En ayant à l'idée que le seuil désiré de performance de ce test est fixé à 40 points, il apparaît peu approprié pour plus de la moitié de la population. La deuxième colonne de la quatrième ligne et la cinquième ligne montrent les distributions relatives aux évaluations PISA, SERCE et TIMSS de pays d'Amérique latine. Les distributions sont proches de la normale, même si le niveau des scores varie entre les enquêtes. L'écart-type de la performance du Chili semble cependant plus élevé dans le cadre du test TIMSS 2003. Enfin, dans la dernière ligne, nous avons représenté les distributions des scores pour l'Afrique du Sud. Dans les évaluations TIMSS 1999 et 2003, les distributions sont clairement en-dessous du seuil minimum de 400 points. Cela remet clairement en cause la légitimité d'une telle participation. C'est d'ailleurs en partie pour cette raison que l'Afrique du Sud n'a pas souhaité prendre part à l'enquête TIMSS 2007. Cependant, la faible diversité des scores obtenue dans l'évaluation SACMEQ II montre aussi les limites d'un test standardisé : les scores sont proches les uns des autres, ce qui limite fortement l'analyse de la distribution des scores et de leur explication par des facteurs individuels et scolaires.

Graphique 7 *Distributions kernel des scores des populations testées dans diverses enquêtes régionales et internationales*




Kernel density estimate
Normal density
Burkina Faso PASEC XVIII INI Grade 2 REA
Density
.025
.02
.015
.01
.005
0
0 10 20 30 40 50 60 70 80 90 100
Score initial/100 français
Kermel = epanechnitov, bandwidh = 3.9563
Burkina Faso PASEC XVIII FIN Grade 2 REA
.02
.015
.01
.005
0
0 10 20 30 40 50 60 70 80 90 100
Score final/100 français
Kermel = epanechnitov, bandwidh = 4,6755
Burkina Faso PASEC XVIII FIN Grade 5 MAT
Density
.025
.02
.015
.01
.005
0
0 100 200 300 400 500 600 700 800
Score final/100 maths
Kermel = epanechnitov, bandwidh = 3.2625
Colombie PISA 2009 Maths
.005
004
.003
.002
.001
0
0 100 200 300 400 500 600 700 800
Plausible value in math
Kermel = epanechnitov, bandwidh = 11.3221
Colombie SERCE Math Grade 6
Density
.005
004
.003
.002
.001
0
0 100 200 300 400 500 600 700 800
Puntaje del estudiante estandarizado
Kermel = epanechnitov, bandwidh = 12.8629
Chili TIMSS 2003 Grade 8 Maths
.004
.003
.002
.001
0
0 100 200 300 400 500 600 700 800
1 st plausible value mathematics
Kermel = epanechnitov, bandwidh = 12.0112



Source : auteurs.

Des problèmes inhérents à la pondération des élèves peuvent aussi être importants. Pour plusieurs raisons, dont la plus importante est l'économie de coûts, les enquêtes utilisent des pondérations afin de représenter le plus justement possible les populations testées. Il est inutile de tester l'ensemble d'une classe d'âge ou d'une population scolarisée dans un grade spécifique. En effet, les techniques modernes d'échantillonnage permettent de prédire fidèlement la performance d'une population à partir d'un échantillon prédéfini. Afin de la représenter le plus fidèlement, une pondération spécifique est également souvent nécessaire. Par exemple, lorsque certaines zones sont difficilement accessibles dans les pays testés, les élèves de ces zones se voient conférer des pondérations plus importantes que les autres. D'autres facteurs sont également utilisés pour déterminer la valeur des pondérations. Les critères les plus communs

sont le genre de l'élève, le lieu géographique, le type d'école (public/privé, général/technique/professionnel/religieux) ou encore sa taille. Cependant, la détermination de la pondération nécessite de connaître avec exactitude la population scolarisée durant l'année scolaire testée. Ceci peut s'avérer pour le moins complexe pour les pays qui ont des difficultés à établir des statistiques sur l'éducation. Des biais importants peuvent donc apparaître, d'autant plus que les enquêteurs utilisent le plus souvent des données statistiques issues des années scolaires précédentes. Malheureusement, à ce jour, très peu d'analyses ont été faites sur la pertinence des pondérations. Une étude récente sur le cas de l'Autriche montre comment la définition des pondérations peut altérer les résultats scolaires des élèves (Neuwirth, 2006). Ce dernier souligne que l'Autriche a mal défini les pondérations dans l'enquête PISA 2000, ce qui aurait conduit à des résultats erronés lors de la comparaison avec les scores de 2003. Comme le montre le tableau 17, l'analyse de la variation des scores avec les pondérations données par l'OCDE montrerait une diminution très importante de la performance des élèves. Ce serait notamment le cas des garçons en sciences, qui perdraient près de 36 points entre 2000 et 2003. Or, d'après l'analyse de Neuwirth, après avoir utilisé des mesures plus adéquates de la pondération des élèves, la baisse serait largement réduite, et rendue non significative d'un point de vue statistique. Plus particulièrement, l'évolution des scores des garçons en sciences ne serait revue à la baisse qu'à hauteur de 17,4 points, soit près de la moitié de celle trouvée avec des pondérations erronées.

Tableau 17 *Variation des scores suite à des problèmes de pondération – Autriche, PISA 2000 et PISA 2003*

	Lecture		Mathématiques		Sciences	
	Filles	Garçons	Filles	Garçons	Filles	Garçons
Scores donnés dans les rapports						
PISA 2000	520,3	494,6	503,0	530,1	513,9	525,7
PISA 2003	514,4	467,1	501,8	509,4	492,3	489,7
Variation	-5,9	-27,5	-1,2	-20,7	-21,6	-36,0
Scores ajustés par Neuwirth						
PISA 2000	509,2	475,8	492,5	512,0	502,2	507,1
PISA 2003	514,4	467,1	501,8	509,4	492,3	489,7
Variation	5,2	-8,7	9,3	-2,6	-9,9	-17,4

Source : Neuwirth (2006).

2.2. Nature des contenus évalués

2.2.1. Collecte des données

La collecte des données varie sensiblement entre les enquêtes. Les évaluations EGRA présentent des différences notables, notamment dans leur procédure d'évaluation orale directe et rapide, contrairement aux évaluations internationales de l'IEA ou encore de l'OCDE qui sont traditionnellement réalisées à l'écrit avec papier et crayon. En effet, les enquêtes autres qu'EGRA ont recours à des évaluations à l'écrit, ce qui suppose que les élèves sachent lire et écrire[37]. Pour autant, dans les pays en développement principalement, une certaine proportion d'élèves peut être analphabète (ou peu alphabétisée), ce qui entre en contradiction avec un test écrit. Par ailleurs, les enquêtes évaluent des niveaux généralement supérieurs au grade 3 ; or, le fait de pouvoir tester des élèves du grade inférieur peut permettre de détecter de possibles problèmes afin de les résoudre[38]. Une principale avancée d'EGRA a donc été d'évaluer oralement les élèves dans un délai très court (inférieur à 15 minutes). Grâce à cet outil, il devient possible de mesurer le degré de maîtrise de la langue testée, avant même d'évaluer le niveau de compétence de l'élève. Au contraire d'EGRA, le test PASEC du grade 2 s'effectue en grande partie à l'écrit. Or, une certaine proportion des élèves semble répondre au hasard à l'intégralité du test, étant donné la faiblesse de leurs scores. Ainsi, il apparaît impossible de savoir si l'élève a répondu faux à la question, ou s'il n'a tout simplement pas compris la question, faute de maîtrise suffisante de la langue. Si l'évaluation au début de la scolarité obligatoire semble primordiale pour identifier les élèves en difficulté, il importe de mieux redéfinir les méthodes d'évaluation.

Parfois, les données socioéconomiques sont collectées directement auprès des élèves du primaire ou auprès de leurs parents, voire même par correspondance. Cela renforce un certain scepticisme quant à la confiance à accorder à ce type de données. En particulier, dans une enquête comme TIMSS, on demande à des élèves du grade 4, c'est-à-dire âgés d'environ 9 ans, de répondre à des questions relatives au niveau socioéconomique de leurs parents[39]. La procédure est identique dans PASEC, SACMEQ et LLECE, où l'on tente d'approximer le niveau socioéconomique des parents à partir de questions telles que l'accès à l'eau potable, ou encore la possession (par le ménage) d'une voiture ou d'un téléviseur. Le motif du questionnement direct des enfants, et non de leurs parents, est clairement financier : il est moins coûteux de soumettre un

[37] En 2009, l'OCDE a réalisé une évaluation directement sur ordinateur. Les principaux résultats peuvent être trouvés dans OCDE (2011).

[38] Cependant, l'enquête PASEC évalue les élèves du grade 2 tandis que LLECE-SERCE évalue le grade 3.

[39] Toutefois, l'enquête TIMSS 2011 devrait comporter un questionnaire destiné aux parents, comme c'est le cas pour PIRLS. Ceci est sûrement dû au fait que PIRLS et TIMSS seront tous les deux effectués la même année.

second questionnaire à l'élève plutôt que d'en envoyer un spécifiquement aux parents. Le tableau 18 montre la nature des questionnaires soumis lors des évaluations. Tous les tests incluent un questionnaire contextuel pour les élèves et les directeurs d'écoles. Cependant, seules les enquêtes PIRLS, PISA[40] et LLECE proposent de mieux évaluer les caractéristiques des élèves en demandant aux parents davantage d'informations.

L'enquête PISA est la seule à ne pas contenir de questionnaire dédié aux enseignants. Cela renvoie plus à la conséquence de ne pas stratifier selon des classes entières, mais plutôt une classe d'âge. Enfin, seuls les tests de l'IEA imposent aux responsables des systèmes éducatifs de répondre à un questionnaire sur l'état du système éducatif, ce qui entraîne la rédaction d'une véritable encyclopédie des systèmes éducatifs des pays participants aux différentes évaluations (Kennedy *et al.,* 2007 ; Mullis *et al.* 2009). Pour l'enquête EGRA, les questionnaires n'apparaissent pas dans l'ensemble des évaluations. De plus, leur contenu diffère entre les pays, ce qui est surtout dû au changement des bailleurs entre les différents tests effectués.

Tableau 18 *Types de questionnaires soumis dans différentes évaluations*

Évaluation	Questionnaire élève	Questionnaire parents	Questionnaire maître	Questionnaire directeur d'école	Questionnaire ministère
TIMSS	G4* : 17 questions G8* : 33 questions		G4 : 41 questions G8 : 33 questions	G4 : 21 questions G8 : 22 questions	G4 : 44 questions G8 : 44 questions
PIRLS	24 questions	22 questions	42 questions	27 questions	19 questions
PISA	37 questions	15 questions**		29 questions	
SACMEQ	38 questions		38 questions	41 questions	
LLECE	Grade 3 : 20 questions Grade 6: 40 questions	20 questions**	38 questions	47 questions	
PASEC	G2 : 13 questions G5 : 15 questions		94 questions + fiche de suivi du temps scolaire élèves	84 questions + fiche de suivi du temps scolaire enseignants	
EGRA	25 questions		51 questions	50 questions	

** G4 : Grade 4 ; G8 : Grade 8. ** Le questionnaire « Parents » n'a pas été distribué à l'ensemble des pays.*
Source : auteurs, à partir des rapports cités.

[40] Seuls quelques pays ont choisi d'envoyer de tels questionnaires aux parents à partir de 2006.

Des différences sont aussi à souligner en ce qui concerne le type de questions posées. Dans les enquêtes internationales et régionales sur les élèves, la plupart des questions sont à choix multiples. À la différence de ces enquêtes, l'évaluation IALS contient uniquement des questions ouvertes en vue de maintenir l'authenticité du matériel testé et des processus cognitifs évalués (Grisay et Griffin, 2005). En effet, les enquêtes sur les acquis des élèves, telles que PISA, PIRLS ou SACMEQ utilisent généralement peu de questions ouvertes, mais davantage des questions à choix multiples. Les pays participants pouvant être tentés de préparer les élèves à ces questions, les enquêtes régionales et internationales ne divulguent pas l'intégralité des tests au grand public[41]. À la différence des autres évaluations, l'enquête EGRA diffuse librement les tests soumis aux élèves[42].

Le tableau 19 présente les contenus des évaluations. La durée du test varie de 15 minutes pour EGRA à plus de 4 heures pour SACMEQ. Dans cette dernière enquête, le temps de passation des tests n'est pas un critère d'évaluation. Le seuil de 4 heures est fixé arbitrairement par les évaluateurs, car dans la plupart des cas les élèves terminent le test avant ce délai. La part des questions à choix multiples est très majoritaire dans les enquêtes régionales (SACMEQ, LLECE et PASEC), tandis qu'elles ne représentent qu'un tiers du contenu du test PISA. Plus spécifiquement, le test PASEC inclut peu de questions ouvertes, en 5e année, ce qui ne permet pas (comme pour l'enquête LLECE) de différencier les élèves analphabètes de ceux ayant de faibles compétences. À l'inverse, en 2e année, particulièrement pour les mathématiques, les questions ouvertes dominent. Par ailleurs, comme les tests sont le plus souvent réalisés en plusieurs langues, il est important de souligner qu'il n'existe pas de procédure de traduction méthodique dans les enquêtes PASEC[43] et EGRA. Dans le cas du PASEC, on peut même remarquer que les items en lecture n'apparaissent pas comparables entre le français et l'anglais, d'après les auteurs du rapport du Cameroun (CONFEMEN, 2008*b*, p. 39).

[41] Quelques items sont exposés dans les évaluations TIMSS (Mullis *et al,* 2008 ; Martin *et al,* 2008), PISA (OCDE, 2009*d*).

[42] Ces tests peuvent être trouvés à l'adresse : https://www.eddataglobal.org/documents/index.cfm?fuseaction=showdir&ruid=6&statusID=3.

[43] Une refonte prochaine des tests devrait permettre de remédier à ces problèmes de traduction.

Tableau 19 *Nombre d'items dans les évaluations*

Enquête	Durée du test (en minutes)	Items QCM (en %)	Items questions ouvertes (en %)	Nombre items QCM	Nombre items questions ouvertes	Nombre items total	Tests de traduction
TIMSS	90	54	46	G4 : 96 G8 : 117	G4 : 83 G8 : 98	G4 : 179 G8 : 215	Oui
PIRLS	80	51	49	64	62	126	Oui
PISA (1)	120	33	67	38	70	108	Oui
SACMEQ (2)	240 max	Majorité	Faible	Majorité	Faible	R : 83 M : 63	Oui
LLECE (3)	60	91	9	87	9	96	Oui
PASEC (4)	G2 :80 G5 : 100	G2 : 33 G5 : 88	G2 :66 G5 : 22	G2 :18+8 G5 : 29+25	G2 : 18+32 G5 :6+10	G2 : 36+40 G5 : 35+35	Non
EGRA	15	-	-	-	-	-	Non

Notes :
(1) Le nombre d'items est relatif au domaine des sciences du test passé en 2006. Celui-ci varie selon les vagues et le domaine testé, mais les proportions sont identiques pour le type de questions proposées.
(2) Il n'a pas été possible de déterminer exactement le nombre d'items pour chaque type.
(3) Les chiffres correspondent aux items en mathématiques et au grade 6. Cependant les proportions sont très proches entre les domaines et les grades.
(4) Le nombre d'items présenté en premier concerne la lecture, tandis que le second concerne les mathématiques. Seuls les tests en fin d'année sont pris en compte dans le tableau, version des tests de 2003 à 2010.

Source : les auteurs.

2.2.2. *Analyse du contenu des items*

La plupart des enquêtes tendent à standardiser les tests en recourant à l'évaluation des élèves en mathématiques et en lecture. Pour autant, plusieurs divergences apparaissent entre les différents tests et l'on constate également que, si le même domaine est testé dans plusieurs enquêtes, l'approche relative à ce domaine peut varier profondément.

Alors que la plupart des enquêtes tendent à évaluer les élèves en mathématiques, ce n'est pas le cas pour la lecture ou encore les sciences. La plupart des enquêtes recensées – à savoir LLECE, TIMSS, PISA, PIRLS, PASEC, SACMEQ – ont recours à un test des élèves en mathématiques. Cette matière est perçue comme étant le domaine le plus standardisé, facilitant l'optique de comparaison internationale. Or, même si les mathématiques sont un outil assez homogène au sein des pays, la dimension qui est évaluée peut varier selon les enquêtes. De plus, le domaine des mathématiques n'est pas évalué de la même manière entre les enquêtes. Ainsi, dans le cas de PISA, les connaissances

acquises sont explorées à partir de trois entrées : 1) leur contenu (espace, changement et relations, quantité et certitude), 2) les compétences acquises qu'elles impliquent (domaines de la reproduction, des associations d'objet et de la réflexion), 3) les situations liées au contexte de l'apprentissage (personnelles, d'éducation dans la classe ou de notions scientifiques liées au programme scolaire). Les items du test tendent à se rapprocher davantage de la vie réelle que des situations typiques de la scolarisation. À la différence de PISA, l'enquête TIMSS évalue les mathématiques sur la base de trois dimensions : le contenu (nombres, mesure, géométrie, proportionnalité, fonctions, relations et équations, données, probabilités, statistiques, analyse élémentaire, validation et structure), la performance en termes de prévision (connaissance et utilisation des procédures routinières, analyse et résolution de problèmes, raisonnement et communication mathématiques) et les contextes scolaires (comme les attitudes du groupe pédagogique, l'intérêt pour soutenir l'attention scolaire et les schémas de pensée pédagogiques)

Depuis sa conception, TIMSS a modifié son plan de travail afin de mieux prendre en compte les changements survenus dans les programmes scolaires et pédagogiques des pays participants. L'évaluation se base ainsi essentiellement sur les programmes scolaires plus que sur les compétences nécessaires dans la « vraie vie ».

Dans le cas de SACMEQ II, un certain nombre d'items ont été tirés de l'enquête TIMSS 1995. Plus généralement, l'enquête SACMEQ II inclut des items sélectionnés de quatre enquêtes précédentes : l'étude sur les indicateurs de qualité de l'éducation du Zimbabwe[44], SACMEQ I, TIMSS et l'enquête sur la littératie de l'IEA[45]. La littératie en mathématiques a été définie comme « *la capacité à comprendre et appliquer les procédures mathématiques et à effectuer les interprétations conjointes en tant qu'un individu mais aussi en tant qu'un citoyen d'une société entière* »[46]. (Shabalala, 2005, p.76). Le test comprend trois domaines et concerne le grade 6 : les nombres (opérations et nombres, racines carrées, arrondis, figures graphiques classiques, fractions, pourcentages et ratios), la mesure (en relation avec la distance, la longueur, la zone, la capacité, la monnaie et le temps) et les données spatiales (figures géométriques, graphiques et tables de données). Il est possible de voir des similitudes avec le test soumis par l'IEA à travers TIMSS au grade 4.

[44] *The Zimbabwe Indicators of the Quality of Education Study.*
[45] *The International Association for the Evaluation of Educational Achievement (IEA) Study of Reading Literacy.*
[46] « *the capacity to understand and apply mathematical procedures and make related judgments as an individual and as a member of the wider society* ».

L'enquête PASEC vise à évaluer, en début et en fin d'année, les élèves des grades 2 et 5. Le test de mathématiques au grade 5 inclut, par exemple, des items qui évaluent les connaissances dans les propriétés des nombres et leur capacité à effectuer des calculs simples (addition et soustraction). Les tests incluent également des items qui demandent aux élèves d'utiliser l'addition, la soustraction, la multiplication et la division dans la résolution de problèmes. D'autres items évaluent les connaissances des élèves dans les décimales, les fractions et les concepts géométriques de base. De façon générale, les questions présentes dans les tests PASEC sont d'ordre scolaire et basées sur des connaissances. Il n'est que très peu fait allusion à des tests de compétence ou d'interprétation de graphiques, de tableaux, comme dans les tests de l'IEA ou de l'OCDE (des items du test du grade 5 en mathématiques sont liés à des problèmes de la vie courante). Toutefois, la prochaine réforme des tests du PASEC devrait intégrer davantage d'items relatifs à ces dimensions. Pour le grade 5, en mathématiques, elle devrait être structurée autour de trois domaines principaux (numération et opérations, mesures, géométrie) et de trois processus (connaître et comprendre, appliquer et résoudre les problèmes). Dans les enquêtes PASEC, les compétences du maître en français sont testées en lui proposant de corriger une dictée type, un score étant établi entre les fautes découvertes et celle inventées par le maître. Le SACMEQ soumet au maître quelques items de mathématiques qui recouvrent son enseignement au grade 6. Enfin, l'enquête LLECE II (ou encore appelée SERCE) évalue le niveau des élèves en mathématiques en découpant ce domaine en cinq domaines (UNESCO-OREALC, 2008, p.14) : numérique, géométrique, domaine de la mesure, des compétences basées sur les informations, et de la variation. On constate ainsi des similitudes avec l'enquête TIMSS. Par ailleurs, une partie des items des tests LLECE est inspirée de ceux de TIMSS au grade 4 [47]. Les items sont basés sur la concordance entre les programmes scolaires et les acquis des élèves. Il est ainsi question de mesurer le niveau des acquis scolaires, plus que les compétences des élèves qui leurs sont utiles dans la vie active (comme le fait PISA).

Au-delà des mathématiques, d'autres contenus scolaires sont également évalués. Le deuxième contenu reste la lecture, ou plus spécifiquement la littératie. Les études sont, dans ce domaine, assez hétérogènes. Ainsi, les enquêtes PIRLS, LLECE, SACMEQ, PASEC et MLA sont essentiellement basées sur les programmes scolaires des pays concernés ; elles évaluent l'adéquation entre ce qui doit être enseigné et ce qui est appris par l'élève. PIRLS, LLECE et SACMEQ tendent à proposer un ensemble commun d'items aux pays participants, contrairement à PASEC et MLA, l'objectif principal de ces

[47] À la différence de TIMSS, l'enquête LLECE 2 – SERCE évalue les élèves des grades 3 et 6. En effet, dans TIMSS, seuls les élèves du grade 4 au niveau primaire sont évalués (les élèves du grade 8 sont au niveau secondaire).

dernières étant, non pas la comparaison internationale, mais l'évaluation diagnostique de la performance des élèves au sein des pays. Cependant, quelques améliorations sont à souligner. En effet, en 2007 et 2008, le PASEC a fait réaliser, sur financement de l'IEA, une analyse des réponses aux items du PASEC (en ayant recours au modèle de Rasch) ainsi qu'une analyse des curricula officiels des pays d'Afrique francophone, de l'Océan Indien et du Liban. L'objectif est une révision des tests PASEC à partir de 2011, pour tenir compte des changements opérés dans les curricula depuis les années 1990 dans la majorité des pays[48]. Par ailleurs, les récentes enquêtes EGRA et EGMA tendent à privilégier non pas le niveau de compétence de l'élève mais plutôt son niveau de connaissances minimum en lecture et mathématiques, ce qui souligne une nouvelle volonté de standardiser des savoirs basiques dans le contexte de l'objectif EPT.

2.2.3. *Disponibilité d'indicateurs de marginalisation*

La marginalisation est une dimension importante à prendre en compte dans le contexte de l'Initiative Fast Track. Effectivement, à mesure que les systèmes éducatifs tendent à augmenter les taux de scolarisation, peuvent apparaître des groupes de populations marginalisées au sein même du système éducatif. La généralisation de la scolarisation primaire peut en effet parfois conduire à scolariser des élèves sans leur transmettre réellement les connaissances prévues dans les curricula. La marginalisation a été le thème central du rapport *EFA GMR 2010.* L'équipe chargée de sa rédaction avait alors pris en compte cinq dimensions principales de la marginalisation de l'élève : l'âge, l'origine ethnique, le sexe, la langue maternelle, la zone d'habitation. La possibilité de suivre la performance des élèves dits « marginalisés » apparaît donc fondamentale dans le cadre de Fast Track. Par ailleurs, il serait également utile d'évaluer les niveaux de compétences des élèves ayant quitté l'école ou n'ayant jamais été scolarisés.

Les variables de marginalisation diffèrent selon les enquêtes analysées et présentées dans le tableau 20. Dans l'enquête TIMSS, plusieurs variables permettent d'apprécier le degré de marginalisation : outre le genre de l'élève, on peut détecter si l'élève parle souvent la langue du test à la maison. Les variables évaluant le niveau socio-économique de l'élève sont principalement le nombre de livres à la maison et le niveau d'éducation des parents. Il est également possible de savoir si l'élève est étranger et si ses parents sont nés à l'étranger[49]. Le type de zone géographique

[48] Voir la première partie de l'ouvrage sur le changement de méthodologie du PASEC et CONFEMEN, 2008*b* (pp.34-51) pour une présentation plus détaillée de la méthodologie PASEC.

[49] Ces informations sont tirées directement du questionnaire soumis aux élèves. Voir section 2.2.1 pour plus d'informations.

dans lequel il est scolarisé est également indiqué[50]. Pour certains pays seulement, il est possible de connaître la zone géographique de l'élève. Le type de zone change avec les pays (dans le cas du Botswana, par exemple, il est possible de découper le pays en six grandes régions). L'enquête PIRLS fournit en grande partie les mêmes informations sur la marginalisation que celles fournies par TIMSS[51].

L'enquête PISA fournit des informations relatives au genre de l'élève, à sa nationalité et à celle de ses parents. À la différence des enquêtes TIMSS et PIRLS, il est possible de connaître précisément l'emploi occupé par chacun des parents[52]. La langue parlée à la maison peut être un indice important pour distinguer les groupes ethniques et les populations immigrées. Par ailleurs, le type d'école (publique ou privée) est explicitement demandé au responsable de l'école. Enfin, il est possible de connaître le type de milieu de l'école (rural ou urbain). Un questionnaire spécifique ayant été distribué aux parents des élèves dans certains pays seulement, il devient possible d'établir une évaluation approchée du revenu le revenu du ménage et de confirmer les réponses données par les élèves sur le niveau d'éducation de leurs parents.

Dans LLECE, plusieurs variables de marginalisation sont disponibles, dont le genre, la localité et le niveau d'éducation des parents. L'avantage principal de cette enquête est l'existence d'un questionnaire pour les parents, qui renforce la précision des données car les élèves testés sont aux grades 3 ou 6. Une faiblesse de cette enquête réside dans l'absence de mesure précise du type de localité. En effet, la variable en question est construite à partir du niveau socioéconomique des familles y vivant. Ceci s'avère assez difficile à mettre en œuvre pour les analystes qui aimeraient comparer les grandes agglomérations avec les villages, indépendamment du niveau socioéconomique des élèves. Le questionnaire parents de l'enquête LLCE permet aussi d'apprécier le travail des enfants et les tâches ménagères qui peuvent interférer sur le temps scolaire.

Dans l'enquête SACMEQ, 14 items sont utilisés afin d'évaluer le niveau socioéconomique de l'élève[53]. Le nombre de biens possédés a été résumé pour chaque élève.

[50] Cette question provient du questionnaire donné au chef d'établissement.

[51] À la différence de TIMSS, l'enquête PIRLS distribue un questionnaire aux parents.

[52] Une question ouverte est posée concernant l'emploi occupé par les parents. Contrairement à PISA, dans TIMSS et PIRLS, une classification simplifiée des emplois est proposée à l'élève et aux parents (pour PIRLS seulement).

[53] Ces items sont : journaux d'actualité, magazines hebdomadaires ou mensuels, radio, téléviseur, magnétoscope, lecteur cassette, téléphone, réfrigérateur, voiture, moto, vélo, eau potable, électricité et une table pour travailler. Le travail de Dolata (2005) a consisté à regrouper ces variables pour obtenir un indice socioéconomique composite.

Le score minimum est de zéro et le score maximum est de 14. Outre cette variable, d'autres informations telles que le genre de l'élève, l'éducation des parents, ou encore la fréquence de la pratique de l'anglais à la maison sont recensées. La possibilité de comparer les régions entre elles étant un objectif de l'enquête SACMEQ, il est possible de délimiter chacun des pays participants en plusieurs zones et d'effectuer des comparaisons intranationales. Cette possibilité est systématiquement offerte pour l'enquête SACMEQ seulement. Il faut toutefois souligner que la variable relative au type de zone géographique est imprécise car elle n'inclut pas le nombre d'habitants et peut donc être interprétée de façon subjective.

L'enquête PASEC ne contenant pas de questionnaire dédié aux parents, les questions posées aux élèves sont assez basiques [54]. Les variables pouvant mesurer la marginalisation sont assez nombreuses : outre le genre de l'élève, certaines évaluent le niveau socioéconomique de la famille (type de maison, accès à l'eau, toilettes, détention d'un téléviseur, etc.) sans toutefois permettre une mesure standardisée entre les différentes enquêtes. Le niveau de vie des parents a été appréhendé par un indicateur de confort matériel, utilisant des informations sur la possession à domicile de certains biens durables comme le réfrigérateur, la télévision, la radio, le téléphone, la voiture, le vélo, etc. Cet indicateur prend également en compte l'usage de certaines infrastructures de base à domicile (robinet, toilettes, électricité, etc.) et le type de matériaux de construction de l'habitat. Cependant, cet indicateur n'a pas été calculé pour l'ensemble des pays. La langue parlée à la maison est également demandée, ce qui permet de différencier le groupe ethnique. Néanmoins, il n'est pas demandé explicitement à quel groupe ethnique appartient l'élève. Le milieu où se situe la résidence consiste simplement à distinguer les zones (rurale ou urbaine) ou le type de localité (ville ou village), ce qui est très limité si l'on désire analyser les différences géographiques. Le découpage selon le nombre d'habitants, comme le fait PIRLS, pourrait permettre une meilleure évaluation de la dimension géographique des écoles évaluées. Dans l'enquête PASEC, en complément, certaines questions des questionnaires administrés au maître et au directeur permettent d'apprécier le milieu socioéconomique et la pratique des langues locales. Le questionnaire administré aux élèves tente, *via* quelques questions, d'évaluer le temps consacré aux activités qui interfèrent sur le temps scolaire.

[54] Pour rappel, l'enquête PASEC teste les élèves du grade 2 et du grade 5. Les enquêtes pour les différents pays testés varient sensiblement selon les années. Les informations contenues dans ce paragraphe proviennent du rapport de l'enquête PASEC, réalisée à Madagascar en 2005 (CONFEMEN, 2007*a*).

L'enquête EGRA prend en compte le genre de l'élève, un indicateur socioéconomique ainsi que la langue parlée à la maison. L' indicateur socioéconomique est construit à partir de variables telles que la jouissance à la maison d'une radio, d'un téléphone fixe, de l'électricité, de toilettes, etc. Néanmoins, il n'y a pas, à travers les différentes évaluations, un indicateur composite pouvant mesurer le niveau socioéconomique des élèves de façon homogène. L'enquête EGRA est l'évaluation la plus hétérogène et la moins standardisée parmi les évaluations analysées dans ce document. Par ailleurs, elle ne propose pas, à la différence de PIRLS, de questionnaires dédiés aux parents.

Tableau 20 *Variables de marginalisation incluses dans les tests*

Enquête	Genre	Niveau éducation des parents	Revenu des parents	Métier des parents	Indice socio-économique	Statut d'immigré	Type de localité
TIMSS	Oui	Oui		Oui		Oui	Oui
PIRLS	Oui	Oui	Oui	Oui		Oui	Oui
PISA	Oui	Oui	Oui	Oui	Oui	Oui	Oui
SACMEQ	Oui	Oui			Oui		Oui (1)
LLECE	Oui	Oui			Oui		Oui (2)
PASEC	Oui	(5)		Oui (4)			Oui (1)
EGRA	Oui						Oui (3)

(1) Mesure subjective sans annoncer la taille de la population.
(2) La mesure de la localité englobe également le niveau socioéconomique des familles, ce qui limite fortement l'utilisation d'un tel indicateur.
(3) Il n'y a qu'une distinction urbain/rural.
(4) Limité puisque complété par l'élève sur de grandes catégories : agriculteur, commerçant, fonctionnaires...
(5) Simplement alphabètes ou non.

Source : les auteurs.

2.3. Analyse de la population cible

2.3.1. *Critères de sélection de la population cible*

Si les enquêtes tendent à évaluer les compétences et savoirs des élèves aux niveaux primaire et secondaire, différents grades peuvent être évalués dans chacun de ces niveaux. Il importe ainsi de synthétiser les informations relatives au grade et à l'âge des élèves au sein des différents tests existants afin de voir dans quelle mesure il existe – ou non – des divergences. Une minorité d'évaluations tendent à tester deux

grades adjacents pour les mêmes élèves, permettant ainsi une analyse en termes de valeur ajoutée. Dans le cas des enquêtes internationales et régionales, seule l'enquête PASEC dispose d'une telle approche[55].

De plus en plus, les enquêtes ont tendance à évaluer les élèves présents dans le même grade plutôt qu'un ensemble d'élèves du même âge. Si l'on exclut l'enquête PISA, les autres enquêtes régionales et internationales évaluent les élèves dans un grade particulier. L'enquête LLECE I évalue deux grades adjacents (3 et 4) mais à la même période, ce qui avait également été le cas de l'enquête TIMSS 1995. Il est pour autant impossible de suivre les élèves sur une période définie car ce sont des élèves différents qui sont évalués sur ces deux grades. Les élèves ont entre 8 et 9 ans selon les pays. Pour autant, la deuxième enquête LLECE, appelée SERCE, a consisté à évaluer deux grades différents : 3 et 6 (UNESCO-OREALC, 2008)[56].

Les enquêtes TIMSS et PIRLS évaluent les élèves du grade 4 (qui ont généralement 9 ans) ; ce choix est justifié par la supposition que les élèves devraient savoir lire tout en étant scolarisés au niveau primaire (Olson *et al.,* 2008).

L'enquête PASEC évalue deux grades et pendant deux périodes chacun. Ces grades sont les niveaux CP2 et CM1 (deuxième et cinquième années du primaire). Une approche en termes de valeur ajoutée est utilisée : les élèves sont testés en début et en fin d'année scolaire pour les deux grades. Cependant, il n'est pas possible d'évaluer finement l'évolution de leur performance entre ces deux périodes car il n'existe pas, de manière systématique, d'items d'ancrage des scores. Ce problème était minimisé par les initiateurs de la méthodologie, Jarousse et Mingat (1993), qui privilégiaient la modélisation multivariée. Le score initial de l'élève sert dans ce cas essentiellement de référence pour capter des effets de structure individuels ou de groupe. Le modèle explicatif du score final reprend comme variable explicative ce score initial et les autres variables. Il n'existe pas non plus d'items d'ancrage des scores entre les différents grades[57]. Toutefois, on doit souligner que pour les seules mathématiques de 5e année environ, un tiers des items du pré test est posé de manière identique au post test. Le fait de tester deux grades différents pose donc question, si les évolutions entre

[55] Bien que le PASEC ait une approche en termes de valeur ajoutée, les grades évalués ne sont pas adjacents, et d'importantes limites quant à l'ancrage des items entre le début et la fin de l'année sont à souligner. La refonte des tests en 2012 devrait permettre de corriger une partie de ces problèmes.

[56] Il n'apparaît d'ailleurs jamais dans les rapports les raisons pour lesquelles les grades évalués ont changé, et dans quelle mesure un ancrage des données est possible entre les deux évaluations.

[57] À l'exception toutefois des analyses en cohortes des PASEC II et III où le test de l'année suivante reprenait des items de l'année précédente.

grades ne sont pas possibles. Au contraire, les enquêtes SACMEQ, TIMSS, PIRLS ou encore PISA et LLECE ont recours à des évaluations ponctuelles, souvent espacées de plusieurs années.

SACMEQ évalue le niveau des élèves du grade 6 du primaire, tandis que TIMSS évalue des élèves à trois niveaux (grade 4, grade 8 et dernière année du secondaire). En ce qui concerne PIRLS, l'évaluation se fait uniquement au grade 4. Le SACMEQ a évalué le grade 6 et non le grade 4, comme c'est le cas pour PIRLS ou TIMSS. Les raisons invoquées par le SACMEQ sont : *(i)* des taux de scolarisation assez élevés au grade 6 (pour les 14 pays ayant participé à l'enquête de 2000-2002) ; *(ii)* les comités nationaux ont estimé que tester les élèves à des niveaux inférieurs au grade 6 pouvait poser problème, ces grades inférieurs étant, dans certains pays, une étape de transition entre l'utilisation – par les enseignants – des langues locales et nationales[58] ; *(iii)* la déduction de ces comités nationaux que la collecte de données auprès d'élèves du grade 6 était plus valide que celle opérée auprès d'élèves de grade inférieur[59] (SACMEQ, 2005, p.39-41).

L'enquête PISA est spécifique dans ce domaine, car ce ne sont pas les élèves d'un même grade qui sont évalués, mais des élèves âgés de 15 ans. Cela entraîne parfois des situations hétérogènes entre pays. En France, par exemple, certains élèves de 15 ans sont au niveau du secondaire supérieur (lycée) tandis que d'autres sont au secondaire inférieur (collège). Il faut ainsi choisir une certaine proportion d'élèves dans les deux types d'institutions, alors que ce choix n'est pas d'actualité dans d'autres pays. En reprenant l'expression de Postlethwaite (2004), on peut même parler de pseudo-enseignants dans ce cas précis, car étant donné que ce n'est pas un groupe homogène qui est évalué (comme c'est le cas dans TIMSS), mais plutôt un groupe d'élèves scolarisés dans différents grades, les enseignants sont plus nombreux et très hétérogènes. Pour une trentaine d'élèves, l'enquête PISA doit parfois évaluer plusieurs classes, donc plusieurs types d'enseignants, ce qui n'est pas le cas pour les enquêtes évaluant un niveau scolaire *(ibid.)*.

L'évaluation EGRA ne teste pas systématiquement le même grade. Même si l'objectif affiché consiste à tester les tous premiers grades afin de détecter, de façon précoce,

[58] Cette étape de transition survient généralement au grade 3, mais dans des zones rurales de certains pays, elle peut se faire au grade 4.

[59] Cela est apparu conforme à la réalité, étant peu vraisemblable que les élèves de grade inférieur puissent connaître le niveau d'éducation de leurs parents ni cerner le niveau socioéconomique du ménage (comme, par exemple, le nombre approximatif de livres à la maison).

les élèves en difficulté, il arrive parfois que les grades supérieurs du primaire soient aussi évalués. Ainsi, dans l'enquête du Liberia de 2009, 780 élèves des grades 3 et 5 ont été évalués, tandis qu'aux Philippines, ce sont 1 426 élèves des grades 1 et 3 qui ont été évalués en novembre 2009. Les populations concernées par ces enquêtes diffèrent donc et ne sont pas non plus toujours représentatives.

2.3.2. *Représentativité des échantillons sélectionnés*

Plus fondamentalement, se pose la question de l'hypothèse de représentativité de l'échantillon de la population ainsi que des items utilisés dans les tests. En effet, les procédures d'échantillonnage peuvent varier d'une enquête à l'autre et il serait pertinent de les comparer entre les différentes évaluations. Il est notamment question de voir dans quelle mesure des enquêtes prennent en compte certains facteurs – tels que le lieu d'habitation ou encore le type d'école – dans l'hypothèse de représentativité, et en négligent d'autres. Si l'on analyse la procédure d'échantillonnage, on peut constater qu'elle diffère selon les types d'évaluation. Les enquêtes TIMSS et PIRLS établissent un échantillonnage en deux étapes : dans un premier temps, un échantillon d'au moins 150 écoles est choisi proportionnellement au nombre d'élèves du grade considéré ; ensuite, on distingue les élèves des grades 4 et 8[60], et les élèves du grade ayant le plus d'élèves âgés de 9 à 13 ans sont sélectionnés[61]. Ainsi, les élèves qui participent à TIMSS sont généralement ceux des grades 4 et 8. La stratification de l'échantillon est très hétérogène au sein même des pays participants à l'enquête TIMSS. Deux critères apparaissent assez souvent : le lieu géographique où se situe l'école, et le statut de l'école (publique, privée, confessionnelle, etc.). Cependant, ces critères de stratification ne sont pas toujours énoncés explicitement dans les rapports techniques, ce qui soulève certaines questions importantes dans la possibilité de comparer les régions de pays entre elles[62] (voir section 2.5.2. sur ce sujet).

[60] Sauf pour PIRLS où seul le grade 4 est évalué.

[61] 9 ans pour le grade 4 et 13 ans pour le grade 8.

[62] Il est parfois possible de trouver, dans les rapports techniques, les indicateurs détaillés des critères d'échantillonnage, mais cela varie fortement entre les enquêtes : pour PISA, quasiment aucune information n'est disponible, hormis des généralités sur la technique d'échantillonnage ; pour TIMSS et PIRLS, les données sont inégalement disponibles selon les années et les pays. Cependant, on constate, notamment avec PIRLS 2006 et TIMSS 2007, une tendance à plus de transparence sur les critères d'échantillonnage (géographique essentiellement).

L'enquête LLECE prend en compte deux critères dans la stratification : le type de zone géographique (grande ville, zone urbaine, zone rurale) et le type d'école (publique ou privée). Il n'est pas fait explicitement référence à des zones géographiques, comme c'est le cas dans TIMSS, PIRLS et SACMEQ. Au moins 4 000 élèves sont choisis parmi une centaine d'écoles. 40 élèves sont choisis par école et sont répartis pour moitié dans les deux grades testés (grades 3 et 4).

Dans le cas de PISA, la procédure d'échantillonnage est plus spécifique. Le critère principal de choix des élèves se base sur leur âge, indépendamment de leur niveau scolaire et du type de l'institution. Un échantillon représentatif de tous les élèves de 15 ans est choisi parmi les élèves testés (incluant différents types d'écoles, telles que les institutions professionnelles). Le grade le plus bas est le grade 7 mais plusieurs grades peuvent être concernés pour un même pays. Dans certains échantillons, on peut observer jusqu'à 7 grades différents, ce qui témoigne d'une diversification précoce des systèmes éducatifs ainsi que d'un fort taux de redoublement dans le pays en question. À titre d'exemple, le tableau 21 montre les différents grades d'un échantillon de pays ayant participé à l'enquête PISA 2009. La forte proportion d'élèves de grades inférieurs peut avoir des effets néfastes sur le score moyen des pays (Wu, 2010). En se basant uniquement sur le critère de l'âge, l'OCDE n'apparaît pas mettre en avant cette hétérogénéité possible des systèmes éducatifs quant à la politique du redoublement, ainsi qu'à l'évolution des voies d'enseignement avant l'âge de 15 ans. Ainsi, dans certains pays (comme l'Allemagne), le système éducatif tend à spécialiser les élèves avant l'âge de 15 ans, tandis que dans d'autres (comme la France), cette spécialisation débute plus tard.

Tableau 21 *Répartition des élèves selon les grades, sélection de pays, PISA 2009 (en %)*

	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11	Grade 12	Grade 13
Albanie	0,41	2,00	45,26	52,18	0,15	0,00	0,00
Argentine	4,00	11,76	19,98	59,72	4,53	0,00	0,00
Brésil	6,55	1,80	39,81	32,80	3,04	0,00	0,00
Colombie	4,10	9,86	20,86	44,74	20,44	0,00	0,00
France	1,40	3,00	33,99	57,54	3,98	0,09	0,00
Indonésie	1,38	5,82	45,09	42,78	4,52	0,41	0,00
Japon	0,00	0,00	0,00	100,00	0,00	0,00	0,00
Mexique	1,07	4,55	21,78	71,95	0,65	0,00	0,00
Tunisie	6,86	14,43	26,62	46,96	5,13	0,00	0,00
Turquie	0,48	2,26	25,18	67,91	3,92	0,24	0,00
Uruguay	6,23	11,08	21,82	56,25	4,62	0,00	0,00
États-Unis	0,00	0,08	10,80	69,16	19,87	0,10	0,00
Moyenne (1)	1,15	5,47	34,19	52,52	6,53	0,14	0,00

Note : Les données utilisées n'ont pas été pondérées.
(1) La moyenne est calculée sur l'ensemble des pays participant à PISA 2009.
Source : les auteurs, à partir des données brutes PISA.

Afin de voir dans quelle mesure la présence de plusieurs grades pouvait avoir un effet sur la performance des élèves, nous avons réalisé une analyse de régression de la scolarisation dans chaque grade sur la performance en mathématiques. Nous désirons en effet savoir si le fait d'être scolarisé dans un grade particulier constitue un avantage ou un inconvénient dans la performance en mathématiques. Les résultats sont présentés dans le tableau 22. Le grade non présent est le grade 13. Ainsi, il faut interpréter les coefficients comme des différences de performance vis-à-vis du grade absent. Le modèle 1 consiste en une régression multiple incluant comme variables les grades 7 à 12 avec comme méthode d'estimation les moindres carrés ordinaires. Les erreurs types ont été corrigées par la méthode de White et par la répétition des informations par école *(cluster)*. Il apparaît clairement qu'être dans le grade 7 fait perdre près de 200 points par rapport à un élève du grade 13. L'écart se resserre pour les grades plus élevés. Dans le modèle 2, nous avons voulu savoir si ces différences

étaient également vérifiées dans les pays. Pour ce faire, des variables indicatrices pour les pays ont été incluses dans le modèle. Les écarts sont quelque peu réduits, mais restent significatifs. Il apparaît ainsi que la présence de plusieurs grades pour un même pays, ainsi que son ampleur peuvent avoir des effets sur sa performance moyenne[63].

Tableau 22 *Effet grade sur la performance en mathématiques, PISA 2009, ensemble des pays*

	Modèle 1		Modèle 2	
	Coefficient	Significativité	Coefficient	Significativité
Grade 7	-194,64	(164,18)***	-159,66	(85,05)***
Grade 8	-139,44	(240,64)***	-136,58	(82,61)***
Grade 9	-73,11	(303,04)***	-81,13	(50,73)***
Grade 10	-40,00	(218,98)***	-31,17	(19,668)***
Grade 11	-37,16	(67,09)***	-3,06	(2,16)**
Grade 12	9,59	(2,52)**	32,58	(8,87)***
Constante	525,25	(0,02)	399	(0,02)

Note : Seule la première valeur plausible du score de mathématiques a été utilisée dans notre modèle.
*Les nombres entre parenthèses sont les t de Student[64]. * significatif à 10 %, ** significatif à 5 %, significatif à 1 %.*
Source : les auteurs, à partir des données brutes PISA.

L'enquête SACMEQ évalue les élèves du grade 6. La technique d'échantillonnage utilisée est proche de celle de TIMSS et PIRLS. Celle-ci s'effectue en deux étapes. L'échantillon est composé d'un nombre d'écoles et, à l'intérieur de celles-ci, de 20 élèves du grade 6. L'échantillonnage a exclu les « petites écoles », c'est-à-dire celles qui comptaient moins de 20 ou 15 élèves du grade 6 selon les pays. La justification

[63] Il est évident qu'il faudrait approfondir ce point en le contrôlant par des facteurs socioéconomiques de l'élève, et voir dans quelle mesure les effets grades relevés dans nos estimations ne sont pas dus exclusivement à ces facteurs. Par ailleurs, il est logique d'avoir des différentiels de performance aussi élevés entre les grades, puisque le redoublement témoigne en théorie d'une moindre performance de l'élève. Les élèves inscrits à 15 ans dans les grades les moins élevés devraient logiquement obtenir des scores plus faibles. Ce qui importe ici, c'est la diversité des cas possibles entre les pays participants qui pourrait potentiellement biaiser leur performance.

[64] La statistique t permet de tester l'hypothèse (nulle) selon laquelle la valeur des coefficients de régression n'est pas significativement différente de 0 (en d'autres termes, qu'il existe bien une relation entre la variable dépendante et la variable indépendante en question). La valeur que doit atteindre le test de Student pour que l'on puisse rejeter l'hypothèse nulle dépend du nombre d'observations et du niveau de confiance recherché (de 90 % à 99 % en général).

de cette exclusion est que ces écoles représentaient une petite partie de la population ciblée et qu'elles étaient géographiquement isolées (voir section 2.3.3. sur la marginalisation). Une autre raison concernait l'augmentation des coûts générée par la prise en compte de ces écoles dites « isolées » [65]. Plus précisément, la stratification a d'abord été explicite (basée sur les régions qui composaient le pays testé) puis implicite (basée sur la taille de l'école incluant les élèves du grade 6). Il est donc clair que l'optique de comparaison intranationale à été privilégiée par rapport à celle de la comparaison entre les différents types d'écoles. C'est une différence importante vis-à-vis d'autres enquêtes telles que LLECE, PASEC ou encore PISA.

L'évaluation PASEC s'intéresse à deux populations cibles : les enfants inscrits en deuxième année du primaire et ceux inscrits en cinquième année du primaire. Il faut rappeler que l'objectif principal du PASEC est d'identifier les facteurs qui agissent positivement ou négativement sur les apprentissages des élèves (CONFEMEN, 2008*c*). L'enquête PASEC a le même cadre d'analyse que celui de Lockheed *et al.* (1991) et, surtout, de Jarousse et Mingat (1993). Dans cet ouvrage, la variété des conditions matérielles et organisationnelles de scolarisation est décrite en identifiant les différents acteurs qui interviennent dans le processus d'apprentissage (élèves, familles, maîtres et directeurs) par leurs caractéristiques. Par conséquent, afin de répondre à son objectif de variété des situations scolaires observées, les tests élaborés par le PASEC pour mesurer les apprentissages scolaires visent en premier à discriminer les niveaux des élèves entre eux. L'échantillonnage s'est effectué à deux degrés (Ross, 1995 ; Houngbedji, 2009). La procédure d'échantillonnage est celle d'un sondage stratifié à deux degrés, ou sondage stratifié en grappes. Le principe de cette procédure est de retenir, dans un premier temps, un ensemble d'écoles proportionnellement à leurs effectifs en 2e et 5e années. Ces écoles sont choisies par stratification, de façon à être représentatives de l'ensemble du système éducatif du pays. Cependant, ici, il n'est pas systématiquement fait référence à une région géographique. Cette stratification peut concerner le type d'école ou encore le type de zone géographique (rural, urbain) sans pouvoir différencier de manière précise la zone géographique. Lorsqu'une école est choisie, le PASEC procède ensuite au tirage d'un nombre fixe de 15 élèves par niveau testé [66]. Afin de satisfaire les contraintes méthodologiques du PASEC, basées sur l'hypothèse d'un taux d'hétérogénéité intra classe de 0,3, un minimum de 150

[65] Dans certains pays tels que l'Ouganda, certaines écoles ont été écartées du fait qu'elles se situaient dans des zones de conflit. Dans la plupart des pays, les écoles dites 'spéciales' ont aussi été écartées de l'évaluation, mais peu de critères réels viennent définir cette classification dans les rapports nationaux du SACMEQ II. Voir par exemple le rapport du Kenya (Onsumo *et al,* 2005).

[66] Ce nombre a été porté à 20 dans certaines enquêtes.

écoles est nécessaire. Parfois ce nombre est dépassé ; c'est par exemple le cas de Maurice, testé en 2006 dans le cadre du PASEC : au lieu des 150 écoles requises, 225 écoles furent choisies. La stratification explicite consiste à découper le pays en plusieurs zones : quatre zones sont définies mais non explicitées, celles-ci contiennent uniquement des écoles « classiques » c'est-à-dire non ZEP. Une strate concerne Rodrigues, qui fait partie de l'Ile Maurice, et une sixième strate concerne l'ensemble des écoles ZEP (à l'exception de celles de Rodrigues). Enfin, la dernière strate concerne les écoles privées non-ZEP hors Rodrigues. On constate donc que, dans le cas de Maurice – et du PASEC en général –, la stratification concerne davantage le type d'école que la région géographique proprement dite. Il apparaît dans cet exemple impossible de comparer les différentes régions du pays entre elles, comme c'est le cas dans SACMEQ (voir CONFEMEN, 2008*c* pour plus d'informations sur l'enquête PASEC à Maurice).

L'évaluation EGRA n'affiche pas la volonté de sélectionner systématiquement des échantillons représentatifs dans les différents tests. Ceci renvoie souvent à l'hétérogénéité des financements obtenus pour conduire l'enquête, mais aussi à l'organisme qui effectue le test en question. Si, dans certains tests, les élèves sont représentatifs de la population (exemple du Rwanda, en juin 2009), dans d'autres, seules quelques dizaines d'élèves composent l'échantillon et réduisent tout autant la portée du test (exemple du Nigeria, en novembre 2009, où seuls 50 élèves ont été testés).

2.3.3. *Personnes exclues des tests*

Pour des raisons diverses, les enquêtes excluent certains groupes d'élèves, notamment quatre : *(i)* les élèves habitant dans des zones géographiques difficilement atteignables ; *(ii)* les élèves handicapés non scolarisés dans des établissements généraux ; *(iii)* réfugiés ; *(iv)* les élèves non scolarisés.

Dans TIMSS et PIRLS, le taux d'exclusion n'est pas supposé dépasser 5 % de la population attendue. Les raisons de l'exclusion sont de quatre types (Olson *et al.,* 2008, chapitre 5) : *(i)* les écoles sont géographiquement isolées ; *(ii)* elles accueillent très peu d'élèves ; *(iii)* le curriculum ou la structure des écoles diffèrent du curriculum général ; *(iv)* les écoles sont spécifiques pour répondre à certaines nécessités (en particulier pour les élèves en situation de handicap mental et physique, ou encore les élèves non natifs du pays et ne sachant ni parler ni lire la langue du test).

Pour l'OCDE, dans le cadre de l'enquête PISA, le taux d'exclusion ne doit pas dépasser 5 % de la population évaluée. PISA a exclu plusieurs groupes de populations : les élèves en situation de handicap mental et/ou physique, et les élèves ayant des problèmes de langage/écriture (notamment les élèves nés dans un autre pays, ceux qui ont un

niveau très faible dans la langue du test, et ceux qui ont reçu moins d'une année d'instruction dans la langue du test). D'autres critères ont également conduit à une exclusion, notamment pour des élèves dyslexiques. Cependant, ce type d'exclusion a été étudié au cas par cas et ne pouvait se faire sans l'accord du consortium constitué par des experts de différents pays (OCDE, 2009*b*). Ainsi, selon Wuttke (2008), le Danemark, l'Espagne la Finlande, l'Irlande et la Pologne ont exclu les élèves dyslexiques de l'enquête PISA 2003. Le Danemark a également exclu les élèves atteints de dyscalculie. De façon plus surprenante, le Luxembourg a exclu les nouveaux immigrés.

Dans le cas de l'enquête SACMEQ, les petites écoles ont été exclues. Cependant, la définition d'une petite école change selon les pays : tandis qu'une école est considérée comme petite si elle contient moins de 10 élèves du grade 6 au Lesotho ou aux Seychelles, ce minimum passe à 20 au Botswana ou encore en Tanzanie. Les écoles « spéciales » sont également exclues[67]. En Ouganda, les zones militaires de conflit n'ont pas été prises en compte dans l'enquête SACMEQ II. Le pourcentage de la population exclue varie sensiblement selon les pays : alors qu'elle est inférieure à 0,1 % de la population ciblée en Maurice, elle est égale à 3,9 % au Botswana (SACMEQ, 2005).

L'enquête PASEC n'est pas homogène quant à la procédure d'exclusion des élèves et peu d'informations sont disponibles sur la proportion d'élèves concernés. En règle générale, les écoles exclues sont celles où le français n'est pas la langue principale d'enseignement. Cependant, ce critère est moins valide depuis quelques années, car des pays tels que le Liban ou le Cambodge prennent part aux tests PASEC. Nous avons analysé trois rapports récents du PASEC pour observer les informations qui étaient disponibles sur les échantillons. Les résultats sont présentés dans le tableau 23. Très peu d'informations sont précisées dans les rapports sur les populations exclues, ce qui paraît assez paradoxal étant donné le niveau économique des pays testés. Il apparaît, par exemple, que, dans le rapport Burkina Faso, seules les écoles bilingues et satellites ont été exclues. Mais l'on ignore quelles sont leurs proportions dans la population globale. Toutefois, les rapports indiquent que certaines difficultés ont été rencontrées dans la sélection des écoles. Ceci paraît d'autant plus crédible que les évaluateurs ont souvent beaucoup de difficultés à anticiper la répartition de la population scolaire pour l'année du test. Par ailleurs, le choix du PASEC de toujours partir sur une base de sondage des écoles à partir d'une liste des écoles reconnues par le ministère explique des aléas sur la prise en compte du secteur privé dans les premières enquêtes. L'analyse des strates semble encore incomplète, même si les rapports annoncent avoir sélectionné les régions dans leur plan de sondage. Aucun

[67] Cependant, la nature de ces écoles « spéciales » n'est pas explicitement définie pour chaque pays.

rapport n'effectue d'analyse sur les différences de performance entre grandes zones régionales des pays, au contraire de l'enquête SACMEQ. Enfin, la représentativité des échantillons est souvent remise en cause du fait de choix dans les poids donnés à certaines parties de la population, mais aussi aux pertes importantes d'élèves en cours d'année. Ceci ne doit toutefois pas être perçu comme un point négatif, dans la mesure où les autres évaluations ignorent très souvent la question des élèves absents aux tests.

Tableau 23 *Analyse comparative de la procédure d'échantillonnage dans l'enquête PASEC*

	Bénin	Cameroun	Maurice	Burkina Faso
Année de test	**2004/2005**	**2004/2005**	**2006**	**2006/2007**
Informations sur les populations exclues	Non	Non	Non	Oui (écoles bilingues, satellites)
Problèmes rencontrés	Oui (erreurs dans la définition des types d'écoles l'année de test)	Oui (difficultés à tester la validité des questionnaires en anglais)	Non	Non
Analyse des strates précise	Oui (les répartitions d'effectifs sont représentées)	Non	Oui (stratification par type de zone d'éducation)	Oui (stratification par région géographique)
Représentativité des échantillons	Partielle (les écoles privées ainsi que les écoles multigrades sont surreprésentées)	Partielle (les zones anglophones sont sous-représentées)	Partielle (la stratification ne permet pas une représentativité équilibrée)	Oui
Degré de perte d'élèves	Important (16 % en 2e année et 13 % en 5e année)	Faible (< 5 %)	Important (> 11 %)	Important en 2e année (> 8 %), faible en 5e année (~5 %)
Taux de réponse	Indéterminé	Faible (< 95 % en 2e année, < 93 % en 5e année)	Indéterminé	Faible (~90 %)

Source : les auteurs à partir de rapports nationaux.

2.4. Qualité des tests

2.4.1. *Validité des tests*

La validité d'un test souligne le degré de cohérence entre ce qui devrait être mesuré et les stratégies de collecte de données et des instruments de mesure. Bien que la validité des tests soit primordiale, les spécialistes de l'éducation ne se sont intéressés à ce sujet que pour les évaluations internationales.

Sjoberg (2007) a sévèrement critiqué la validité des enquêtes internationales, et en particulier de PISA. En analysant un item dans le test en mathématiques, il montre que la validité de ce dernier peut être remise en question. En effet, alors que PISA est une enquête censée évaluer les compétences dans la vie active du futur travailleur, la question posée a plusieurs incohérences. Cette question concerne l'application d'une formule qui relie la taille du pied et le nombre de pas nécessaires pour parcourir une distance. Non seulement, la meilleure réponse à la question posée revient à appliquer simplement la formule donnée, mais toute analyse critique n'apporte pas de points, elle semble même ignorée dans les réponses possibles. Par ailleurs, tandis que PISA se veut être une enquête basée sur la vie de tous les jours, dans la question analysée par Sjoberg *(ibid.),* la taille du pied est de 80 centimètres ! L'auteur en déduit qu'un élève qui applique bêtement la formule aura tous les points, alors que celui qui aura une vision critique de la question sera sanctionné. Les tests de PISA n'étant pas rendus publics, il apparaît difficile de mesurer le degré de « réalisme » de l'ensemble du test.

D'autres auteurs ont exprimé des limites quant à la validité des enquêtes internationales, en particulier celles de l'IEA. Russell (1981 et 1982) a critiqué l'utilisation d'un score unique des tests appliqués à des pays différents afin de reporter des scores de performance des élèves. L'auteur explique ainsi que les tests de l'IEA ne peuvent pas être valides, car la couverture des curricula est très incomplète et incorrecte pour être effectuée entre les pays. Ainsi, dans leur analyse de la correspondance des items entre l'enquête TIMSS et le système éducatif national de l'Afrique du Sud, Howie et Hugues (2000) ont constaté que les items en sciences ne se retrouvaient que dans 17 % du curriculum national du grade 7, tandis que 50 % étaient proches de l'enseignement du grade 8. Ces différences sont probablement plus fortes pour d'autres pays, comme le Yémen lors de sa participation à l'enquête TIMSS 2007. Cependant, à notre connaissance, aucune analyse précise n'a été effectuée dans ce domaine.

Un avantage des enquêtes régionales concerne la validité de l'enquête. En effet, les pays testés partageant de nombreux items, le test doit être plus cohérent. C'est ce que l'on peut observer avec l'enquête SACMEQ, où la corrélation entre le score des élèves

basés sur les items de chaque curriculum et le score général de SACMEQ II s'étalait entre 0,98 et 1. Cela signifie que la quasi-totalité des items testés sont déjà inclus dans les curricula de l'ensemble des pays testés (SACMEQ, 2005).

Les erreurs de traduction peuvent également être source de problème dans le cas d'études internationales ; ces derniers peuvent être résolus en suivant deux méthodes principales. La méthode la plus populaire est de faire effectuer une traduction par deux traducteurs indépendants (ou plus). Chacun traite les documents originaux (en général en anglais) dans la langue de destination ; puis les versions indépendantes sont comparées entre elles et fusionnées dans une version finale nationale. L'autre méthode consiste à effectuer une seule traduction dans la langue de destination, puis une autre traduction dans la langue originale. Les deux versions (originale et retraduite dans la langue originale) sont ensuite comparées, et les possibles déviations sont corrigées. PISA a introduit une méthode alternative de traduction parallèle en français et en anglais. Ces deux versions sont traduites dans la langue de destination par des équipes de traducteurs du pays concerné. La traduction s'effectue indépendamment entre les deux langues d'origine (français et anglais). Ensuite, la comparaison entre les deux documents traduits permet d'obtenir une version finale des tests et questionnaires. Or, les tests PASEC et EGRA ne semblent pas recourir à ces méthodes de traduction. Dans PASEC, par exemple, les scores des élèves du Cameroun n'apparaissent pas comparables entre le français et l'anglais. La même observation peut être faite à propos de la participation de Maurice au test PASEC (CONFEMEN, 2008*c*).

On peut noter quelques faiblesses en ce qui concerne la validité de l'enquête IALS sur le domaine de la traduction. Blum *et al.* (2001) ont ainsi montré que la traduction du questionnaire IALS en langue française était biaisée : la traduction francophone pour la France apparaissait plus complexe que la traduction francophone pour la Suisse francophone. La France ayant participé deux fois à l'enquête IALS (en 1995 et 1998 ; Carey, 2000), environ 40 % de l'échantillon français (300 personnes) ont été interrogés dans le questionnaire original français, tandis que 60 % (422 personnes) ont été interrogés dans la version suisse francophone. D'après Blum *et al.,* certains problèmes proposés dans le questionnaire français ne figuraient pas dans la version suisse. Ceci confirme les critiques relatives à IALS selon lesquelles la traduction n'a pas été correctement effectuée, justifiant, en grande partie, le retrait de la France de l'enquête (même si la traduction était de la responsabilité de l'équipe technique française).

2.4.2. Fiabilité des tests

La fiabilité (*reliability*) d'un test renvoie à la consistance des données, en référence à la qualité des instruments, des procédures et analyses utilisés pour collecter et interpréter les données. En particulier, une hypothèse principale retenue dans l'ensemble des tests est que les répondants de tous les pays font de leur mieux lorsqu'ils sont testés. Or, d'après Sjoberg (2007), dans le cadre de PISA, dans une grande partie des sociétés modernes, plusieurs élèves ne réalisent pas leur meilleure performance s'ils trouvent que les items sont longs, illisibles, irréalistes et ennuyeux, et en particulier si de mauvais résultats n'ont pas de conséquences négatives pour eux. Les élèves de Taïwan ou de Singapour, par exemple, se sentent plus concernés *(high stakes)* par la réussite, tandis que ceux de Norvège ou du Danemark n'auraient pas grand intérêt à répondre de leur mieux au test PISA *(low stakes),* ce qui remet en doute la crédibilité d'un test international de renommée tel que PISA.

Le même type de problème a été souligné par Blum *et al.* (2001) pour l'enquête IALS. La France, la Grande Bretagne et la Suède, qui y ont participé deux fois, montrent qu'un biais relatif à l'attitude des personnes interrogées peut être détecté entre l'enquête originale et le second test. Lors de ce dernier, chaque personne a répondu à un questionnaire dans lequel un tiers des questions étaient identiques au questionnaire du premier test IALS. La proportion de bonnes réponses pour chaque individu variait d'un test à l'autre, autant pour la France (si l'on compare sa participation en 1994 et 1998) que pour la Grande Bretagne (entre 1996 et 1998). Ces résultats renforcent l'hypothèse selon laquelle il existerait une relation entre le comportement de la personne interrogée lors de l'enquête et son niveau de littératie.

2.5. Utilisation des résultats aux évaluations

2.5.1. *Analyse normative de la performance sous forme de benchmarks*

Au-delà de la simple évaluation des élèves dans des domaines de compétences variés, les tests récents élaborent des critères précis permettant d'évaluer le niveau de chaque élève. De façon générale, il est possible de distinguer deux types de tests : les tests à référence normative[68] et les tests à référence critériée[69]. La plupart des tests sont aujourd'hui des tests à référence critériée.

[68] Un test à référence normative *(norm-referenced)* est un test, une enquête ou une évaluation qui fournit une estimation de la position de l'individu testé par rapport à une population prédéfinie, en relation avec la dimension mesurée. L'objectif d'un test à référence normative peut être, par exemple, de voir si les élèves peuvent prononcer un certain nombre de mots prédéfini à la minute.

[69] Un test à référence critériée *(criterion-reference)* est un test qui permet de déduire, à partir des scores issus des tests, si la personne évaluée a acquis – ou non – les connaissances ou compétences désirées.

L'enquête TIMSS fournit des indicateurs de *benchmarks*. Quatre différents points de *benchmarks* sont disponibles : le *benchmark* avancé *(Advanced International Benchmark)* correspond à un seuil de 625, le *benchmark* élevé *(High International Benchmark)* à 550, le *benchmark* intermédiaire *(Intermediate International Benchmark)* à 475 et le *benchmark* faible *(Low International Benchmark)* à 400. Ainsi, dans le cas du grade 4 et en mathématiques, si un élève obtient un score supérieur à 400, cela signifie qu'il possède les connaissances basiques en mathématiques, telles que la compréhension de l'addition et de la soustraction de nombres entiers. Il connaît les triangles, il peut lire les graphiques et tableaux simples. Les pays qui ont la plus grande proportion d'élèves atteignant le *benchmark* avancé sont des pays asiatiques (près de 41 % des élèves de Singapour atteignent le niveau avancé). Au contraire, dans la plupart des pays en développement, très peu d'élèves atteignent ce niveau. Ainsi, dans le cas du Yémen, au grade 4 et en mathématiques, aucun élève n'a atteint le niveau avancé, moins d'une dizaine d'élèves ont atteint le niveau élevé, seulement 1 % des élèves a atteint le niveau intermédiaire et moins de 6 % des élèves ont atteint le niveau faible.

L'enquête PIRLS fournit la même approche que celle utilisée dans TIMSS et les seuils fixés sont identiques. Tandis que les élèves de Singapour réussissent presque tous à atteindre le *benchmark* faible[70] (400 points), ils ne sont que 26 % à l'atteindre dans le cas du Royaume du Maroc.

L'enquête PISA propose de séparer des élèves en six niveaux différents. Les seuils pour chacun des niveaux dans le cas des sciences sont les suivants : niveau 6 (707,9), niveau 5 (633,3), niveau 4 (558,7), niveau 3 (484,1), niveau 2 (409,5) et niveau 1 (334,9). Par exemple, moins de 2 % des élèves de Finlande ont un niveau en-dessous du niveau 1, tandis qu'ils sont plus de 50 % dans ce cas au Kirghizistan.

L'enquête LLECE-SERCE distingue quatre différents niveaux de performance des élèves : le niveau 4 correspond à un score supérieur ou égal à 621 points ; le niveau 3 à 558 points ; le niveau 2 à 489 points, et le niveau 1 à 391 points[71]. Près de 10 % des élèves, tous pays confondus, ont obtenu un score inférieur au seuil 1, tandis qu'ils sont plus de 11 % à avoir obtenu un score supérieur au seuil 4. De fortes disparités existent entre les pays : plus de la moitié des élèves du grade 3 de Cuba (54,36 %) ont obtenu un score supérieur au niveau 4 en mathématiques, tandis qu'ils sont moins de 2 % dans le cas du Nicaragua (voir UNESCO-OREALC, 2008, pp.23-25).

[70] *Low International Benchmark.*

[71] Les seuils délimités par le LLECE sont en fait plus précis : niveau 4 (621,68 points) ; niveau 3 (558,54 points) ; niveau 2 (489,01 points) et niveau 1 (391,50 points).

Dans l'enquête SACMEQ, les élèves ainsi que leurs enseignants ont été testés et évalués sur une échelle de plusieurs niveaux. En mathématiques, cinq niveaux sont délimités afin d'évaluer la performance des élèves. Ainsi, un élève qui atteint le niveau 1 est capable d'identifier les formes graphiques simples, de reconnaître les unités de mesure, et d'effectuer des opérations simples telles que l'addition ou la soustraction de nombres à deux chiffres (SACMEQ, 2005). Pour la lecture, la performance de l'élève est également classée en 5 différents niveaux. Par ailleurs, un niveau « minimum » et un niveau « désiré » ont été définis par un comité d'experts indépendants avant la collecte des données. Le niveau minimum indique un niveau qui permettrait à l'élève de poursuivre sa scolarisation l'année suivante ; le niveau « désiré » indique que l'élève est capable de faire face aux impératifs de l'année suivante. Seulement 1 % des élèves du grade 6 du Malawi ont atteint le niveau « désiré » en lecture, tandis qu'ils sont 37 % au Mozambique. Sur l'ensemble des élèves, environ 4 sur 10 ont atteint le niveau « minimum » tandis qu'ils sont 1 sur 10 à avoir atteint le niveau « désiré » (Greaney et Kellaghan, 2008, p.132).

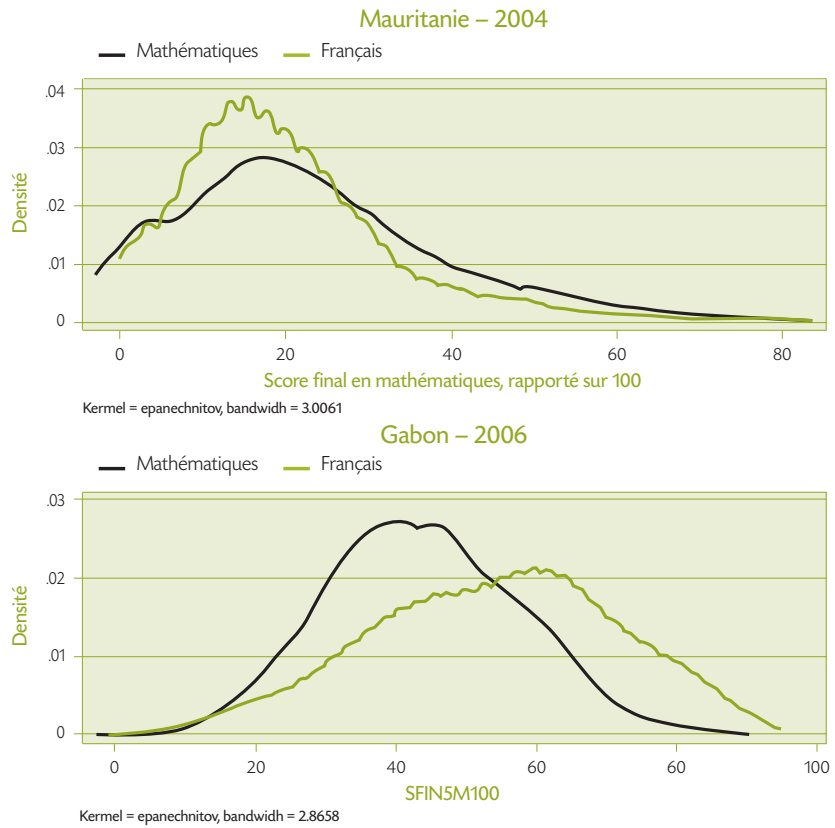
L'enquête PASEC adopte depuis quelques années une approche spécifique de l'échec scolaire, en fixant une norme en-dessous de laquelle l'élève est considéré en échec scolaire. Les tests PASEC présentent plusieurs questions à choix multiples, auxquelles un élève qui ne comprend pas bien les exercices proposés peut être amené à répondre au hasard[72]. L'équipe du PASEC calcule alors la probabilité de donner une réponse juste à chaque item et, par là-même, la note que l'élève peut espérer obtenir en répondant au hasard. C'est cette note qui est retenue comme seuil. Un élève ayant un score inférieur ou égal à ce seuil minimum est considéré comme en difficulté ou en situation d'échec scolaire.

Par conséquent, ce seuil n'est pas déterminé par des spécialistes des curricula ni des psychopédagogues, mais uniquement sur une estimation statistique. Par exemple, on peut voir que, d'après cette approche, près de 70 % des élèves de grade 5 en Mauritanie seraient en échec scolaire en français contre environ 63 % en mathématiques. Au contraire, seuls 7,5 % des élèves du grade 5 au Gabon seraient en échec scolaire en français, ce chiffre passant à 5 % en mathématiques (CONFEMEN, 2008*a*). Ici, il n'est pas du tout fait référence à une norme nationale qui déterminerait le niveau minimum nécessaire de connaissances dans le pays, mais plutôt à une logique statistique basée sur une hypothèse probabiliste.

[72] Cette approche est utilisée également dans les SAT et ETS aux États-Unis.

L'ampleur de ces différences d'un pays à un autre est illustrée dans le graphique 8, qui montre les densités de distribution des résultats aux tests finaux de français et de mathématiques, en 5e année, pour la Mauritanie (test en 2004) et le Gabon (test en 2006). La batterie de tests étant identique dans les deux cas, le graphique montre la dissemblance des deux distributions. Si l'on tient compte du nombre d'items dans les tests (34 en mathématiques et 41 en français) et du nombre de ces derniers, présentés non comme question ouverte mais comme question à choix multiples, on peut (en fonction de ce qui a déjà été précisé) déterminer l'espérance mathématique d'une note obtenue par la seule réponse au hasard dans les QCM, en fonction du nombre de choix. Cette note plancher serait de 14/100 en français et de 15/100 en mathématique.

Graphique 8 *Densités de kernel de la Mauritanie et du Gabon, 5e année, PASEC*



Source : les auteurs.

À côté des *benchmarks,* certaines enquêtes fixent également des normes spécifiques signifiant des niveaux spécifiques des élèves. EGRA, par exemple, définit les seuils limites en référence aux normes fixées aux États-Unis. Ainsi, EGRA peut être considéré comme un test à référence normative. Dans cette évaluation, les enfants lisant moins de 40 lettres par minute à la fin de l'école pré-primaire devraient être considérés en situation de risque d'illettrisme ou en grande difficulté concernant l'apprentissage de la lecture, tandis que ceux qui lisent moins de 27 mots sont considérés comme l'étant définitivement. Or, le résultat de certains pays montre que les élèves à la fin du grade 2 n'ont été capables de détecter correctement que 23 lettres en anglais en une minute[73]. Par conséquent, le niveau moyen de lecture d'un nombre défini de lettres des élèves de fin de grade 2 de ce pays correspond environ à la moitié du niveau considéré comme étant à risque pour un élève à la fin de la scolarisation pré-primaire.

2.5.2. *Comparabilité dans les enquêtes*

Comparaisons internationales

De plus en plus, les ministères de l'Éducation demandent à connaître non seulement le score moyen des élèves mais également les divergences possibles entre les régions du pays, ou encore entre les différents groupes de populations.

Par ailleurs, un score moyen donne une information intéressante seulement si la comparaison internationale est possible. La possibilité de comparaison internationale est renforcée dans le cas des enquêtes régionales et internationales. C'est d'ailleurs le principal avantage de ce type d'enquête. Pour autant, ces comparaisons ne sont pas possibles pour toutes les enquêtes. C'est le cas des enquêtes PASEC et EGRA. Pour PASEC, quelques ajustements ont toutefois été effectués pour permettre une telle comparaison. Le principal inconvénient de cette enquête est lié au fait qu'elle se déroule de façon quasi-autonome pour chaque pays et à une période différente des autres évaluations. Au mieux, seuls deux ou trois pays sont évalués durant la même période, ce qui ne permet pas de procéder à une comparaison internationale[74].

Pour l'enquête EGRA (ou EGMA), il n'est pas non plus possible de recourir à une comparaison entre les performances des pays, mais les raisons de cette absence ne sont pas clairement explicitées (RTI International, 2009). En effet, dans le guide de l'enquête EGRA, il est expliqué que l'acquisition de l'anglais s'avère être plus

[73] Le nom exact du pays considéré dans cet exemple n'est pas explicitement donné dans le document consulté (Gove, 2009).

[74] Un récent travail de l'équipe du PASEC permet toutefois une comparaison des scores de 11 pays (CONFEMEN, 2011).

complexe et ainsi plus long que d'autres langues telles que le grec, le finlandais ou encore le français. Pour cette raison, aucune analyse comparative n'est menée explicitement entre les performances des pays ayant des langues testées différentes[75]. En ce qui concerne la comparaison entre pays testés dans la même langue, EGRA semble pouvoir être comparé, mais ceci reste encore incertain à ce jour, du fait de différences de pratique de la langue (vocabulaire et dialectes pouvant différer). À titre d'exemple, la même version d'instruments ayant été utilisée pour le Pérou et le Nicaragua, il serait possible de comparer la performance de ces deux pays.

Au sein des autres enquêtes (SACMEQ, TIMSS, PIRLS, PISA, LLECE), la comparaison internationale est considérée comme un objectif principal. Cependant, une différence existe entre les enquêtes internationales et régionales : tandis que les premières tendent à afficher clairement des classements de pays dès les premières pages de leurs rapports officiels[76], les secondes privilégient une optique moins directe et se focalisent davantage sur d'autres facteurs.

Un autre point essentiel, mais qui pose des problèmes importants, concerne les comparaisons internationales basées sur des tests validant des apprentissages réalisés lorsque la langue d'enseignement n'est pas la langue parlée dans le milieu local. Prenons un exemple à travers les tests PASEC, dans lesquels diverses questions permettent d'évaluer la pratique du français dans le milieu scolaire et familial[77]. Les traitements réalisés autour de ces variables posent de nombreuses questions, liées au manque de précision dans les réponses recueillies[78]. Bien sûr, dans l'analyse internationale des résultats, les différences sont importantes. Ainsi, sur un protocole d'enquête totalement homogène, il est constaté qu'au Gabon, le français est utilisé comme langue d'expression dans près de 40 % des familles des élèves, alors qu'en Mauritanie ce chiffre est inférieur à 2 %. On ne peut que regretter un manque de précision quant au recueil de ces variables ; par ailleurs, et surtout, il est évident que ces questions de pratique des langues recoupent des variables socioéconomiques sur l'environnement de l'école.

[75] La présentation d'EGRA propose toutefois une analyse comparative mais sans préciser le nom des pays concernés. Voir RTI International (2009), p.2.

[76] Par exemple, dans le cas de PIRLS 2006, les résultats sont présentés dès la page 37. Voir Mullis *et al.* (2007).

[77] Le questionnaire élève comprend une question sur la pratique des langues dans la vie familiale. Il en est de même avec le questionnaire complété par l'enseignant, qui précise sa pratique de la langue d'enseignement dans la vie courante et l'éventuelle utilisation accessoire de la langue locale comme support à son enseignant. Le directeur de l'école doit, quant à lui, préciser le niveau de pratique de la langue d'enseignement dans la population desservie par l'école.

[78] Il est, par exemple, aisé de montrer des cohérences entre les pourcentages de pratique du français donnés par la moyenne des élèves et celle fournie par le directeur.

Comparaisons intranationales

Outre une comparaison internationale, on peut également s'intéresser aux différences intranationales. Généralement, les pays ou territoires sont découpés en différentes zones administratives, comprennent plusieurs groupes ethniques, ou encore des populations parlant des langues différentes ou de différentes cultures. Ces différences existent dans la plupart des pays et peuvent sembler plus sensibles dans les pays en développement. Or, les enquêtes internationales telles que TIMSS, PIRLS ou PISA ne permettent pas, de façon systématique, de telles comparaisons. Les comparaisons intranationales ne font pas non plus partie de leurs objectifs principaux. Pour certains pays, toutefois, ce type de comparaison est possible. On peut citer les cas de l'Algérie, du Botswana ou encore de la Fédération de Russie pour TIMSS 2007. En ce qui concerne PISA, cette optique est encore moins présente et seuls deux pays offrent une telle possibilité dans PISA 2006 (Italie et Espagne)[79]. Citons également, pour TIMSS, PIRLS et PISA, le cas de pays à organisation fédérale (Belgique, Canada, Suisse, etc.) qui ont adapté leur échantillonnage afin qu'il soit représentatif de chaque communauté[80].

La comparaison entre différentes régions au sein d'un pays est possible dans l'enquête SACMEQ et cette possibilité est même inscrite dans les objectifs premiers de l'évaluation. Pour autant, l'enquête PASEC ne permet pas toujours une telle analyse, ce qui limite sa portée pour les comparaisons intranationales. Or, cette possibilité peut être primordiale pour les pays africains, dans la mesure où des groupes ethniques sont parfois présents dans plusieurs pays limitrophes.

À côté des comparaisons géographiques, la totalité des enquêtes permet une possibilité de comparer selon le type de zone dans laquelle est située l'école (rurale ou urbaine). De façon non systématique, les enquêtes TIMSS, PIRLS et PISA permettent aussi de connaître le type d'école (publique, privée ou autre fonctionnement). Les enquêtes régionales facilitent ce type de distinction, qui entre toujours en considération lors de la procédure d'échantillonnage de la population.

[79] Voir OCDE (2009*b*), p.247-312 pour une présentation détaillée des résultats pour ces deux pays. Cependant, certains pays étendent leur participation à l'enquête PISA et ont la possibilité de procéder à de telles comparaisons, mais ceci se fait sur leur propre initiative. On peut citer, en guise d'exemples, les États-Unis ou encore la Suisse, qui sont des pays organisés sur le modèle d'une fédération d'États.

[80] Avec le cas extrême de la Belgique, où souvent les systèmes scolaires francophone et néerlandophone sont considérés comme totalement disjoints, l'un pouvant participer à la vague de l'enquête, l'autre non.

Comparaisons temporelles

Le dernier type de comparaison est temporel. En effet, chaque ministère de l'Éducation désire connaître l'évolution du niveau des élèves et/ou des adultes dans un champ de compétences/savoirs donné. Les enquêtes récentes, avec l'utilisation de la méthode d'IRT permettent aujourd'hui une telle comparabilité : c'est le cas de TIMSS, PIRLS, PISA, SACMEQ, et LLECE. Toutefois, même dans l'enquête PISA, il est impossible de comparer la performance des élèves en mathématiques entre 2000 et 2006, car chaque cycle PISA est spécialisé dans un domaine, ce qui nécessite trois cycles au minimum pour obtenir une comparaison entre domaines[81]. Une telle comparaison n'a pas été systématiquement effectuée dans le cadre du PASEC : seuls trois pays ont dégagé quelques analyses de l'évolution des performances, et ce de manière très limitée (voir CONFEMEN, 2009, tableau 15, p.35 ; CONFEMEN, 2007*b*, tableau 3.7, p.46 ; et CONFEMEN, 2008*d*, tableaux 3.8 et 3.12). Ces comparaisons étaient importantes car elles montraient que, dans certains cas, la marche vers l'EPT se réalise avec un maintien de la qualité scolaire (et pas dans d'autres cas), avec des écarts sensibles entre littératie et numéracie.

En ce qui concerne EGRA ou EGMA, étant donné que les tests visent à évaluer un niveau minimum en lecture ou mathématiques, la comparaison temporelle devrait être possible, mais elle n'a pas encore été effectuée. Deux études pilotes nommées EGRA Plus, menées en 2009 au Kenya et au Liberia, consistent à donner des outils de politique éducative en suivant la performance des élèves dans le temps. Il est ainsi possible, pour certains pays seulement, d'espérer obtenir des comparaisons temporelles de la performance des élèves. Dans le cas du Liberia, par exemple, cette enquête a permis de suivre l'évolution de la performance des élèves pendant une période allant de novembre 2008 à juin 2010 (Piper et Corda, 2011). Cependant, d'après le rapport publié par RTI International, de nombreuses limites sont à souligner dans l'évaluation EGRA[82].

Six pays ou territoires ont participé à la fois à SACMEQ I et SACMEQ II (Kenya, Malawi, Maurice, Namibie, Zambie et Zanzibar). Grâce à la méthodologie utilisée dans SACMEQ, il est possible de suivre l'évolution de la performance des élèves du grade 6 en lecture seulement[83]. L'enquête SACMEQ I a eu lieu en 1995/1996 (sauf

[81] Il est ainsi possible de comparer la performance en lecture entre 2000 et 2006, et la performance en mathématiques entre 2003 et 2006. Aucune comparaison n'est possible en sciences entre 2000 et 2006. Voir notamment OCDE (2010) pour plus d'informations.

[82] Il convient, en particulier, de noter le taux de déperdition des élèves entre les différentes évaluations.

[83] Les élèves n'ont pas été testés en mathématiques dans l'enquête SACMEQ I.

pour le Malawi et le Kenya qui furent évalués en 1998) et SACMEQ II, en 2000 (sauf pour Maurice et le Malawi, testés respectivement en 2001 et 2002) . Avec la publication des résultats de l'évaluation SACMEQ III, en 2010, une comparaison de l'évolution de la performance des pays ayant participé à cette enquête est rendue possible sur la période 1995-2007.

Il serait également intéressant de pouvoir comparer la performance de pays ayant participé à des enquêtes différentes. En toute théorie, cela devrait être possible pour certaines, dans la mesure où les enquêtes SACMEQ et LLECE ont repris quelques items similaires à ceux utilisés dans TIMSS et PIRLS. Or, ce type d'ajustement n'est jamais présenté, même si cela est techniquement possible. Le travail de Brown *et al.* (2005) montre que les enquêtes PISA et TIMSS fournissent des résultats assez similaires, ce qui renforce l'idée d'une comparaison entre les enquêtes. Effectivement, bien que les approches d'évaluation soient différentes entre PISA et TIMSS sur différentes dimensions[84], au final, les performances des pays diffèrent peu.

[84] Les plus significatives sont le type de population testée (d'un côté des élèves d'un même grade [TIMSS, PIRLS], de l'autre des élèves d'un même âge [PISA]) et, surtout, la finalité du test qui diffère (d'un côté on souhaite évaluer le niveau de connaissances dans une matière déterminée [TIMSS, PIRLS], de l'autre on souhaite évaluer le niveau de compétence de l'élève dans un domaine déterminé [PISA]).

Conclusion et recommandations

Ces dernières années, beaucoup de ministères de l'Éducation de pays développés et en développement ont trouvé un intérêt à participer aux enquêtes évaluant la qualité de l'éducation. Dans ce document, nous avons présenté plusieurs aspects relatifs aux évaluations des élèves et des adultes. Plusieurs dimensions ont été explorées, telles que la fréquence des enquêtes, les domaines testés, la nature des populations évaluées, les différentes méthodologies d'évaluation utilisées, et les variables de marginalisation présentes.

1. Cibler les attentes pour mieux cerner les choix en matière d'évaluation

Comme le soulignent Grisay et Griffin (2005), les enquêtes nationales sont plus aptes que les enquêtes internationales à donner des informations spécifiques sur les systèmes éducatifs. Ainsi, ces premières donneront davantage d'informations aux autorités de l'éducation pour *(i)* savoir si les aspects d'un nouveau curriculum ont été réellement appliqués dans les écoles, *(ii)* connaître la proportion d'élèves qui satisfont aux standards nationaux ; et *(iii)* connaître les effets locaux négatifs possibles d'une innovation nationale alternative. Les enquêtes nationales peuvent également répondre aux questions telles que « Combien coûte notre système éducatif ? », « Qui paie pour s'éduquer ? », ou encore « Obtiennent-ils un bon rendement de l'éducation ? ». Les enquêtes nationales permettent donc d'appréhender un certain nombre de points que les enquêtes internationales ne peuvent complètement étudier.

À l'inverse, d'autres points peuvent être traités à partir des enquêtes internationales. En particulier, les enquêtes internationales peuvent : *(i)* informer les autorités nationales sur la possibilité d'autres formes d'organisations scolaires qui « font mieux » que leur propre système éducatif (en termes de performance des élèves, ou encore dans le domaine de l'instruction délivrée, de la qualification des enseignants et/ou l'efficacité de l'utilisation des ressources) ; *(ii)* indiquer si l'organisation des systèmes éducatifs des autres pays implique des disparités faibles pour la qualité de l'éducation fournie, et des effets différents du statut socioéconomique des élèves ou d'autres facteurs tels que l'ethnie ou le genre ; et *(iii)* montrer si l'évolution dans le temps d'un indicateur est positive (ou négative) entre plusieurs pays.

Cependant, les enquêtes internationales présentent plusieurs limites (Greeney et Kellaghan, 2008). Avant tout, il est important de souligner que, malgré l'augmentation du nombre de pays participant aux enquêtes internationales, les domaines de compétence évalués sont souvent les mêmes, ou du moins très similaires. Les enquêtes standardisées ne mesurent qu'une petite partie de ce qu'apprennent réellement les élèves à l'école, n'évaluant généralement que les mathématiques, les sciences et la lecture, et excluant toutes les autres matières qui comportent plus de spécificités nationales dans les programmes.

Par ailleurs, une procédure d'évaluation qui mesure de façon standardisée des compétences dans plusieurs dizaines de pays est pour le moins très difficile à mettre en œuvre. Bien que des éléments génériques dans les curricula soient proches dans l'ensemble des pays du monde dans les enquêtes de l'IEA, des différences fortes peuvent subsister. Celles-ci peuvent se trouver à la fois dans le contenu enseigné, mais également dans la méthode d'enseignement. Comme nous l'avons vu pour l'Afrique du Sud (Howie, 2000), des différences fortes apparaissent entre le curriculum du pays et le contenu des items. Le degré de standardisation d'une enquête dépend donc du degré de proximité entre les items de l'enquête considérée et le curriculum national : plus les items standardisés d'une enquête internationale sont éloignés du curriculum des pays, plus le risque de mal évaluer un système éducatif est élevé. Sauf, évidemment, à se replacer dans une logique proche de l'option de PISA où ce sont des compétences universelles qu'il est nécessaire d'évaluer, alors la question est inversée et l'évaluation internationale peut conduire à une réflexion sur l'adaptation du curriculum national.

Un autre problème relatif aux enquêtes internationales peut renvoyer à la difficulté de transposition des relations trouvées dans un pays. Bien que les enquêtes soient conçues pour permettre une comparabilité entre les pays, il s'avère en effet délicat d'isoler des facteurs organisationnels et d'émettre des généralités quant aux effets de certains *inputs* sur la réussite des élèves. Comme le soulignent Hanushek et Woessmann (2010), les relations trouvées entre certains *inputs* (tels que la taille des classes) et le niveau de performance des élèves dans certains pays, ne se vérifient pas de façon similaire dans d'autres pays. Le contexte dans lequel s'effectue l'évaluation est donc primordial, ce qui limite d'autant plus les comparaisons internationales et donc l'intérêt de participer à une enquête internationale. Il a été démontré, par exemple, que les facteurs familiaux influencent la performance des élèves dans la plupart des pays. Cependant, l'amplitude de cet effet varie considérablement entre les pays (voir par exemple OCDE et UNESCO Institute for Statistics, 2003 ; Wilkins *et al.,* 2002). Si l'on suit ces analyses, la variété des systèmes éducatifs, et surtout de leurs résultats en termes d'acquisition des élèves, est beaucoup plus forte dans les pays en développe-

ment car ces systèmes évoluent dans des sociétés plus diversifiées et souvent plus inégalitaires (par exemple du fait d'une prévalence de VIH-Sida, de la malnutrition, de carences en matière d'infrastructures scolaires, etc.). Cette diversité paraît plus faible dans le cas des pays développés, où le développement correspond aussi à une certaine convergence sociale entre pays. Si les conditions initiales d'enseignement diffèrent fortement entre les pays, il est évident que les comparaisons internationales ne peuvent être considérées comme strictement valides. Ces faiblesses renforcent les avantages liés à une participation aux enquêtes régionales telles que SACMEQ en Afrique subsaharienne ou encore LLECE en Amérique latine.

Il est nécessaire de souligner également que les populations et les échantillons des élèves ayant participé aux enquêtes ne sont pas totalement comparables entre les pays. Par exemple, la comparaison internationale peut être biaisée si la prise en compte de populations spécifiques diffère entre les pays. La prise en compte (ou non) des élèves dans les écoles spécialisées peut parfois modifier fortement le niveau moyen des élèves d'un pays. Le degré de représentativité des échantillons est également fondamental pour le bien-fondé d'une comparaison internationale. Si un pays choisit délibérément, ou non, de ne pas évaluer certaines écoles ou certaines classes, le niveau moyen des élèves évalués de ce pays ne peut pas être considéré comme mesurant réellement le niveau moyen de tous les élèves. Par ailleurs, si des différences fortes existent dans l'abandon scolaire, la date d'administration des questionnaires peut biaiser fortement l'évaluation des acquis des élèves. Or, alors que l'abandon scolaire est proche de zéro dans le niveau primaire pour la plupart des pays développés, il peut atteindre jusqu'à 50 % dans certains pays en développement. De ce fait, on trouvera moins d'hétérogénéité dans les systèmes éducatifs des pays en développement que dans ceux des pays développés.

Une autre difficulté peut également survenir, surtout dans le cas de pays en développement. En effet, si un pays connaît des difficultés administratives et d'organisation de l'évaluation, le temps requis pour effectuer chacune de ces tâches peut s'avérer indisponible dans certains pays : si ces tâches sont aisément effectuées dans la plupart des pays développés, de sérieuses difficultés peuvent survenir dans des pays moins riches. Par exemple, Howie (2000) a montré l'ensemble des difficultés qu'ont connu les administrateurs d'Afrique du Sud pour participer à l'enquête TIMSS (problèmes de suivi des courriers, de téléphone, de fonds monétaires pour les trajets, etc.). Ces difficultés sont aussi à ajouter à d'autres, relatives au manque d'informations sur la population scolarisée, à sa répartition géographique mais aussi au manque de compétences en statistiques chez les fonctionnaires pour élaborer une bonne représentativité de la population nationale (voir tableau 24).

Tableau 24 *Avantages et inconvénients des évaluations sur les élèves*

Enquête	Avantages	Inconvénients
TIMSS	(1) Nombre élevé de pays (2) Analyse précise des curricula (3) Stratification régionale (4) Clarté des données et rapports (5) Publication d'une encyclopédie des systèmes éducatifs	(1) Trop peu de pays en développement (2) Les curricula sont essentiellement basés sur ceux des pays développés (3) La représentativité des régions n'est pas systématique (4) Absence d'un questionnaire parents au grade 4
PIRLS	(1) Seule évaluation sur la lecture au niveau international (2) Disponibilité d'un questionnaire pour les parents (3) Possibilité de suivre l'évolution des pays sur le moyen terme (depuis 2001)	(1) Trop peu de pays en développement (2) Participation irrégulière des pays, empêchant un suivi permanent des performances (3) Trop peu de diffusion auprès des chercheurs et des médias
PISA	(1) Augmentation croissante du nombre de pays (70 pays prévus pour 2012) (2) Introduction de l'informatique dans l'évaluation des élèves (3) Mesure des compétences génériques plus que des contenus scolaires (4) Politique de communication efficace (5) Possibilité d'étendre l'enquête au primaire dans les années à venir	(1) Focalisation sur un groupe d'âge et non un grade particulier (2) Absence de comparaison temporelle pour tous les domaines et entre toutes les vagues [85] (3) Quasi absence de comparaison des régions au sein des pays (4) Absence de questionnaire pour les enseignants (5) Analyse normative trop focalisée sur les élites et les pays de l'OCDE
SACMEQ	(1) Enquête réunissant la plupart des pays anglophones d'Afrique (2) Politique de publication plus cohérente depuis 2010 (3) Données standardisées grâce à une coopération avec l'IIEP (4) Possibilité de comparer l'évolution de la performance des régions au sein des pays	(1) Absence de test dans des langues locales et focalisation sur l'anglais (2) Diffusion des données trop tardive empêchant des analyses précises (3) Faible communication des résultats (4) Question de l'élimination des écoles à faibles effectifs au grade 6(<20)
LLECE	(1) Deux grades évalués permettant des analyses plus poussées (2) Publication des résultats sous forme de rapports de synthèse (3) Questionnaire pour les parents (4) Technique de l'IRT	(1) Publications en espagnol et trop peu en anglais (2) Diffusion des données quasiment absentes pour les chercheurs (3) Grades testés qui varient selon les tests (4) Impossibilité de comparer la performance des élèves entre les deux vagues (5) Politique de communication peu efficace. (6) Impossibilité de comparer la performance des élèves entre les grades

[85] Une tentative récente existe avec PISA-TRENDS (voir OCDE, 2010).

Enquête	Avantages	Inconvénients
PASEC	(1) Seule évaluation à mesurer la performance des élèves sous forme de valeur ajoutée (2) Nombre croissant de pays participants (y compris du Moyen-Orient et de l'Asie) (3) Évaluation de deux grades différents (4) Enquête ancienne ayant débuté au début des années 1990 (5) Enquête réunissant la plupart des pays francophone d'Afrique subsaharienne	(1) Difficultés à comparer la performance des pays entre eux et dans le temps pour les pays ayant pris part à plusieurs vagues de tests (2) Questionnaires maîtres et directeurs d'écoles trop longs et inexploitables (3) Absence de possibilité d'évaluer la valeur ajoutée des élèves de manière précise (pas de procédure d'ancrage) (4) Impossibilité de comparer la performance des élèves entre les grades (5) Participation non synchrone des pays (6) Qualité médiocre des rapports nationaux
EGRA	(1) Seule enquête qui évalue les élèves à l'oral (2) Capacité de distinguer précisément les élèves analphabètes et ceux comprenant la langue testée (3) Évaluation aux tous premiers grades de l'école primaire (4) Participation de pays où aucune information n'est disponible (Nigeria, Bangladesh) (5) Publication de rapports nationaux rapide et détaillée (7) Publication des items servant aux tests	(1) Absence de possibilité de comparer les résultats entre les différents tests (2) Grande variabilité des méthodes d'échantillonnages, de la qualité des tests (3) Absence de diffusion publique des données (4) Analyses trop limitées dans les rapports remettant en cause la crédibilité des tests

Source : les auteurs.

2. Adapter le PASEC au vu de ses faiblesses

L'équipe de la CONFEMEN tentant d'adapter, depuis 2011, le test du PASEC aux techniques modernes d'évaluation, il apparait souhaitable de proposer quelques pistes d'évolution de ce test. Pour ce faire, nous présentons, dans le tableau 25, les principales limites de cette évaluation et les solutions proposées pour y remédier.

Tableau 25 *Préconisations pour l'amélioration du test PASEC*

Problèmes	Solutions proposées
Grades testés : le PASEC évalue à deux grades différents (2 et 5). Cette méthode est un avantage puisque l'on peut mesurer, sur deux grades, la performance des élèves d'une même école. Cependant, il n'y a pour le moment aucune possibilité de mesurer l'évolution des compétences entre ces 2 grades.	• Une solution consiste à introduire dans le pré-test du grade 5 une dizaine d'items provenant du post-test du grade 2. Il serait par conséquent possible d'évaluer l'évolution de la performance standardisée des élèves d'un grade à l'autre, de comparer le degré de valeur ajoutée entre les grades.
Une autre limite concerne l'absence d'items d'ancrage entre le pré-test et le post-test. En effet, jusqu'aux dernières évaluations, il reste impossible de savoir comment a évolué la performance des élèves entre le début de l'année et la fin d'année, seul le test de français de 5e année permet cette comparaison avec environ 1/3 d'items communs, mais centrés sur la compréhension de l'écrit..	• Une première solution – la moins coûteuse – serait d'abandonner cette perspective et de soumettre un test plus léger aux élèves au début de l'année (pré-test), en le composant par exemple d'une vingtaine d'items seulement. Si ce pré-test doit uniquement servir à contrôler le niveau initial de l'élève, il resterait peu utile de soumettre un questionnaire entier. • La deuxième solution – la plus efficace mais pas forcément la moins coûteuse – consisterait à introduire dans le post-test au moins un tiers des items du pré-test. , Par ce biais on saurait exactement comment a évolué la performance de l'élève au cours de l'année. Plus fondamentalement, on pourrait dériver trois types de valeurs ajoutées : celles de l'élève et de son milieu socioéconomique, celle de sa classe et de son enseignant, et une partie qui resterait inexpliquée (aléas du temps, etc.).

Problèmes	Solutions proposées
En ce qui concerne la disponibilité des différents questionnaires contextuels, on remarque plusieurs problèmes qui devraient être corrigés, car ils sont à la fois inefficaces et porteurs de coûts financiers importants. En premier lieu, on remarque que le questionnaire contextuel de l'élève du grade 2 est clairement inapproprié pour un tel âge.	• Une première solution reviendrait à demander aux parents de répondre à certaines questions contextuelles afin de mieux mesurer le niveau socioéconomique de l'élève. Il est en effet assez difficile d'imaginer qu'un élève du grade 2 puisse savoir précisément le type de biens dont sa famille dispose à la maison. Bien évidemment, la question des problèmes postaux et d'acheminement vers les parents est posée dans cette hypothèse, mais il reste toutefois possible de s'inspirer de l'évaluation LLECE, qui parvient à soumettre un tel questionnaire. • Une autre solution serait de demander oralement à chaque élève le type de biens qu'il possède en utilisant des représentations graphiques ou autres, comme le font les évaluateurs d'EGRA. Cette approche est déjà en partie adoptée par le PASEC mais le questionnaire soumis reste trop basé sur du texte, empêchant sans doute l'élève de bien saisir l'enjeu des questions, surtout en 2e année.
Une autre faiblesse importante du PASEC renvoie à la longueur illogique des questionnaires maître et directeur d'école. Si on les étudie de plus près, plusieurs questions sont clairement sans utilité et les réponses sont difficilement vérifiables.	• La solution préconisée serait de s'inspirer des questionnaires effectués dans PIRLS ou dans SACMEQ en y ajoutant au besoin quelques questions propres aux pays francophones. La lecture des questionnaires laisse tout de même assez perplexe tout spécialiste de l'éducation. • Il est aussi très important de diminuer le nombre de questions dans ces questionnaires.
Par ailleurs, on note l'absence d'indices de ressources scolaires standardisés, d'un indice de climat scolaire et d'un indice de degré d'implication financière des familles dans la gestion de l'école.	• En premier lieu, il conviendrait de s'inspirer du questionnaire SACMEQ pour dégager un indice de ressources scolaires. Le récent travail de Saito (2005) dans le cadre du SACMEQ peut servir de base pour l'établissement d'un indice composite.

Problèmes	Solutions proposées
	• En ce qui concerne l'établissement d'un indice de climat scolaire, il serait possible de s'inspirer du travail pionner de PIRLS dans ce domaine. Comme cette évaluation concerne le grade 4, les questionnaires pourraient être utilisés dans le cadre du PASEC avec des ajustements. • Il y a aussi un besoin d'établir une variable d'implication financière des familles dans la gestion de l'école. Il faudrait pour cela insérer des questions appropriées dans le questionnaire du directeur d'école, mais aussi dans celui des enseignants. Par ailleurs, la mise en place d'un questionnaire dédié aux parents pourrait permettre d'améliorer sensiblement la précision des contributions, comme le fait PISA depuis 2006 pour une sélection de pays. • La disponibilité croissante de bases exhaustives sur les moyens affectés aux écoles pourrait permettre des appariements de fichiers ; ceci avait été possible à titre expérimental pour l'enquête Mauritanie.
Une autre faiblesse principale du PASEC renvoie à l'absence d'un indice de niveau de vie standardisé sur l'ensemble des évaluations. Il demeure risqué de penser aboutir à un tel indice en ayant recours à un questionnaire élève au grade 2.	• On pourrait introduire un questionnaire pour les parents ou encore poser oralement les différentes questions à l'élève. • Pour la construction d'un tel indice, il serait possible de s'inspirer du travail de Dolata (2005) dans ce domaine. La méthodologie utilisée par PISA est inappropriée pour le PASEC car elle suppose qu'il existe un marché du travail bien structuré dans le pays. Ainsi, le SACMEQ devrait servir de référence à une telle approche. Une dizaine de questions pourraient être reprises du questionnaire élève du SACMEQ afin de les introduire dans le questionnaire des élèves du grade 5 du PASEC.

Source : les auteurs.

3. Redéfinir l'évaluation en éducation en Afrique

L'évaluation des acquis scolaires en Afrique n'est pas récente. Comme nous l'avons souligné au cours de ce rapport, les évaluations telles que celle du MLA ou encore du SACMEQ ont débuté dès les années 1990. Il n'en reste pas moins que subsiste aujourd'hui la question cruciale de l'utilité de ces évaluations pour les décideurs politiques.

En effet, la mise en application des recommandations – quand elles existent – des rapports issus des différentes évaluations sur les acquis reste posée. Si l'on prend l'exemple des évaluations nationales, il apparaît très souvent qu'elles sont plus désirées par les bailleurs de fonds que les pays eux-mêmes. Ceci est compréhensible dans la mesure où les pays n'ont pas de capacité bureaucratique suffisante pour remplir les conditions nécessaires afin d'évaluer les élèves. L'autre raison principale à ce désintéressement des ministères renvoie également au manque de dispositions financières et aux contraintes budgétaires : compte tenu du contexte de la scolarisation primaire universelle, les gouvernements s'attardent davantage sur l'accès et la survie au cours de l'éducation primaire, que sur les acquis des élèves.

Un certain revirement semble cependant se profiler de la part des bailleurs de fonds, Banque mondiale en tête. Au-delà du simple accès à la scolarisation primaire, la Banque, *via* son projet *Education for All Fast Track Initiative* (rebaptisé, depuis fin 2011, *Global Partnership for Education*), semble désormais se focaliser sur l'analyse de l'évolution de la qualité de l'éducation. L'action croissante des parents et des ONG va aussi dans le sens d'une recherche de qualité par l'évaluation des systèmes éducatifs et le débat que ces enquêtes induisent.

La principale faiblesse que l'on peut constater sur les évaluations est le manque de coordination entre celles-ci. Par exemple, dès lors qu'un pays entreprend d'évaluer le niveau d'acquis de ses élèves, il peut recourir à plusieurs types d'évaluations :

- une évaluation certificative nationale tentant d'évaluer dans quelle mesure les élèves ont atteint (ou pas) un niveau prédéfini. Ce type d'évaluation est courant en Afrique subsaharienne, et en particulier dans les anglophones ; il permet souvent aux élèves ayant validé ce concours d'accéder à un niveau supérieur. Cependant, ce type d'évaluation ne fournit pas d'informations sur le réel niveau des élèves, il ne sert qu'à certifier l'acquisition d'un niveau prédéfini ;
- une évaluation nationale sur les acquis des élèves qui va consister à mesurer le niveau des élèves dans les domaines les plus courants (mathématiques, lecture ou encore

sciences). Ce type d'évaluation est difficile à mettre en place au sein des pays africains, essentiellement du fait d'un manque de compétences dans le domaine de l'évaluation et de difficultés à estimer correctement la population des élèves à tester ;

- une évaluation de type EGRA, où seule une partie de la population est évaluée afin de dresser un diagnostic du niveau initial des élèves. Ce type d'évaluation a tendance à se généraliser en Afrique subsaharienne, étant donné le faible coût de ce test et la volonté, de la part des bailleurs de fonds, de renforcer l'évaluation au sein des pays africains ;
- une évaluation de type PASEC ou SACMEQ, où le pays obtiendrait, en plus d'une simple mesure du niveau d'acquis de ses élèves, une base comparative avec les autres pays de la région. La plupart des pays africains ont eu recours à ces évaluations régionales. Cependant, le degré d'implication de ces évaluations sur les politiques éducatives pose question.

La question principale réside, dans ce dernier cas, à se demander quelle préconisation pourrait être apportée pour la participation d'un pays à l'une ou plusieurs de ces évaluations. Bien qu'il n'en ait pas été question au cours de ce rapport, la question du coût des évaluations est primordiale : il est non seulement financier, mais se mesure également en termes de temps disponible par expert national. Bien souvent, les personnes capables d'organiser une telle évaluation sont peu nombreuses au sein des pays en développement. Dès lors, il convient d'effectuer des choix pertinents en termes de participation à une évaluation. Cette dimension du coût, entraînant une restriction du nombre d'évaluations, ne permet pas le développement d'un potentiel national stable de capacité d'évaluation, tout juste permet-il le développement d'une culture d'évaluation.

Si, une fois ces contraintes prises en compte, il n'est guère possible d'espérer un développement massif du potentiel d'évaluation, au moins pourrait-on envisager des ponts entre les diverses méthodes afin de les enrichir par des analyses croisées. Dans une possibilité d'accroissement des analyses secondes, l'évaluation certificative pourrait être croisée avec des enquêtes internationales. Ainsi, pour un pays disposant d'un examen de fin d'études primaires, l'on pourrait croiser les résultats de celui-ci avec ceux d'une enquête internationale menée dans les dernières années de ce cycle primaire. Si cet essai ne serait possible que pour un nombre très limité de pays, au moins mériterait-il d'être tenté, compte tenu de l'intérêt des informations ainsi produites, et de son coût de réalisation modéré.

Liste des sigles et abréviations

ABE LINK	*Assistance to Basic Education / Linkages in Education aud Health*
ACER	*Australian Council for Educational Research*
ADEA	*Association for the Development of Education in Africa*
AELE	Association européenne de libre-échange
AFD	Agence Française de Développement
AHELO	*Assessment of Higher Education Learning Outcomes*
ALL	*Adult Literacy and Life Skills Survey (ELCA)*
BASDA	*Basic Academic Skills Diagnostic Assessment*
BEC	*Botswana Examination Council*
BGCS	*Botswana General Certificate of Secondary Education*
BREDA	Bureau régional pour l'éducation en Afrique
BRR	*Balances Repeated Replication*
CAE	*Council for Aid to Education*
CBE	*Certification of Basic Education*
CEE	*Common Entrance Examination*
CITE	Classification internationale type de l'éducation
CLA	*Collegiate Learning Assessment*
CONFEMEN	Conférence des ministres de l'Éducation des pays ayant le français en partage
CRT	*Criterion Referenced Testing*
DEP-PAGE	Direction de l'enseignement primaire et projet d'appui à la gestion de l'éducation
DFID	*Department for International Development*
EEOS	*Equality of Educational Opportunity Survey*
EFA	*Education for All*

EFA FTI	*Education for All – Fast Track Initiative*
EGMA	*Early Grade Mathematics Assessment*
EGRA	*Early Grade Reading Assessment*
EHAE	Évaluation hybride sur les acquis des élèves
EIAA	Enquête internationale sur l'alphabétisation des adultes (IALS)
EIAE	Évaluation internationale sur les acquis des élèves
EICA	Évaluation internationale sur les compétences des adultes
ELCA	Enquête sur la littératie et les compétences des adultes (ALL)
EN	*Evaluaciones Nacionales*
ENAE	Évaluation nationale sur les acquis des élèves
ENLACE	*Evaluacion Nacional del Logro Académico en Centros Escolares*
EPT	Éducation pour tous
EQUIS	*European Quality Improvement System*
ERAE	Évaluation régionale sur les acquis des élèves
ESQAC	*Educational Standards and Quality Assurance Center*
ETS	*Educational Testing Service*
EXCALE	*Examen de la Calidad y el Logro Educativos*
FILNA	*Fidji Island Literacy Numeray Assessment*
GCM	Graphèmes lus correctement en une minute
GMR	*Global Monitoring Report* (UNESCO)
HSRC	*Human Sciences Research Council*
IAEP	*International Assessment of Education Progress*
IALS	*International Adult Literacy Survey* (EIAA)
ICFES	*Instituto Colombiano para el Fomento de la Educación Superior*
ICT	*Information and Communications Technology*
IDB	*Inter-American Development Bank*

IEA	*International Association for the Evaluation of Educational Achievement*
IEQ	*Improving Educational Quality*
IIEP	*International Institute for Educational Planning*
INEADE	Institut national d'étude et d'action pour le développement de l'éducation
INEE	*Instituto Nacional para la Evaluación de la Educación*
INEP	*Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira*
IRM	*Item Response Modeling*
IRT	*Item Response Theory*
IS	Institut de statistique (UNESCO – UIS)
JCE	*Junior Certificate Examination*
LAMP	*Literacy Assessment and Monitoring Program*
LLECE	*Latin American Laboratory for Assessment of the Quality of Education*
MCT	*Minimum Competency Testing*
MEPS	Ministère des Enseignements primaire et secondaire
MEPSP	Ministère de l'Enseignement primaire, secondaire et professionnel
MdE	Ministère de l'Éducation
MDG	*Millennium Development Goals*
MIM	Mots inventés lus correctement en une minute
MLA	*Monitoring Learning Achievement*
MST	*Minimum Standard Tests*
MTM	Mots/texte lus correctement en une minute
NAAA	*National Assessment of Academic Ability*
NAEA	*National Assessment of Educational Achievement*
NAEAS	*National Assessment of Educational Achievements of Students*
NAEP	*National Assessment of Educational Progress*
NAGB	*National Assessment Governing Board*

NAPE	*National Assessment of Progress in Education*
NAPEMR	*National Assessment of Primary Education Mathematics and Reading*
NAPLAN	*National Assessment Programme – Literacy and Numeracy*
NAS	*National Assessment System*
NASA	*National Assessment of Students' Achievement*
NAT	*National Achievement Test*
NCERT	*National Council of Educational Research and Training*
NCES	*National Center for Education Statistics*
NCLB	*No Child Left Behind* (Loi)
NCRA	*National Criterion-Referenced Assessment*
NEAT	*National Elementary Achievement Test*
NEMP	*National Education Monitoring Project*
NEU	*National Examinations Unit*
NIER	*National Institute for Educational Policy Research*
NST	*National Standardized Test*
NTIC	Nouvelles technologies de l'information et de la communication
OCDE	Organisation de coopération et de développement économiques
OMS	Organisation mondiale de la santé
ONE	*Operativo Nacional de Evaluación*
ONG	Organisation non gouvernementale
PAES	*Prueba de Aptitudes y Aprendizaje para Estudiantes de Educación Media*
PASEC	Programme d'analyse des systèmes éducatifs des pays de la CONFEMEN
PAST	*Program for Achievement School Test*
PAT	*Progressive Achievement Tests*
PEP	*Primary Education Project*
PIAAC	Programme pour l'évaluation internationale des compétences des adultes
PIB	Produit intérieur brut

PIRLS	*Progress in International Reading Literacy Study*
PISA	*Program for International Student Assessment*
PMA	Pays les moins avancés
PRONERE	*Programa Nacional de Evaluación del Rendimiento Escolar*
PSLE	*Primary School Leaving Examination*
QCEA	*Qatari Comprehensive Educational Assessment*
QCM	Questionnaire à choix multiples
RCA	République centrafricaine
RDC	République démocratique du Congo
RIES	*Research Institute for the Educational Sciences*
RLS	*Reading Literacy Study (The)*
SABE	*Strengthening Achievement in Basic Education*
SACMEQ	*Southern and Eastern Africa Consortium for Monitoring Educational Quality*
SAEB	*National System for the Evaluation of Basic Education*
SAT	*Scholastic Aptitude Test*
SECE	*Sistema de Evaluación de la Calidad de la Educación*
SEDEP	Service de développement et d'évaluation de programmes de formation
SERCE	*Second Regional Comparative and Explanatory Study*
SIMCE	*Sistema de Medición de la Calidad de la Educación*
SIMECAL	*Sistema de Medición de la Calidad*
SINEA	*Sistema de Información, Monitoreo y Evaluación de Aprendizajes*
SINECA	*Sistema Nacional de Evaluación de la Calidad de los Aprendizajes*
SNE	*Sistema Nacional de Evaluacion*
SNEPE	*Sistema Nacional de Evaluación del Proceso Educativo*
SNERS	Système national d'évaluation du rendement scolaire
SPELL	*Samoa Primary Education Literacy Levels*

SSME	*Snapshot of School Management Effectiveness*
STAR	*Supplementary Tests of Achievement in Reading*
TIC	Technologies de l'information et de la communication
TIMSS	*Trends in International Mathematics and Science Study*
UBEC	*Universal Basic Education Commission*
UE	Union européenne
UMCE	*Unidad de Medición de la Calidad de la Educación*
UNESCO	Organisation des Nations unies pour l'éducation, la science et la culture
UNICEF	*United Nations Children's Fund*
UIS	*UNESCO Institute for Statistics* (IS)
USAID	*United States Agency for International Development*
ZEP	Zone d'éducation prioritaire

Bibliographie

Abadzi, H. (2006), *Efficient Learning for the Poor,* Banque mondiale, Washington, D.C.

Banque Mondiale (1995), *Priorities and Strategies for Education,* Washington D.C.

Barro, R.J. (2001), "Education and Economic Growth" in Helliwell, J.F. (Ed.), *The Contribution of Human and Social Capital to Sustained Economic Growth and Well-Being* (pp.14-41), OCDE, Paris.

Barro, R.J. (1991), "Economic Growth in a Cross Section of Countries", *Quarterly Journal of Economics,* 151, 407-443.

Benavot, A. et E. Tanner (2007), "The Growth of National Learning Assessments in the World: 1995-2006", article commandé pour le Rapport de suivi de l'EPT 2008, *'Education For All by 2015: Will We Make It?'.*

Benhabib, J. et M. Spiegel (1994), "The Role of Human Capital in Economic Development: Evidence from Aggregate Cross-Country Data", *Journal of Monetary Economics,* 34, 143-179.

Blum, A., H. Goldstein, et F. Guérin-Pace (2001), "International Adult Literacy Survey (IALS): an Analysis of Adult Literacy", *Assessment in Education,* Vol. 8, No. 2, pp. 225-246.

Boissiere, M., J. B. Knight et R.H. Sabot (1985), "Earnings, Schooling, Ability, and Cognitive Skills", *The American Economic Review,* 75(5), pp.1016-1030.

Bonora, D. et P. Vrignaud (1998), "Literacy Assessments and International Comparisons", *in* Wagner, D. (Ed.), *Literacy Assessment for Out-Of-School Youth and Adults,* UNESCO/ International Literacy Institute, Philadelphia.

Borghans, L., H. Meijers et B. Ter Weel (2008),"The Role of Noncognitive Skills in Explaining Cognitive Test Scores", *Economic Inquiry,* 46(1), pp.2-12.

Bottani, N. et P. Vrignaud, P. (2005), « La France et les évaluations internationales », Rapport établi à la demande du Haut conseil de l'évaluation de l'école, Paris.

Bourdieu, P. et J.-C. Passeron (1964), *Les héritiers – les étudiants et la culture,* Le sens commun, Editions de Minuit, Paris (traduction anglaise, en 1979 : *The Inheritors: French Students and their Relations to Culture,* University of Chicago Press)

Brown, G., J. Micklewright, S.V. Schnepf et R. Waldmann (2005), "Cross-National Surveys of Learning Achievement: How Robust Are the Findings?" Southampton Statistical Sciences Research Institute, *Applications and Policy Working Paper,* A05/05.

Burstein, L. (1992), *The IEA Study of Mathematics III: Student Growth and Classroom Processes,* Pergamon Press, Oxford.

Carey, S. (Ed.), (2000), *Measuring Adult Literacy – the International Adult Literacy Survey in the European Context,* Office for National Statistics, Londres.

Casassus, J., J.E. Froemel, J.C. Palafox et S. Cusato (1998), *First International Comparative Study of Language, Mathematics, and Associated Factors in Third and Fourth Grades, First Report,* Latin American Laboratory for Evaluation of the Quality of Education, Santiago, Chili.

Center For Global Development (2006), *When Will we Ever Learn? Improving Lives through Impact Evaluation,* Disponible à l'adresse : www.cgdev.org/files/7973_file_WillWeEverLearn.pdf

Centre International d'Études Pédagogiques – Ciep (2007), « Les évaluations en éducation au niveau international : impacts, contradictions, incertitudes ». Réflexions et données extraites du séminaire international « L'évaluation au service de la qualité en éducation : pratique et enjeux », Paris.

Chabbott, C. (2006), "Accelerating Early Grades Reading in High Priority EFA Countries: A Desk Review", Disponible à l'adresse : http://www.equip123.net/docs/E1-EGRinEFACountriesDeskStudy.pdf

Chinapah, V. (2003), "Monitoring Learning Achievement (MLA) Project in Africa", ADEA Biennal Meeting, 3-6 décembre 2003, Grand Baie, Maurice.

Chudgar, A. et T.F. Luschei (2009), "National Income, Income Inequality, and the Importance of Schools: A Hierarchical Cross-National Comparison", *American Educational Research Journal,* 46 (3), pp.626-658

CONFEMEN (2011), « Synthèse des résultats PASEC VII, VIII et IX », PASEC, Dakar.

CONFEMEN (2009), « Étude PASEC Burkina Faso. Les apprentissages scolaires au Burkina Faso : Les effets du contexte, les facteurs pour agir », PASEC, Dakar.

CONFEMEN (2008*a*), « Vers la scolarisation universelle de qualité pour 2015, Évaluation diagnostique. GABON », PASEC, Dakar.

CONFEMEN (2008*b*), « Rapport PASEC Cameroun 2007 », PASEC, Dakar.

CONFEMEN (2008*c*), « L'enseignement primaire : la qualité au cœur des défis, Évaluation diagnostique PASEC Maurice », PASEC, Dakar.

CONFEMEN (2008*d*), « Rapport PASEC Madagascar 2004 », PASEC, Dakar.

CONFEMEN (2007*a*), « Quelques pistes de réflexion pour une éducation primaire de qualité pour tous, Évaluation diagnostique du Madagascar », PASEC, Dakar.

CONFEMEN (2007*b*), « Évaluation PASEC Sénégal », PASEC, Dakar.

CONFEMEN (2004), « Les enseignants contractuels et la qualité de l'enseignement de base I au Niger : quel bilan ? », *Document de travail* PASEC, Dakar, Disponible à l'adresse : http://www.confemen.org/le-pasec/rapports-et-documents-pasec/les-rapports-du-pasec/

Coombs, P.H. (1985), *The World Crisis in Education: The View from the Eigthies,* Oxford University Press, Oxford.

Coulombe, S. et J.-F. Tremblay (2006), "Literacy and Growth", *Topics in Macro-economics,* 6(2), Berkeley Electronic Press, Berkeley, Disponible à l'adresse : http://www.sba.muohio.edu/davisgk/growth%20readings/10.pdf

Dickes, P. et P. Vrignaud (1995), « Rapport sur les traitements des données françaises de l'enquête internationale sur la littératie », Rapport pour le ministère de l'Éducation nationale. direction de l'Evaluation et de la Prospective, Paris.

Dolata, S. (2005), "Construction and Validation of Pupil Socio-Economic Status Index for the SACMEQ Education Systems", SACMEQ *Working Document,* Paris.

Duru-Bellat, M et J.-P. Jarousse (2001), "Portées et limites d'une évaluation des politiques et des pratiques éducatives", *Éducation et Sociétés,* 8(2), 97-109, De Boeck Université, Bruxelles.

Elley, W.B. (Ed.), (1992), *How in the World do Student Read?,* Grindeldruck GMBH, Hambourg.

ENCINAS-MARTIN, M. (2006), "A Global Survey of Educational Evaluation: International, Regional, and National Assessments of Student Learning", Article commandé pour le Rapport de suivi de l'EPT 2007, *"Strong Foundations: Early Childhood Care and Education".*

EURYDICE (2009), Les évaluations standardisées des élèves en Europe: objectifs, organisation et utilisation des résultats, Agence exécutive Éducation, Audiovisuel et Culture, Comission européenne, Bruxelles, dispoible à l'adresse : http://eacea.ec.europa.eu/education/eurydice/documents/thematic_reports/109FR.pdf

FLIELLER, A. (1989), "Application du modèle de Rasch à un problème de comparaison de générations", *Bulletin de Psychologie,* XLII, pp.86-91.

GAMERON, A. et D.A. LONG (2007), "Equality of Educational Opportunity: a 40-year Retrospective" *in* TEESE, R., S. LAMB et M. DURU-BELLAT (Eds.), *International Studies in Educational Inequality, Theory and Policy, Vol. 1, Educational Inequality, Persistence and Change,* Springer, Dordrecht.

GIPPS C. et P. MURPHY (1994), *A Fair Test? Assessment, Achievement and Equity,* Open University Press, Buckingham, Philadelphia.

GOASTELLEC, G. (2003), "Le SAT et l'accès aux études supérieures : le recrutement des élites américaines en question", *Sociologie du travail,* 45, pp. 473-490.

GOLDSTEIN, H. (2004), "International Comparisons of Student Attainment: Some Issues Arising from the PISA Study", *Assessment in Education,* 11. pp.319-330.

GOLDSTEIN, H., G. BONNET et T. ROCHER (2007), "Multilevel Stuctural Equation Models for the Analysis of Comparative Data on Educational Performance", *Journal of Educational and Behavioral Statistics,* 32, 3, pp.252-286.

GOVE, A. (2009), *EGRA FAQs,* Disponible à l'adresse : https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&ID=95.

GREENEY, V. et T. KELLAGHAN (2008), "Assessing National Achievement Levels in Education", *National Assessment of Educational Achievement,* Volume 1, Banque mondiale, Washington, D.C.

GRISAY, A. et P. GRIFFIN (2005), "What Are the Main Cross-National Studies?" in ROSS, K.N. et I.J. GENEVOIS (Eds.), *Cross-National Studies of the Quality of Education: Planning Their Design and Managing Their Impact,* pp.67-104, IIEP, Paris.

Groupe Indépendant d'Évaluation de la Banque Mondiale (2006), *From Schooling Access to Learning Outcomes–An Unfinished Agenda: An Evaluation of World Bank Support to Primary Education,* Washington, DC.

Guerin-Pace, F. et A. Blum. (1999), « L'illusion comparative. Les logiques d'élaboration et d'utilisation d'une enquête internationale sur l'illettrisme », *Population,* 54e année, n° 2., pp.271-302.

Gurgand, M. (2000), « Capital humain et croissance : la littérature empirique à un tournant ? », *Économie Publique,* vol. 6, p.71-93.

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

Hambleton, R.K., H. Swaminathan et H.J. Rogers (1991), *Fundamentals of Item Response Theory.:* Sage, Newbury Park, CA.

Hanushek, E.A. (2006), "School Resources" *in* Hanushek, E.A. et F. Welch (Eds.), *Handbook of the Economics of Education,* Chapter 4, 865-908, Elsevier.

Hanushek, E.A. (2003), "The Failure of Input-Based Schooling Policies", *Economic Journal,* vol. 113(485), pp. F64-F98

Hanushek, E.A. et D.D. Kimko (2000), "Schooling, Labor-Force Quality, and the Growth of Nations", *American Economic Review,* 90(5), 1184-1208.

Hanushek, E. A et J.A. Luque (2003), "Efficiency and Equity in Schools Around the World", *Economics of Education Review,* vol. 22(5), pp. 481-502

Hanushek, E.A. et L. Woessmann (2010), "The Economics of International Differences in Educational Achievement", *NBER Working Paper,* 15949, NBER, Cambridge, MA.

Hanushek, E.A. et L. Woessmann (2007), "The Role of Education Quality in Economic Growth", *World Bank Policy Research Working Paper,* 4122, Banque mondiale, Washington, D.C.

Harlen, W. (2007), Assessment of Learning, Sage, Londres.

Heyneman, S. et W. Loxley (1983), "The Distribution of Primary School Quality Within High and Low Income Countries", *Comparative Education Review* 27 (1), pp.108–118.

Houngbedji, K. (2009), « Guide méthodologique PASEC. Module pondération », CONFEMEN/PASEC.

Howie, S. (2000), "TIMSS-R in South Africa: a Developing Country Perspective", article présenté lors de la réunion annuelle de l'American Educational Research Association, 24-28 avril, New Orleans.

Howie, S. et C. Hughes (2000), "South Africa", *in* Robitaille, D., A. Beaton et T. Plomb (Eds.). *The Impact of TIMSS on the Teaching and Learning of Mathematics and Science,* pp. 139-145. Pacific Educational Press, Vancouver, BC.

Husén, T. (Ed.) (1967), *A Comparison of Twelve Countries: International Study of Achievement in Mathematics* (Vols. 1-2), Almquist & Wiksell., Stockholm.

Jarousse, J.-P. et A. Mingat (1993), *L'école primaire en Afrique : analyse pédagogique et économique ; le cas du Togo,* L'Harmattan, Paris.

Kellaghan, T. (2004), "Public Examinations, National and International Assessments, and Educational Policy", *Mimeo,* Educational Research Centre, St Patrick's College, Dublin.

Kellaghan, T. et V. Greaney (2004), *Assessing Student Learning in Africa,* Banque mondiale, Washington, D.C.

Kellaghan, T. et V. Greaney (2001), "The Globalisation of Assessment in the 20th Century", *Assessment in Education,* 8(1), 87-102.

Kennedy, A.M., I.V.S. Mullis, M.O. Martin et K.L. Trong (2007), *PIRLS 2006 Encyclopedia. A Guide to Reading Education in the Forty PIRLS 2006 Countries,* Boston College Press, Boston.

Korda, M. (2009), "Early Grade Reading Assessment (EGRA) Plus: Liberia", Document prepare pour la reunion annuelle de la Comparative and International Education Society (CIES), 25 mars, Charleston (Caroline du Sud), Disponible à l'adresse: https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&ID=172.

Kudo, I. et J. Bazan (2009), "Measuring Beginner Reading Skills. An Empirical Evaluation of Alternative Instruments and their Potential Use for Policymaking and Accountability in Peru", *World Bank Policy Research Working Paper,* n°4812, Banque mondiale, Washington, D.C.

Laveault, D. et J. Grégoire (2002), *Introduction aux théories des tests en sciences humaines,* 2e édition, De Boeck Université, Bruxelles.

Lazear, E.A. (2003), "Teacher Incentives", *Swedish Economic Policy Review,* 10(3), 179-214.

Lee, D.-W. et T.H. Lee (1995), "Human Capital and Economic Growth: Tests Based on the International Evaluation of Educational Achievement", *Economics Letters,* 47(2), 219-225.

Leuven, E., H. Oosterbeek et H. Ophen (Van) (2004), "Explaining International Differences in Male Skill Wage Differentials by Differences in Demand and Supply of Skill", *Economic Journal,* 114(495), 466-486.

Lewis, E.G. et C.E. Massad (1975), The Teaching of English as a Foreign Language in Ten Countries, Almquist & Wiksell (Stockholm) et John Wiley (New York).

Lockheed, M.E., A.M. Verspoor., D. Bloch, P. Englebert, B. Fuller, E. King, J. Middleton, V. Paqueo, A. Rodd, R. Romain et M. Welmond (1991), *Improving Primary Education in Developing Countries,* Banque mondiale, Washington, D.C.

Lord, F. et M.R. Novick (Eds) (1968), *Statistical Theories of Mental Test Scores,* Addison-Westley, Reading, MA.

Mankiw N., D. Romer et D. Weil (1992), "A Contribution to the Empirics of Economic Growth", *Quarterly Journal of Economics,* 107, 407-437.

Martin, M.O., I.V.S. Mulli et P. Foy (2008), TIMSS 2007 *International Science Report,* Boston College Press, Boston.

Mincer, J. (1974), *Schooling, Experience, and Earnings,* National Bureau of Economic Research Press, New York.

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

Mislevy, R.J. et N. Verghelst (1990), "Modeling Item Responses when Different Subjects Employ Different Solution Strategies", *Applied Psychological Measurement,* 17, pp.297-334.

Mislevy, R.J., K.M. Sheehan et M. Wingersky (1993), "How to Equate Tests with Little or No Data", *Journal of Educational Measurement,* 30, pp.55-78.

Mulligan, C.B. (1999), "Galton Versus the Human Capital Approach to Inheritance", *Journal of Political Economy,* 107(6), S184-S224.

MULLIS, I.V.S. et L.B. JENKINS (1990), *The Reading Report Card 1971-1988 : Trends from the Nation's Report Card,* Educational Testing Service, Princeton, New Jersey.

MULLIS, I.V.S., M.O. MARTIN et P. FOY (2008), *TIMSS 2007 International Report,* Boston College Press, Boston.

MULLIS, I.V.S., M.O. MARTIN, A.M KENNEDY et P. FOY (2007), *PIRLS 2006 International Report,* Boston College Press, Boston.

MULLIS, I.V.S., M.O. MARTIN, J.F. OLSON, D.R. BERGER, D. MILNE et G.M. STANCO (2009), *TIMSS 2007 Encyclopedia,* Boston College Press, Boston.

MURNAME, R.J., J.B. WILLET, Y. DUHALDEBORDE et J.H. TYLER (2000), "How Important Are the Cognitive Skills of Teenagers in Predicting Subsequent Earnings?", *Journal of Policy Analysis and Management,* 19(4), 547-568.

NEUWIRTH, E. (2006), "PISA 2000: Sample Weights Problems in Austria", *OECD Education Working Papers,* No. 5, OCDE, Paris.

NIDEGGER, C. (2008), *PISA 2006 : Compétences des jeunes romands. Résultats de la troisième enquête PISA auprès des élèves de 9e année.* INRP, Neuchâtel.

OCDE (2011), *PISA 2009 Results: Students On Line: Digital Technologies and Performance, Volume VI.* OCDE, Paris.

OCDE (2010), *PISA 2009 Results: Learning Trends: Changes in Student Performance since 2000, Volume V.* OCDE, Paris.

OCDE (2009*a*), *PISA 2006 Science Competencies for Tomorrow's World,* Volume 2, OCDE, Paris.

OCDE (2009*b*), *PISA 2006 Technical Report,* OCDE, Paris.

OCDE (2009*c*), *Top of the Class. High Performers in Science in PISA 2006,* OCDE, Paris.

OCDE (2009*d*), *PISA Take the Test: Sample Questions from the OCDE's PISA Assessments,* OCDE, Paris.

OCDE (2007), *PISA 2006 Science Competencies for Tomorrow's World,* OCDE, Paris.

OCDE et STATISTIQUE CANADA (2009), EIAA.

OCDE et Statistique Canada (2005), *Learning a Living: First Results of the Adult Literacy and Life Skills Survey,* Ottawa et Paris.

OCDE et Statistique Canada (2000), *La littératie à l'ère de l'information : Rapport final de l'Enquête internationale sur la littératie des adultes,* Ottawa et Paris.

OCDE et UNESCO Institute for Statistics (2003), *Literacy Skills for the World of Tomorrow: Further Results from PISA 2000,* OCDE et UNESCO IS, Paris et Montreal.

Office Fédéral De Statistique – OFS (2007), « PISA 2006 : Les compétences en sciences et leur rôle dans la vie, Rapport national », Série Statistique de la Suisse, Neuchâtel.

Oliver, R. (1999), "Fertility and Women's Schooling in Ghana", in *The Economics of School Quality Investments in Developing Countries,* pp. 327-344, ed. P. Glewwe, St. Martin's, New York.

Olson, J.F., M.O. Martin et I.V.S. (2008), *TIMSS 2007 Technical Report,* International Association for the Evaluation of Educational Achievement (IEA), TIMSS et PIRLS International Study Center, Boston College, Boston.

Onsomu, E., Nzomo, J. et C. Obiero (2005), *The SACMEQ II Project in Kenya: A Study of the Conditions of Schooling and the Quality of Education,* SACMEQ, Harare, Zimbabwe.

Piper, B. et M. Corda (2011), *EGRA Plus: Liberia. Program Evaluation Report,* USAID, RTI International.

Pôle de Dakar (2007), "Les acquisitions scolaires et la production d'alphabétisation de l'école primaire en Afrique : approches comparatives", *Note thématique* n°2, UNESCO-BREDA, Dakar

Postlethwaite, T.N. (2004), *Monitoring Educational Achievement,* UNESCO International Institute for Educational Planning (Fundamentals of Educational Planning, 81), Paris.

Postlethwaite, T.N. et K.N. Ross (1992), *Effective Schools in Reading: Implications for Educational Planners: An Explonatory Study, The IEA Study of Reading Literacy II,* IEA, Hambourg.

Pouezevara, S., M. Sock, et A. Ndiaye (2010), *Évaluation des compétences fondamentales en lecture au Sénégal, Rapport d'analyse,* RTI International et FocusAfrica, Dakar.

Pritchett, L. (2001), "Where Has All the Education Gone?", *World Bank Economic Review,* 15, 367-391, Banque mondiale, Washington, D.C.

Psacharopoulos, G. et H. Patrinos (2004), "Returns to Investment in Education: A Further Update", *Education Economics,* 12(2), 111-134.

Reimers, F. (2000), "Educational Opportunity and Policy in Latin America", in *Unequal Schools, Unequal Chances,* Ed. F. Reimers, Harvard University Press, Cambridge, MA.

Reuchlin, M. (1997), *La psychologie différentielle,* Nouvelle édition entièrement refondue, PUF, Paris.

Rojas, C. et J.M. Esquivel (1998), "Los Sistemas de Medicion del Logro Acamedico en Latino América", *LCSHD Paper 25,* Banque mondiale, Washington, D.C.

Ross, K.N. (1995), "From Educational Research to Educational Policy: An Example from Zimbabwe", *International Journal of Educational Research,* 23(4), pp.301-403.

Ross, K.N. et T.N. Postlethwaite (1991), *Indicators of the Quality of Education: A Study of Zimbabwean Primary Schools,* Ministère de l'Education et de la Culture (Harare) et International Institute for Educational Planning, Paris.

Rothstein, J. (2010), "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement", *The Quarterly Journal of Economics,* MIT Press, vol. 125(1), pp. 175-214.

RTI International (2009), *Early Grade Reading Assessment Toolkit,* Document préparé pour le Bureau du développement humain, de la Banque mondiale, Washington, D.C.

Russell, H. (1982), "Total Score. Subscore Group: Comments", Memorandum IEA (Maths-NZ) A./362. Department of Education, Wellington.

Russell, H. (1981), "Validity Patterns and the Total Test Score Variable", *Mimeo,* OISE, Toronto.

Sacmeq (2010), "SACMEQ III Project Results: Pupil Achievement levels in reading and mathematics", *Working Document,* No. 1, SACMEQ.

Sacmeq (2005), *The SACMEQ II Project in Kenya: A Study of the Conditions of Schooling and the Quality of Education. Kenya Working Report,* Harare, Zimbabwe.

SAITO, M. (2005), "The Construction of a 'SACMEQ School Resources Index' Using Rasch Scaling", Article présenté lors de la conference de la SACMEQ, 28 septembre, Paris.

SAKELLARIOU, C. (2006), "Cognitive Ability and Returns to Schooling in Chile", article rédigé pour VEGAS, E. et J. PETROW (2008), "Raising Student Learning in Latin America, The Challenge for the 21st Century", *Latin American Development Forum Series,* Banque mondiale, Washington, D.C.

SEYMOUR, P. H. K., M. ARO et J.M. ERSKINE (2003), "Foundation Literacy Acquisition in European Orthographies", *British Journal of Psychology,* 94, 143–174, Disponible à l'adresse : http://onlinelibrary.wiley.com/doi/10.1348/000712603321661859/pdf

SHABALALA, J. (2005), *The SACMEQ II Project in Swaziland: A Study of the Conditions of Schooling and the Quality of Education,* Ministère de l'Éducation et de la Culture (Harare) et International Institute for Educational Planning (Paris).

SIKA, G. L. (2011), *Impact des allocations en ressources sur l'efficacité des écoles primaires en Côté d'Ivoire,* Thèse de sciences économiques, IREDU/Université de Bourgogne.

SJOBERG, S. (2007), "PISA and 'Real Life Challenges': Mission Impossible?", *in* S.T. HOPMANN, G. BRINEK et M. RETZL (Eds.), PISA *According to* PISA, *Does PISA Keep What It Promises?,* Disponible à l'adresse :
http://www.univie.ac.at/pisaaccordingtopisa/pisazufolgepisa.pdf

STANOVICH, K. E. (1986), "Matthew Effects in Reading: Some Consequences of Individual Differences in the Acquisition of Literacy", *Reading Research Quarterly,* 21, 360-406.

THOMAS, D. (1999), "Fertility, Education and Resources in South Africa", *in* C.H. BLEDSOE, J.B. CASTERLINE, J.A. JOHNSON-KUHN et J.G. HAAGA (Eds), *Critical Perspectives on Schooling and Fertility in the Developing World,* National Academic Press, Washington, D.C.

THORNDIKE, R.L. (1973), *Reading Comprehension Education in Fifteen Countries: An Empirical Study,* Almquist et Wiksell, Stockholm.

UIS-UNESCO (2010), « Programme d'évaluation et de suivi de l'alphabétisation (LAMP) Mise à jour n°1 », Montréal.

UNESCO (2004), *EFA Global Monitoring Report 2005: The Quality Imperative,* Paris.

UNESCO (2003), Table ronde ministérielle sur "l'éducation de qualité" des 3-4 octobre 2003, dans le cadre de la 32e assemblée générale, Paris.

UNESCO (2000), « Cadre d'action de Dakar : L'Education pour tous. Tenir nos engagements collectifs », Forum mondial sur l'éducation, Dakar, Sénégal, 26-28 avril 2000, UNESCO, Paris.

UNESCO IIEP (2010), "In Search of Quality: What the Data Tell Us", *IIEP newsletter*, Vol XXVIII, No. 3, septembre-décembre 2010, Paris.

UNESCO-OREALC (2008), Student Achievement in Latin America and the Caribbean, Results of the Second Regional Comparative and Explanatory Study (SERCE), Laboratorio Latinoamericano de Evaluacion de la Calidad de la Educacion, Santiago, Chili.

USAID (2009), EDDATA II, *Manuel pour l'évaluation des compétences fondamentales en lecture.*

Vaniscotte F. (1996), *Les écoles de l'Europe – systèmes éducatifs et dimension européenne,* INRP, Paris.

Vegas, E. et J. Petrow (2008), *Raising Student Learning in Latin America, The Challenge for the 21st Century, The World Bank, Latin American Development Forum Series,* Banque mondiale, Washington, D.C.

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

Vrignaud, P. (2005), « L'INETOP : 75 années dans l'histoire de l'évaluation psychologique et pédagogique en France », *L'Orientation Scolaire et Professionnelle.*

Wainer, H. et D. Thissen (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,* 15, pp.22-29.

Walker, D.A. (1976), *The IEA Six-Subject Survey: An Empirical Study of Education in Twenty-One Countries,* Almquist et Wiksell (Stockholm) et John Wiley et Sons (New York).

Wilkins, J.L.M., M. Zembylas et K.J. Travers (2002), "Investigating Correlates of Mathematics and Science Literacy in the Final Year of Secondary School" *in* D.F. Robitaille et A.E. Beaton (Eds.), *Secondary Analysis of the TIMSS Data,* 291-316, Kluwer Academic, Dordrecht.

Wu, M. (2010), "Comparing the Similarities and Differences of PISA 2003 and TIMSS", *OECD Education Working Papers,* No. 32, OCDE, Paris.

Wuttke, J. (2008), "Uncertainties and Bias in PISA" in S.T. Hopmann, G. Brinek et M. Retzl (Eds.), PISA According to PISA. Does PISA Keep What It Promises?, Disponible à l'adresse : http://www.univie.ac.at/pisaaccordingtopisa/pisazufolgepisa.pdf (pp.241-264).

Précédentes publications de la collection

À SAVOIR N° 1 : La régulation des services d'eau et d'assainissement dans les PED
The Regulation of Water and Sanitation Services in DCs

À SAVOIR N° 2 : Gestion des dépenses publiques dans les pays en développement
Management of public expenditure in developing countries

À SAVOIR N° 3 : Vers une gestion concertée des systèmes aquifères transfrontaliers
Towards concerted management of cross-border aquifer systems

À SAVOIR N° 4 : Les enjeux du développement en Amérique latine
Development issues in Latin America

À SAVOIR N° 5 : Transition démographique et emploi en Afrique subsaharienne
Demographic transition and employment in Sub-Saharan Africa

À SAVOIR N° 6 : Les cultures vivrières pluviales en Afrique de l'Ouest et du Centre
Rain-fed food crops in West and Central Africa

À SAVOIR N° 7 : Les paiements pour services environnementaux
Payments For Ecosystem Services

À SAVOIR N° 8 : Les accords de libre-échange impliquant des pays en développement ou des pays moins avancés

À SAVOIR N° 9 : Comment bénéficier du dividende démographique ?
La démographie au centre des trajectoires de développement
How Can We Capitalize on the Demographic Dividend?
Demographics at the Heart of Development Pathways

À SAVOIR N° 10 : Le risque prix sur les produits alimentaires importés – Outils de couverture pour l'Afrique

À SAVOIR N° 11 : La situation foncière en Afrique à l'horizon 2050

À SAVOIR N° 12 : *Contract Farming in Developing Countries – A Review*

À SAVOIR N° 13 : Méthodologies d'évaluation économique du patrimoine urbain : une approche par la soutenabilité

À SAVOIR N° 14 : *Creating Access to Agricultural Finance – Based on a horizontal study of Cambodia, Mali, Senegal, Tanzania, Thailand and Tunisia*

À SAVOIR N° 15 : *The Governance of Climate Change in Developing Countries*

Qu'est-ce que l'AFD ?

Établissement public, l'Agence Française de Développement (AFD) agit depuis soixante-dix ans pour combattre la pauvreté et favoriser le développement dans les pays du Sud et dans l'Outre-mer. Elle met en œuvre la politique définie par le Gouvernement français.

Présente sur quatre continents où elle dispose d'un réseau de 70 agences et bureaux de représentation dans le monde, dont 9 dans l'Outre-mer et 1 à Bruxelles, l'AFD finance et accompagne des projets qui améliorent les conditions de vie des populations, soutiennent la croissance économique et protègent la planète : scolarisation, santé maternelle, appui aux agriculteurs et aux petites entreprises, adduction d'eau, préservation de la forêt tropicale, lutte contre le réchauffement climatique…

En 2011, l'AFD a consacré plus de 6,8 milliards d'euros au financement d'actions dans les pays en développement et en faveur de l'Outre-mer. Ils contribueront notamment à la scolarisation de 4 millions d'enfants au niveau primaire et de 2 millions au niveau collège, et à l'amélioration de l'approvisionnement en eau potable pour 1,53 million de personnes. Les projets d'efficacité énergétique sur la même année permettront d'économiser près de 3,8 millions de tonnes d'équivalent CO_2 par an.

www.afd.fr

Agence Française de Développement
5, rue Roland Barthes – 75598 Paris cedex 12
Tél. : 33 (1) 53 44 31 31 – **www.afd.fr**
Dépôt légal : 4e trimestre 2012
ISSN : 2105-553X

Renforcer la mesure sur la qualité de l'éducation

Analyse comparative des évaluations sur les acquis des élèves au sein des pays en développement

Les réflexions en cours sur le développement et les inégalités accroissent l'intérêt de mesurer les savoirs et les compétences des populations jeunes. Parmi les objectifs du Millénaire pour le développement, la lutte contre l'analphabétisme se fonde sur une éducation de qualité. La question de la création d'outils de mesure de la qualité de l'éducation est donc essentielle pour valider, universellement et équitablement, l'acquisition des apprentissages fondamentaux. Cette question est directement liée au développement, depuis 50 ans, de méthodes d'évaluation des acquisitions à l'origine des grandes enquêtes internationales comme PIRLS, TIMSS et PISA.

Dans cet effort de connaissance sur la qualité de l'éducation, cet ouvrage analyse les objectifs des principales enquêtes d'évaluation scolaire et les méthodes de mesure qui y sont mises en œuvre. Les auteurs montrent comment ces outils s'adaptent à des terrains complexes, où l'information et la culture d'évaluation sont parfois limitées. Une attention toute particulière est ici portée à l'Afrique subsaharienne.

AUTEURS

Nadir ALTINOK
BETA (Bureau d'économie théorique et appliquée),
IREDU (Institut de recherche sur l'éducation), Université de Lorraine
nadir.altinok@univ-lorraine.fr

Jean BOURDON
IREDU (Institut de recherche sur l'éducation),
CNRS (Centre national de la recherche scientifique), Université de Bourgogne
jbourdon@u-bourgogne.fr

CONTACTS

Véronique SAUVAT
Division Recherche économique et sociale, AFD
sauvatv@afd.fr

Valérie TEHIO
Division Education et formation professionnelle, AFD
tehiov@afd.fr

A SAVOIR

afd AGENCE FRANÇAISE DE DÉVELOPPEMENT