ÉDUCATION & FORMATIONS

Évaluation des acquis : principes, méthodologie, résultats

n°86
87
mai 2015

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

MINISTÈRE
DE L'ÉDUCATION
NATIONALE, DE
L'ENSEIGNEMENT
SUPÉRIEUR ET DE
LA RECHERCHE

ÉDUCATION & FORMATIONS

Évaluation des acquis : principes, méthodologie, résultats

n°86
87
mai 2015

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

MINISTÈRE
DE L'ÉDUCATION
NATIONALE, DE
L'ENSEIGNEMENT
SUPÉRIEUR ET DE
LA RECHERCHE

Cet ouvrage est édité par **le ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche**

Direction de l'évaluation,de la prospective et de la performance
61-65, rue Dutot – 75 732 Paris Cedex 15

Direction de la publication : Catherine Moisan

Coordination : Thierry Rocher et Caroline Simonis-Sueur

Secrétariat de rédaction : Aurélie Bernardi

Conception graphique : délégation à la communication du ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche

Composition – PAO : Tony Marchois

Impression : Ovation

ISSN 0294-0868
ISBN 978-2-11-138951-9
Pour la version numérique,
ISBN 978-2-11-138958-8
Dépôt légal : mai 2015

PRÉSENTATION //

Catherine Moisan
Directrice de l'évaluation, de la prospective et de la performance (DEPP)

La revue *Éducation & formations* fait peau neuve. Avec ce numéro, elle se présente avec un nouveau visage et de nouvelles modalités de fonctionnement.

Sur la forme, la revue adopte désormais une nouvelle maquette, plus moderne, plus lisible, et surtout mieux adaptée à l'édition sur Internet. Rappelons que la revue est disponible gratuitement, en ligne, rendant ainsi accessible au plus grand nombre le résultat d'études scientifiques dans le domaine de l'éducation.

Au-delà de la physionomie de la revue, de nouvelles procédures de relecture et d'expertise des articles ont été définies. Engagée depuis plusieurs numéros, la formalisation de ces procédures a été mise en place et rigoureusement suivie pour ce numéro spécial. Chaque article de la revue est expertisé par deux relecteurs : un spécialiste du domaine concerné et un statisticien de la DEPP. Ce double regard, interne et externe, permet d'asseoir le caractère scientifique de la revue tout en conservant l'apport indispensable des compétences spécifiques de la DEPP, qui permettent de garantir la qualité des analyses statistiques produites.

Avec cette nouvelle version, la revue *Éducation & formations* conforte ainsi la place singulière qu'elle occupe dans le paysage des revues scientifiques en éducation, au croisement de la recherche scientifique et des missions de statistique publique, en lien avec la demande sociale. Gageons que cette nouvelle formule contribue à faire progresser de façon éclairée le débat public en éducation.

AVANT-PROPOS

Thierry Rocher

Le thème de l'évaluation est un sujet récurrent de débat dans le domaine de l'éducation. Encore est-il important de bien préciser ce dont il s'agit, tant les formes d'évaluations sont multiples : évaluation des élèves dans la classe par les enseignants, évaluation certificative, formative, évaluation diagnostique, auto-évaluation, etc. Très clairement, ce double numéro spécial de la revue *Éducation & formations* porte sur les « évaluations standardisées » des acquis. Organisées le plus souvent sous forme de « programmes », ces évaluations se distinguent par le fait qu'elles ambitionnent de fournir une mesure objective, scientifique, des acquis des élèves, la plus indépendante possible des conditions d'observation, de passation, de correction. En ce sens, elles sont « standardisées ». En outre, ces programmes d'évaluations apparaissent comme des évaluations externes, et non internes comme par exemple le contrôle continu ou l'évaluation formative. Si leurs objectifs peuvent sensiblement varier, ces évaluations ont en commun de pouvoir rendre compte des acquis des élèves au-delà du niveau individuel et, en particulier, d'apprécier les résultats du système éducatif pris dans sa globalité. Leurs modalités de conception sont relativement partagées, car tendues vers l'objectivité de la mesure et la comparabilité des résultats, entre groupes d'élèves, dans le temps, etc. Bien que leurs fondements méthodologiques restent assez méconnus en France, les résultats de ces évaluations sont largement diffusés et commentés dans le champ médiatique et politique. Ce numéro spécial est donc important, car il donne à voir les différentes facettes des évaluations standardisées des acquis des élèves, de leur conception à l'exploitation des résultats. Les cadres et les formateurs en particulier en verront tout l'intérêt.

Les treize articles de ce numéro sont organisés en trois parties. La première comprend quatre articles et livre un panorama général des problématiques, des usages et des concepts spécifiques aux évaluations standardisées. La seconde regroupe quatre articles portant plus spécifiquement sur des aspects méthodologiques. Enfin, cinq articles constituent la dernière partie qui rend compte de différents résultats issus de ces évaluations.

Pour ouvrir ce numéro, Bruno Trosseille et Thierry Rocher (*Les évaluations standardisées des élèves – perspective historique*) proposent une analyse historique du développement des programmes d'évaluations standardisées en France. Différentes périodes se sont succédé depuis leur naissance. Ces périodes ont correspondu à des objectifs spécifiques assignés à ces évaluations, objectifs qui doivent être clairement distingués sous peine de créer des confusions néfastes. Les fonctions de ces évaluations (servent-elles à observer, à sélectionner, à diagnostiquer, à piloter ?) ainsi que leurs usages (servent-elles aux élèves eux-mêmes, aux enseignants, aux responsables politiques ?) sont des questions fondamentales. Les auteurs

montrent à travers l'histoire récente que la poursuite d'objectifs différents pour une même évaluation peut conduire à une utilisation inefficace, voire dévoyée de leurs résultats.

La question des usages des résultats des évaluations est indissociable de la question de la mesure des compétences. Les problématiques de mesure relèvent du domaine de la psychométrie, domaine spécifique très peu investi en France, que ce soit dans le monde académique ou dans celui de la statistique publique. L'article de Thierry Rocher (*Mesure des compétences – Méthodes psychométriques utilisées dans le cadre des évaluations des élèves*) vise précisément à donner aux lecteurs de la revue un aperçu des procédures psychométriques, avec un souci pédagogique, sans toutefois sacrifier la rigueur statistique. L'article montre en particulier que tout instrument de mesure est un « construit », reposant sur un ensemble d'hypothèses et de méthodes spécifiques. Une illustration de ces méthodes est donnée à travers la présentation des procédures psychométriques employées dans le cadre du programme d'évaluations standardisées Cedre (Cycle des évaluations disciplinaires réalisées sur échantillons).

Ces méthodes sont globalement partagées au niveau international. L'ingénierie à mettre en place pour assurer la qualité et la comparabilité des données relève d'une forme de technicité bien décrite dans un autre contexte par Christophe Dierendonck, Amina Kafaï, Antoine Fischbach, Romain Martin et Sonja Ugen (*Les épreuves standardisées – Élément-clé du pilotage du système éducatif luxembourgeois*). Le programme d'évaluations standardisées mis en place au Luxembourg se distingue par ses objectifs, dans la mesure où il vise à aider au pilotage des établissements scolaires. Les établissements sont en effet accompagnés dans la rédaction d'un plan d'action reposant en partie sur l'analyse des données tirées des évaluations standardisées. Tout le défi est là, dépasser la méfiance envers un outil pouvant être perçu comme un contrôle, associer les acteurs aux constats et les accompagner dans la définition et la mise en place d'actions ciblées. La taille du Luxembourg favorise sans doute la réalisation de cet exercice difficile, mais précisément son expérience pourrait sans doute inspirer les acteurs du pilotage local en France, souvent demandeurs de dispositifs d'évaluations standardisées.

Pour conclure cette première partie de cadrage général, Fabrice Murat et Thierry Rocher (*L'évaluation des compétences des adultes – Quelles contraintes ? Quelles spécificités ?*) nous invitent à sortir du cadre scolaire et à considérer les évaluations d'adultes. Dans ce domaine, deux traditions statistiques, peu amenées à se côtoyer, doivent converger, à savoir celle des enquêtes-ménage, très maîtrisée par l'Insee, et celle de l'évaluation des compétences, naturellement plus développée à la DEPP. Une perspective de comparaison avec les évaluations des élèves montre les spécificités de l'évaluation des compétences des adultes, notamment le fait d'interroger des individus, chez eux, sur leurs compétences, ce qui peut réveiller de mauvais souvenirs scolaires chez certains d'entre eux. La question de leur implication dans

la situation d'évaluation est alors centrale. Les auteurs détaillent deux dispositifs d'évaluation – l'un national (IVQ), l'autre international (Piaac) – complémentaires, mais qui affichent des différences méthodologiques liées à différents objectifs ainsi qu'aux leçons que la France a pu tirer de précédentes enquêtes sur le sujet.

Les quatres articles suivants portent sur des questions méthodologiques précises : méthodes de sondage, questions de motivation, seuil de maîtrise et passage au numérique.

En France, les programmes d'évaluations standardisées ont pour objectif le pilotage d'ensemble du système éducatif ; ils sont donc principalement réalisés sur échantillons. Dès lors, des questions d'ordre technique se posent. Elles sont abordées par Émilie Garcia, Marion Le Cam et Thierry Rocher (*Méthodes de sondages utilisées dans les programmes d'évaluations des élèves*) : comment tirer un échantillon « représentatif » de la population ? Comment éviter de tirer deux fois la même école pour deux évaluations différentes ? Comment tenir compte des non-réponses ? Comment mesurer l'erreur d'échantillonnage inhérente à cette démarche ? Les auteurs proposent des réponses à travers une présentation pédagogique des méthodes et selon une démarche empirique, reposant sur des simulations. En filigrane, les auteurs montrent tout l'intérêt de tenir compte des informations disponibles en amont sur les élèves ou sur les établissements, pour améliorer la qualité des échantillons et des résultats. Notons avec eux au passage que ces informations sont aujourd'hui inégalement disponibles dans notre système : présentes dans le second degré, quasi-absentes dans le premier degré.

Une fois les élèves sélectionnés, se pose la question de leur participation et de leur implication dans l'activité d'évaluation. De nombreux travaux, notamment issus de la psychologie sociale, ont montré l'importance du contexte de l'évaluation sur les résultats eux-mêmes. En l'occurrence, les programmes d'évaluations standardisés en cours en France ne comportent aucun enjeu pour les élèves. Dans un système largement dominé par la notation, les élèves ne sont pas notés à l'issue de la passation d'une évaluation standardisée, n'ont pas connaissance de leurs résultats individuels et ne reçoivent pas d'incitations financières (à l'inverse de certains pays pour l'évaluation PISA). Dès lors, quel degré d'implication peut-on attendre des élèves ? Comment le mesurer ? Et quel lien peut-on faire entre leur niveau d'implication et leurs résultats ? C'est à ces questions que tentent de répondre Saskia Keskpaik et Thierry Rocher (*La motivation des élèves français face à des évaluations à faibles enjeux – Comment la mesurer ? Son impact sur les réponses*) en proposant un instrument de mesure de la motivation des élèves envers l'évaluation.

À supposer que les élèves aient témoigné d'un degré de motivation satisfaisant et que l'on puisse les classer selon un score en fonction de leurs réussites et de leurs échecs aux items qu'ils ont passés, certains programmes doivent alors produire un résultat « normatif ». C'est le cas des indicateurs calculés

dans le cadre de la LOLF (Loi organique relative aux lois de finances) qui reposent sur les pourcentages d'élèves qui maîtrisent les compétences du socle commun. Nicolas Miconnet et Ronan Vourc'h (*Détermination de standards minimaux pour évaluer les compétences du socle commun*) montrent que le statisticien ne peut pas répondre seul à la question de la détermination de seuils de réussite. En effet, il s'agit de se prononcer sur la « maîtrise » d'une compétence, au regard des objectifs visés et de ce qui est proposé dans l'évaluation. Les auteurs décrivent et appliquent des méthodes spécifiques développées pour répondre à cette problématique. Elles visent à croiser le jugement d'« experts » pédagogiques avec un corpus de données empiriques. Étant donné la valeur normative du résultat produit par l'évaluation, il est important que ces méthodes soient explicitées.

Le dernier article de cette partie axée sur la méthodologie porte sur un sujet d'actualité : la transition des évaluations, habituellement passées sous un format papier-crayon, vers un format numérique. Pascal Bessonneau, Philippe Arzoumanian et Jean-Marc Pastor (*Une évaluation sous forme numérique est-elle comparable à une évaluation de type « papier-crayon » ?*) présentent les résultats de deux expériences consistant à comparer les résultats obtenus aux mêmes épreuves, envisagées selon les deux formats différents. Si les auteurs parviennent à identifier certains paramètres pouvant expliquer les écarts obtenus, ils montrent qu'il n'est pas aisé de contrôler l'effet du passage à un environnement numérique sur la difficulté de la tâche proposée. C'est l'un des défis à venir : répondre à la préoccupation légitime de continuité des séries de résultats, indépendamment du format d'interrogation. Cependant, au-delà de cette attitude « conservatrice » visant à une simple dématérialisation des évaluations, notamment pour des raisons financières, les futures évaluations devraient grandement profiter du passage au numérique en termes de contenu (format des items, données de navigation des élèves, items interactifs, etc.).

Après ces considérations méthodologiques, les cinq articles suivants portent sur des analyses et des résultats issus de données d'évaluations.

Tout d'abord, l'article de Sylvie Beuzon, Émilie Garcia et Corinne Marchois (*Les compétences des élèves français en anglais en fin d'école et en fin de collège – Quelles évolutions de 2004 à 2010 ?*) rend compte des résultats des évaluations Cedre en anglais, la focale étant mise sur l'évolution du niveau de compétence des élèves de 2004 à 2010. La période envisagée se révèle intéressante. À l'école, les résultats positifs peuvent être rapprochés de la politique d'apprentissage précoce de la deuxième langue. Au collège, en revanche, le tableau est moins glorieux : malgré les efforts de l'institution sur la place de l'oral dans l'enseignement de l'anglais, le niveau baisse et les inégalités augmentent. L'autre apport important de l'article est de montrer l'importance du caractère standardisé des évaluations. Ainsi, en fin de troisième, Cedre, tout comme l'enquête européenne ESLC, montre qu'au plus 40 % des élèves maîtrisent le niveau A2 du cadre de référence européen, alors que la validation de ce même niveau, telle qu'enregistrée dans le livret personnel de compétences (LPC) dépasse 90 %...

Les comparaisons diachroniques ont l'avantage de révéler des points forts et des points faibles, en coupe, mais elles ne permettent pas de saisir la dynamique des évolutions de compétences. La DEPP conduit depuis les années 1960 des études longitudinales, appelées « panels ». Le panel d'élèves entrant en sixième à la rentrée 2007 est sans doute le panel le plus riche qu'ait conduit la DEPP : environ 35 000 élèves, évalués en fin de sixième, trois ans plus tard, en fin de troisième, deux prises d'informations sur les familles, des évaluations sur les aspects cognitifs mais également affectivo-motivationnels. Ce panel constituera très certainement une source majeure de connaissance pour la recherche en éducation dans les années futures. À partir de ces données, Linda Ben Ali et Ronan Vourc'h ont conduit une analyse sur l'évolution, de la sixième à la troisième, du niveau des élèves dans les différents domaines évalués, en fonction de leurs caractéristiques familiales. Cette étude est importante, car elle permet de renouveler les analyses sur les inégalités sociales à l'école. Elle montre que les inégalités sociales sont relativement figées de la sixième à la troisième, dans des disciplines telles que la lecture-compréhension et le raisonnement logique, mais qu'elles augmentent en mathématiques et en mémoire encyclopédique. C'est l'intérêt de cette étude de préciser la construction des inégalités sociales, en distinguant différentes dimensions des acquis, alors que les études sur les inégalités sociales portent très majoritairement sur les parcours ou les diplômes.

Il est difficile de réaliser un numéro spécial sur les évaluations des acquis sans faire référence aux évaluations internationales qui occupent une place importante dans ce domaine. Les nombreux rapports publiés rendent compte de résultats très variés, mais ils évoquent assez rarement les résultats sur le contenu même de l'évaluation. Rapidement, l'attention se concentre sur des palmarès globalisants des pays sur une échelle dont on perd de vue la façon dont elle a été construite. C'est précisément l'objectif de l'article d'Éric Roditi et de Franck Salles (*Nouvelles analyses de l'enquête PISA 2012 en mathématiques – Un autre regard sur les résultats*) que de revenir sur les principes de conception des items. Les auteurs sortent d'une analyse selon « le niveau en mathématiques » pour déconstruire le score global et montrer les leçons très utiles à tirer d'un point de vue pédagogique, dès lors que l'on adopte une grille de lecture pertinente et que l'on affine la granularité des analyses.

Toujours dans le domaine des mathématiques, l'article suivant offre néanmoins une toute autre perspective. Stéphane Herrero, Thomas Huguet et Ronan Vourc'h (*Évaluation des compétences des jeunes en numératie lors de la Journée défense et citoyenneté*) font état des résultats obtenus par les jeunes Français au test de numératie, introduit dans les tests de la Journée défense et citoyenneté (JDC) sur un large échantillon. La corrélation obtenue avec les résultats aux tests de lecture passés par tous les jeunes toute l'année montre la spécificité des difficultés des jeunes dans le domaine de l'usage des mathématiques. Cette évaluation se révèle riche d'enseignements, mais plus

généralement, le dispositif d'interrogation, adapté à l'interrogation massive de l'ensemble d'une génération, de façon relativement simple et efficace, renforce la JDC comme lieu d'observatoire de la jeunesse.

Pour clore ce numéro spécial, le dernier article porte sur l'évaluation standardisée au service de la mesure des effets d'une expérimentation. Marion Le Cam et Olivier Cosnefroy (*Évaluation des effets du dispositif expérimental d'enseignement intégré de science et technologie (EIST)*) présentent les résultats de l'évaluation du dispositif EIST. Le dispositif d'évaluation a consisté à suivre les élèves bénéficiant de l'EIST et de comparer leurs progressions à celles obtenues par un échantillon témoin. Le corpus de données est relativement unique : environ 4 000 élèves ont été suivis tout au long du collège, et évalués à cinq reprises, du début de la sixième à la fin de la troisième. Les effets de l'expérimentation ne sont pas concluants, pour des raisons dépassant certainement le seul contenu de l'expérimentation et certainement liées aux conditions de sa mise en œuvre. Si les résultats de l'expérimentation se révèlent décevants, les données recueillies constituent une source très riche d'informations sur les progressions des acquis en science et devraient renseigner sur le développement des compétences et des attitudes à l'égard des sciences au cours du collège.

SOMMAIRE

ÉDUCATION & FORMATIONS N° 86-87

III – ANALYSES ET RÉSULTATS DES ÉVALUATIONS

Ojectifs, construction et usage des évaluations

LES ÉVALUATIONS STANDARDISÉES DES ÉLÈVES

Perspective historique

Bruno Trosseille et **Thierry Rocher**
MENESR-DEPP, bureau de l'évaluation des élèves

Depuis une quarantaine d'années, le ministère de l'Éducation nationale a mis en œuvre des évaluations tantôt « de masse », tantôt sur échantillons. Ces évaluations peuvent avoir deux fonctions principales : de diagnostic lorsqu'elles sont élaborées pour fournir aux enseignants des outils professionnels qui leur sont nécessaires pour adapter leur enseignement en fonction des acquis de leurs élèves ; de bilan lorsque l'objectif est d'observer les acquis des élèves et leur évolution pour le pilotage d'ensemble du système éducatif. La confusion, dans une même évaluation, de ces deux fonctions est potentiellement source d'erreurs et de troubles, tant sur le plan scientifique que sociétal. Après avoir décrit l'histoire entrelacée de ces deux types d'évaluations au sein du Ministère, nous envisageons l'avenir du paysage évaluatif et la façon dont il peut se réorganiser en fonction des différentes finalités qui lui sont aujourd'hui assignées et des défis qu'il devra affronter à l'avenir.

Depuis quatre décennies, la DEPP (et les services et directions qui l'ont précédée)[1] met en place des dispositifs d'évaluation, spécifiques, nationaux, des acquis des élèves reposant sur des épreuves standardisées. Elle est également maître d'œuvre pour la France de diverses évaluations internationales (voir *infra*). Le développement des évaluations standardisées apparaît en effet, aux yeux des responsables des services statistiques, s'appuyant sur les exemples étrangers et internationaux, comme un complément indispensable des statistiques pour rendre compte du système et le piloter. Trois grandes périodes, qui se recouvrent peu ou prou, caractérisent le développement de ces dispositifs au ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche (MENESR).

1. On conviendra, par commodité, d'utiliser le sigle DEPP pour dénommer l'ensemble des services et directions qui ont précédé l'actuelle direction de l'évaluation, de la prospective et de la performance avec des missions d'évaluation (à savoir SEIS, SIGES, SPRESE, DEP, DPD).

Dans une première période, de la fin des années 1970 à la fin des années 1980, les dispositifs mis en œuvre à l'école et au collège, couvrent progressivement, niveau scolaire par niveau scolaire, l'ensemble des disciplines. Il s'agit, au regard des programmes en vigueur, d'établir un constat, d'apprécier l'état du système, de rendre compte des acquisitions des élèves aux responsables de la politique éducative du Ministère. Il s'agit aussi, déjà, de nourrir le débat public, ce qui sera plus nettement affiché au cours de la deuxième période. Dans le cadre d'un « observatoire permanent des acquis des élèves » des dispositifs d'évaluations sont systématiquement organisés, selon une méthodologie rigoureuse [Le Guen, 1991], sur des échantillons d'élèves, en fin d'année scolaire. Ils concernent la cinquième dès 1975, puis le CP dès 1979, et ensuite pratiquement tous les niveaux du primaire au lycée. Durant cette décennie, chaque niveau scolaire fera l'objet, une année donnée, d'une évaluation de type « bilan ».

Une deuxième période, à partir de 1989, est occupée par la mise en place d'évaluations diagnostiques « de masse », conséquences de la loi sur l'éducation de 1989, dite « loi Jospin ». Le rapport qui lui est annexé, relevant que *« moins d'un élève sur deux arrive au collège avec une maîtrise suffisante de la lecture »,* précise déjà l'urgence de la mise en œuvre d'un véritable plan sur l'apprentissage de la lecture et indique que *« cette acquisition fondamentale fera l'objet d'une évaluation auprès de tous les élèves entrant en cours élémentaire deuxième année et en sixième ; elle sera suivie d'actions de soutien ou de reprises d'apprentissage dans chaque école et chaque établissement scolaire »*. Dans son introduction à la revue *Éducation & formations* consacrée aux résultats nationaux des évaluations de septembre 1989, le ministre rappelle l'objectif de cette évaluation *« conçue comme un outil mis à votre disposition pour déceler, de façon précise et dès le début de l'année scolaire, les difficultés de vos élèves et vous permettre, dans toute la mesure du possible d'y apporter rapidement une réponse »* [*Éducation & formations,* 1990]. Les évaluations sur échantillons du type de celles de la première décennie se font alors plus rares (1990, 1995, 1999).

La troisième période, depuis le début du XXIe siècle, voit se systématiser, à côté des enquêtes internationales, notamment PISA[2] qui débute en 2000, des évaluations sur échantillons pour un bilan des acquis des élèves en fin d'école primaire et en fin de collège (Cedre[3]) et pour le calcul des indicateurs de maîtrise des compétences du socle commun destinés aux projets de lois de finances (LOLF) ▸ **Encadré**.
Durant cette période, de nouvelles évaluations « de masse » apparaissent brièvement, ayant pour but le repérage des élèves en difficulté vis-à-vis du socle commun (objectif de mise en place de remédiations ciblées en début de CE1, de 2005 à 2007, puis en début de CM2, en 2007), puis d'évaluations « bilans-diagnostics » (en fin de CE1 et en milieu de CM2 de 2009 à 2012) succédant à l'arrêt des évaluations « de masse » en CE2 et en sixième.
Avant de retracer de façon plus détaillée l'histoire, au MENESR, des deux grands types de dispositifs que sont les évaluations diagnostiques « de masse » et les évaluations de type bilan, faisons un petit détour méthodologique afin d'identifier ce qui les distingue.

2. Programme international pour le suivi des acquis des élèves, mené par l'OCDE.
3. Cedre : cycle des évaluations disciplinaires réalisées sur échantillons.

LE CYCLE DES ÉVALUATIONS DISCIPLINAIRES RÉALISÉES SUR ÉCHANTILLON (CEDRE)

Ce cycle d'évaluations établit des bilans nationaux des acquis des élèves en fin d'école et en fin de collège. Il couvre les compétences des élèves dans la plupart des domaines disciplinaires en référence aux programmes. La présentation des résultats permet de situer les performances des élèves sur des échelles de niveau allant de la maîtrise pratiquement complète de ces compétences à une maîtrise bien moins assurée, voire très faible, de celles-ci. Renouvelées tous les six ans (tous les cinq ans à partir de 2012), ces évaluations permettent de répondre à la question de l'évolution du « niveau des élèves » au fil du temps.

Le Calendrier

2003 – 2009 – 2015 : maîtrise de la langue française et compétences générales.
2004 – 2010 – 2016 : langues étrangères.
2005 : attitudes à l'égard de la vie en société (non repris par la suite).
2006 – 2012 – 2017 : histoire-géographie, éducation civique.
2007 – 2013 – 2018 : sciences expérimentales.
2008 – 2014 – 2019 : mathématiques.

Les élèves concernés

La population visée est celle des élèves de CM2 et de troisième générale des collèges publics et privés sous contrat de France métropolitaine. Pour les écoles, un échantillon représentatif est constitué au niveau national et tous les élèves de CM2 y passent les évaluations. Pour les collèges, des classes de troisième sont sélectionnées aléatoirement en vue d'une représentativité nationale. Pour chaque niveau scolaire, de 5 000 à 10 000 élèves, répartis dans plusieurs centaines de classes, sont évalués. Pour plus de détails, le lecteur est invité à consulter l'article de Garcia, Le Cam, Rocher [dans ce numéro, p. 101].

La comparabilité entre les différentes années d'évaluation

Afin de pouvoir comparer les résultats des enquêtes entre la première itération et les suivantes, une partie des items de la première est reprise à l'identique dans l'évaluation suivante (items dits « d'ancrage »). La mise en œuvre de modèles psychométriques adaptés – les modèles de réponse à l'item [Rocher, dans ce numéro, p. 37] – permet d'assurer la comparabilité entre les enquêtes successives et de mesurer l'évolution dans le temps de la distribution des niveaux de compétence des élèves.

Évaluations orientées élèves *vs* populations

Les évaluations standardisées des acquis des élèves offrent des perspectives diverses, selon que l'on s'intéresse aux élèves pris individuellement comme sujets de leurs apprentissages ou que l'on s'intéresse à eux comme éléments d'une population sur laquelle on cherchera à recueillir des informations destinées à éclairer le fonctionnement du système éducatif. Il est ainsi très important de clarifier, dès la phase initiale de la construction d'un dispositif, l'usage qui sera fait des données à recueillir. Au plan conceptuel, on peut distinguer les dispositifs « d'évaluation diagnostique » et les dispositifs « d'évaluation bilan » qui constituent deux types d'évaluations qui ne se substituent pas l'un à l'autre. Ils diffèrent dans leurs objectifs, dans les modalités de mise en œuvre, dans l'exploitation et l'utilisation des résultats.

Axées sur les élèves, outils professionnels pour les enseignants, les **évaluations diagnostiques** permettent d'établir un diagnostic individuel – évaluer les points forts et les points faibles (freins aux apprentissages), d'aider l'enseignant à définir les actions pédagogiques adaptées à la situation de chacun et à réguler la programmation des apprentissages. Au niveau local, utilisées par les enseignants dans la gestion pédagogique de leur classe, les évaluations diagnostiques, descripteurs des

réussites et échecs de chaque élève, aides à la constitution de groupes de besoin, supports pour la réflexion pédagogique, doivent permettre de cerner individuellement les compétences et les difficultés de chaque élève et d'orienter le travail de chaque élève et de la classe en fonction des résultats. Elles doivent aussi permettre de dégager des priorités de formation continue « de proximité ». En raison de leurs objectifs, de leur conception et de leur renouvellement annuel, ce type d'évaluations n'est pas adapté à la comparaison dans le temps. Toutefois, si les conditions de passation et de correction sont respectées, leurs résultats peuvent être comparés dans l'espace, entre les différentes classes d'une même école ou entre les différents collèges d'un département.

Axées sur des populations, les évaluations-bilans, (comme Cedre, PISA, les indicateurs de la LOLF) sont des outils pour le pilotage d'ensemble du système éducatif. Leur méthodologie de construction s'appuie sur les méthodes de la mesure en éducation et les modèles psychométriques [Laveault et Grégoire, 2002]. Elles concernent de larges échantillons représentatifs d'établissements, de classes et d'élèves. Organisées, le plus souvent, en fin de cycles, elles révèlent, en référence aux objectifs de la politique éducative, les objectifs atteints et ceux qui ne le sont pas. Ces évaluations doivent permettre d'agir au niveau national sur les programmes des disciplines, sur les organisations des enseignements, sur les contextes de l'enseignement, sur des populations caractérisées. Sous certaines conditions méthodologiques, leurs résultats peuvent être comparés dans le temps.

On notera que, depuis les années 1990, le Ministère a souvent entretenu la **confusion entre ces deux types d'évaluations**. Comme le pointent les inspecteurs généraux dans un rapport consacré à l'évaluation des acquis des élèves [IGEN-IGAENR, 2005, p. 15], *« les évaluations CE2/sixième ont eu, dès le début, une fonction ambiguë »*. Ainsi, au niveau national, la DEPP organise entre 1989 et 2007, sur des échantillons représentatifs d'élèves, une remontée des résultats sur les évaluations « de masse » en CE2 et en sixième. Ces résultats nationaux sont présentés comme des repères destinés à aider les enseignants à faire une analyse individuelle des freins rencontrés par leurs élèves dans les apprentissages. S'ils ne peuvent – au sens psychométrique [Dickes, Tournois *et alii,* 1994] – être comparés d'une année à l'autre, ils ont pu, ici ou là, être utilisés comme des indicateurs d'évolution des acquis des élèves, voire du système éducatif. La DEPP juge ainsi parfois nécessaire de rappeler que *« les évaluations nationales CE2 et sixième, tout comme celles d'entrée en seconde de lycée, n'ont de valeur qu'annuelle puisque les supports des évaluations et les objectifs évalués diffèrent chaque année. Aussi, ces résultats ne peuvent-ils en aucun cas être utilisés à des fins de comparaisons d'une année sur l'autre et détournés de leur objet pédagogique »*[4].
Cette confusion a encore été accentuée entre 2009 et 2012 avec la mise en place des évaluations « de masse » en fin de CE1 et en milieu de CM2, évaluations dont les objectifs ont été présentés de façon quelque peu « flottante », hésitant entre le diagnostic et le bilan pour finalement leur assigner ces deux objectifs simultanés.

4. Circulaire n° 2000-091 du 23 juin 2000, *Bulletin officiel,* n° 25 du 29 juin 2000.

Décrivons maintenant de manière plus détaillée la façon dont le Ministère a utilisé et déployé ces deux grands types d'évaluation au cours des trente dernières années. Nous retracerons tout d'abord l'histoire des grandes évaluations diagnostiques « de masse » qui, du fait de leur nature exhaustive, ont pendant une vingtaine d'années été, auprès des enseignants, la face visible du travail de la DEPP en matière d'évaluation. Nous montrerons ensuite pourquoi et comment s'est peu à peu installé dans le paysage un ensemble d'évaluations bilans standardisées sur échantillons dont l'objectif, outre celui de faire un état des lieux des acquis des élèves, est de donner à voir les évolutions de ces acquis, en permettant des comparaisons temporelles et internationales. Enfin, nous envisagerons ce que pourrait être un paysage renouvelé des évaluations au MENESR.

LES ÉVALUATIONS DIAGNOSTIQUES « DE MASSE »

Les évaluations en début de CE2, de sixième et de seconde

Les premières évaluations nationales « de masse » sont mises en place pour accompagner la loi d'orientation sur l'éducation dite « loi Jospin », du 10 juillet 1989 ; à la rentrée 1989 au début du cycle des approfondissements (CE2) et du collège (sixième), en français et en mathématiques. La DEPP pilote la conception et la mise en œuvre de ces évaluations dont les protocoles sont élaborés avec les corps d'inspection, des enseignants et chefs d'établissement, des représentants des directions pédagogiques, des chercheurs et des techniciens. Elles sont alors accompagnées d'un important effort de formation piloté par les directions pédagogiques du Ministère : formation de 400 formateurs de formateurs, conception et large diffusion de documents de « remédiation ». De plus, et c'est une nouveauté dans ce Ministère, la formation à l'utilisation de ces outils est rendue obligatoire pour tous les instituteurs de CE2 et les professeurs de français et de mathématiques de sixième.
Des outils informatisés de saisie et d'exploitation des résultats sont progressivement mis à disposition des enseignants (Casimir, puis J'ADE) ainsi que des statistiques détaillées, des dossiers de synthèse et de commentaires approfondis, d'abord sur support papier, jusqu'en 2002 [Brézillon, Chollet-Remvikos *et alii,* 2003], puis en ligne jusqu'en 2007. Pour Le Guen, responsable du département de l'évaluation des élèves et des étudiants au sein de la DEPP, *« cinq types d'actions relèvent de cette évaluation-diagnostique »* [Le Guen, 1991]. Si les trois premières actions sont clairement orientées élèves et enseignants : aider à mieux connaître les lacunes individuelles des élèves ; fournir une méthodologie et des outils d'évaluation pour les apprentissages de base ; contribuer à une meilleure efficacité en étant un outil de dialogue avec les familles, les deux autres sont « axées populations » : fournir des indicateurs pour le pilotage du système ; rendre compte au niveau national de l'efficacité de l'école.

Huit ans après leur mise en place, analysant l'usage qui en est fait, Levasseur [1996] leur attribue cinq fonctions : outil professionnel, outil de dialogue avec les parents, outil d'adaptation de l'enseignement, outil de formation des enseignants, outil de régulation du système éducatif. Parallèlement, la DEPP propose aux enseignants, de la maternelle au lycée, dans toutes les disciplines et durant toute l'année scolaire,

une « banque d'outils d'aide à l'évaluation ». Son objectif est de donner aux enseignants des outils diversifiés pour analyser les compétences des élèves et de *« leur permettre de faire évoluer les progressions pédagogiques en fonction des besoins objectivement repérés chez les élèves de la classe »*. D'abord sur support papier, cette banque est informatisée et consultable en ligne dès 2002. Elle n'est cependant pas actualisée depuis 2006. Le coût d'élaboration d'une telle banque est en effet très élevé, surtout si l'on souhaite accompagner les outils de résultats statistiques fiables permettant de donner des repères aux utilisateurs.

À la rentrée 1992, de nouvelles évaluations diagnostiques sont proposées au début du lycée, dans quatre disciplines : français, mathématiques, histoire-géographie, première langue vivante pour les secondes générales et technologiques ; français, mathématiques, sciences et techniques industrielles, économie-gestion pour les secondes professionnelles. L'évaluation de tous les élèves à l'entrée en seconde, évaluation de compétences à visée diagnostique, est destinée à faciliter la mise en œuvre des modules et de l'aide individualisée (à partir de 1999) afin de répondre au mieux aux besoins des élèves dans leur diversité. Ayant perduré pendant presque dix années, ces évaluations en classe de seconde ne seront pas reconduites à la rentrée 2002, compte tenu de leur faible usage par les enseignants. En effet, en 2001, Claude Pair, ancien recteur d'académie, examinant à la demande du Haut Conseil de l'évaluation de l'école (HCéé) les « forces et faiblesses de l'évaluation du système éducatif » écrit :
« Il faut distinguer les niveaux CE2 et sixième de celui de seconde. Dans le premier cas, l'opération est acceptée et effectuée quasiment partout ; les résultats sont restitués aux parents. Comme cela nous a été dit, "elle est entrée dans le paysage" [...] *Mais en seconde la situation semble beaucoup moins favorable : d'après les avis recueillis, l'évaluation est loin d'être effectuée partout et elle est exploitée moins souvent encore, les résultats n'étant donc guère communiqués aux collèges dont viennent les élèves.* [...] *En outre, les enseignants de lycée sont peut-être moins réceptifs que ceux de l'école ou du collège au questionnement pédagogique créé par l'hétérogénéité des acquis des élèves. »* [Pair, 2001]
Dans un contexte, déjà, de ressources contraintes, la DEPP doit faire des choix et la mise en place, dès le début des années 2000 (suite à l'avis n° 2 du HCéé, juin 2001), des évaluations bilans, plus tard appelées Cedre, l'engagement dans les évaluations triennales internationales PISA et la construction d'indicateurs nouveaux dans le cadre de la LOLF (voir *infra*), nécessitent un redéploiement de ressources humaines et financières qui décident du sort des évaluations à l'entrée en seconde, jusqu'alors mises en œuvre par la DEPP. Une incursion en cinquième est cependant tentée à la rentrée scolaire 2002. Elle ne sera pas poursuivie.

Alors que le Ministère, souhaitant mettre davantage l'accent sur la prévention de l'illettrisme, prépare dès 2004 de nouvelles évaluations en début de CE1, l'idée s'impose progressivement d'arrêter les évaluations en CE2 et en sixième. Cette décision sera actée en janvier 2007, la circulaire de rentrée[5] indiquant que les évaluations sont supprimées en CE2 et reconduites pour la dernière fois en sixième à la rentrée 2007. Elles sont cependant reconduites en sixième à la rentrée 2008, pour satisfaire

5. Circulaire n° 2007-011 du 9 janvier 2007, *Bulletin officiel*, n° 3 du 18 janvier 2007.

la demande de l'encadrement de terrain qui souhaite leur maintien pour le pilotage pédagogique local. Toutefois, les groupes de concepteurs élaborant les protocoles d'évaluation (cahiers des élèves, consignes de passation et de correction pour les enseignants) ne sont pas reconduits et les protocoles utilisés restent identiques de 2005 à 2008, sans que les résultats relevés sur des échantillons représentatifs varient au cours des trois ou quatre dernières années de leur utilisation.

Pour répondre à la demande de repérage des élèves rencontrant le plus de difficultés, sont donc expérimentées[6] en 2004 et en 2005, et généralisées dès la rentrée 2006, de nouvelles évaluations en début de CE1, en lien avec les programmes personnalisés de réussite éducative (PPRE). Elles s'inscrivent dans le dispositif de prévention de l'illettrisme et en cohérence avec les nouveaux programmes de 2002. Elles seront suivies du même type d'évaluations en début de CM2, sous statut expérimental à la rentrée 2007. Ces évaluations comportent un premier filtre destiné à repérer les élèves (estimés en moyenne à 20 % de la population) nécessitant un regard plus fin sur leurs besoins. Une fois ces élèves repérés, il est demandé aux enseignants de leur administrer un deuxième livret d'évaluation permettant de cerner avec plus de précision leurs difficultés et de les orienter vers les remédiations les plus adaptées, à l'intérieur ou à l'extérieur de la classe. Le cahier des charges de l'évaluation en CE1 précise qu'elle ne fera pas l'objet de statistiques au niveau national et qu'il n'y aura donc pas de remontées vers le Ministère.
Cependant, dès 2006, dans une note interne du 16 octobre, le directeur de cabinet demande de produire pour le 20 décembre *« une synthèse des résultats de l'évaluation nationale de CE1 »* et, pour l'année scolaire suivante, la préparation d'un dispositif fonctionnel de remontée exhaustive des résultats des évaluations aux différents niveaux territoriaux pour *« permettre aux différents échelons de pilotage pédagogique du premier degré d'appréhender finement la situation réelle de nos élèves en matière de maîtrise des compétences du socle commun »*. Il est également demandé de prévoir, sur échantillon représentatif, une synthèse au niveau national pour l'information du ministre et de l'administration centrale.
Ce mélange des genres est mal reçu au niveau local comme en témoigne le rapport établi en mars 2007 par Claus et Mégard [IGEN, 2007, p. 9] sur le suivi de cette évaluation, qui en souligne les ambiguïtés en posant la question de la capacité d'une évaluation diagnostique à devenir un outil de pilotage : *« De manière générale, une évaluation dont le seul objectif est d'aider les enseignants à gérer l'hétérogénéité des élèves peut s'accommoder de différences lors de la passation des épreuves (consignes répétées ou expliquées, temps supplémentaire accordé, etc.). Une évaluation qui aboutit à des moyennes et des comparaisons pour piloter à une échelle supérieure à celle de l'école ne peut se satisfaire de telles approximations. Le changement de cap imposé en 2006 a eu pour conséquence la construction d'indicateurs peu fiables et donc peu exploitables. »*

Les inspecteurs recommandent fortement une clarification des finalités de l'évaluation des élèves au début du CE1, d'autant plus, ajoutent-ils, que cette évaluation sera accompagnée d'une évaluation de même nature au début du CM2. Ils ajoutent :

6. Circulaire n° 2004-015 du 25 janvier 2004, *Bulletin officiel,* n° 6 du 5 février 2004 et circulaire n° 2004-108 du 05 juillet 2004, parue au *Bulletin officiel* n° 28 du 15 juillet 2004.

« En tout état de cause, le Ministère doit impérativement faire preuve de constance. Une décision arrêtée et communiquée aux académies et aux écoles ne peut sans risque de rupture de confiance être abrogée sans raison impérative. Si les résultats d'une évaluation doivent être communiqués, publiés et agglomérés, il convient d'en informer les écoles avant la passation. »

À la rentrée 2007, les élèves de CE1 passent des protocoles modifiés et ceux de CM2 expérimentent un nouveau protocole. La circulaire de rentrée 2007 (voir note 4) précise la vocation essentiellement analytique de ces deux évaluations qui *« visent à repérer et analyser les difficultés et les freins que rencontrent certains élèves de CE1 et de CM2 dans leurs apprentissages dans les domaines de la lecture, de l'écriture et des mathématiques pour acquérir les compétences du socle attendues en fin des paliers 1 et 2* [...]*. Ces évaluations s'insèrent tout naturellement dans le dispositif mis en place localement pour déterminer si un élève doit bénéficier d'un programme personnalisé de réussite éducative (PPRE) et pour envisager les modalités de celui-ci. »*
Bien que soit annoncé le calcul de scores nationaux à partir d'un échantillon représentatif[7], aucun bilan de ces évaluations ne sera finalement réalisé. Elles disparaîtront sans bruit du paysage dès la fin de l'automne 2007 avec la préparation de nouvelles évaluations nationales obligatoires et exhaustives en CE1 et en CM2, qui auront cours de 2009 à 2012 et dont le pilotage sera confié à la DGESCO. Ce transfert d'un pilotage qui n'entre pas dans les missions habituellement dévolues à cette direction peut être compris comme la volonté de donner à cette dernière les moyens de contrôle dont elle souhaite disposer pour s'assurer que les réformes pédagogiques qu'elle impulse sont bien mises en œuvre au niveau local. Les nouveaux programmes de l'école primaire applicables à partir de la rentrée 2008 en sont notamment l'enjeu et, ces nouvelles évaluations, l'instrument. C'est d'ailleurs ce que notent les inspecteurs généraux Claus et Roze dans leur troisième note de synthèse à destination du ministre : « *Ces évaluations révèlent aussi l'écart, qui peut être important, entre ce qui est enseigné et ce qui devrait l'être. En ce sens les évaluations nationales sont un puissant levier pour une mise en œuvre complète des nouveaux programmes.* » [IGEN-IGAENR, 2009, p. 10].

Les évaluations en CE1 et en CM2

Le 5 juillet 2007, la lettre de mission du président de la République à son ministre de l'éducation[8] souhaite l'organisation d'*« une évaluation systématique de tous les élèves tous les ans, afin de repérer immédiatement les élèves en difficulté et de pouvoir les aider ; une évaluation régulière des enseignants sur la base des progrès et des résultats de leurs élèves »*. Ces nouvelles évaluations en CE1-CM2 sont en totale rupture avec les évaluations précédentes à ces niveaux scolaires. Situées en fin d'année scolaire pour le CE1, en janvier pour le CM2, elles sont présentées au départ comme des bilans devant également servir à évaluer les enseignants. La publication des résultats sur Internet, école par école, est même annoncée. Sollicitée par le Cabinet, la DEPP fait part de ses réserves et soulève un certain nombre d'interrogations, notamment sur les usages de l'évaluation, la comparabilité temporelle et la prise

7. Circulaire n° 2007-140 du 23 août 2007, *Bulletin officiel,* n° 30 du 30 août 2007.
8. http://discours.vie-publique.fr/notices/077002457.html

en compte du contexte social de l'école. Devant la levée de boucliers suscitée tant chez les enseignants que chez les parents d'élèves, l'idée de la publication des résultats école par école fait long feu. Toutefois, subsiste chez les enseignants une défiance quant à la vraie nature de ces évaluations, présentées à la fois comme bilan et comme diagnostic, en insistant tantôt sur un aspect, tantôt sur l'autre, et pouvant servir à contrôler leur valeur professionnelle. Cet usage possible de l'évaluation est ressenti comme d'autant plus injuste qu'il ne repose pas sur les progrès réalisés par les élèves, mais uniquement sur leur niveau à un instant donné, sans prendre en considération leur niveau scolaire à leur arrivée dans la classe ni leurs différences socioéconomiques. Cette confusion amène une résistance jamais encore vue chez les enseignants du primaire contre des évaluations malgré une prime de 400 € instituée pour les enseignants des niveaux concernés.

Dans un document d'orientation de novembre 2007[9], ces évaluations sont affichées comme une innovation : *« Il faut également se donner les moyens de connaître et de faire connaître quels sont les acquis des écoliers français à des moments clefs de leur scolarité, notamment par rapport aux pays comparables. C'est pourquoi seront créées* ***deux évaluations nationales témoins qui serviront à mesurer les acquis des élèves au CE1 et au CM2****.* [...] *Leurs constats seront rendus publics et permettront d'apprécier l'évolution de la réussite du système éducatif. »* Cette présentation ne fait aucune référence à l'existence d'évaluations spécifiquement construites pour assurer des comparaisons temporelles ou internationales, telles que les enquêtes nationales du cycle Cedre (voir *infra*), basées sur les programmes et mises en place depuis 2003, et l'enquête internationale PIRLS[10] qui existe depuis 2001.

En complément, la circulaire de rentrée 2008[11] précise que ces nouveaux protocoles d'évaluation *« permettent de dresser un bilan des acquis des élèves en CE1 et en CM2, premiers paliers du socle commun.* [...] *Les résultats scolaires des élèves seront un élément essentiel du pilotage. »* Interrogé sur ces évaluations par le site « Le café pédagogique »[12], le chef du bureau des écoles de la DGESCO déclare : *« Ce sont des évaluations organisées autour des programmes. C'est la grande différence avec les évaluations précédentes. La référence c'est le programme. Par conséquent on a affaire à une évaluation bilan de ce que les élèves ont acquis. En même temps, quand on regarde ce qui n'a pas été réussi on est sur le versant du repérage voire du diagnostic. »* On le voit, pour ces évaluations, la double assignation de bilan et de diagnostic est clairement assumée. Elle sera critiquée tant par les organisations syndicales que par le Haut Conseil de l'Éducation [voir *infra*, HCÉ, 2011] ou encore dans le rapport du groupe UMP de l'Assemblée nationale [Breton et Marc, 2009] qui souligne dans sa conclusion : *«* ***Une clarification des objectifs poursuivis est nécessaire****. En effet, les évaluations mises en place pour les élèves de CM2 sont à mi-chemin entre l'évaluation-bilan et l'évaluation-diagnostic appelant un dispositif de remédiation. Cette question ne semble pas complètement tranchée et une clarification par le ministère de l'Éducation nationale, en concertation avec l'ensemble des acteurs concernés, serait utile sinon indispensable. »*

9. Document d'orientation. Propositions du ministre de l'Éducation nationale, soumises à discussion, pour définir un nouvel horizon pour l'école primaire : http://media.education.gouv.fr/file/40/9/20409.pdf.
10. Le programme international sur la recherche en lecture scolaire, mené par l'IEA (*International Association for the Evaluation of Educational Achievement*) évalue les compétences en lecture des élèves en quatrième année de scolarité obligatoire (CM1 pour la France).
11. Circulaire n° 2008-042 du 4 avril 2008, *Bulletin officiel*, n° 15 du 10 avril 2008.
12. http://www.cafepedagogique.net/lesdossiers/pages/2009/evacm2ministere.aspx

En termes de fiabilité, une étude interne, réalisée par la DEPP lors de la première évaluation de janvier 2009, fait apparaître des distorsions dans les résultats selon que les écoles ont ou non été suivies par les inspecteurs du contrôle qualité, ainsi qu'en fonction des secteurs de scolarisation[13]. Ainsi, on observe une surestimation des élèves par leurs enseignants, et ce de façon plus particulièrement marquée dans le secteur privé, en l'absence de contrôle des procédures de passation et de correction. Dès la deuxième année d'utilisation, les limites de l'exercice, en termes de comparabilité, sont atteintes : les résultats des élèves de CM2 affichent une forte baisse en mathématiques. Cette baisse est en fait due à la plus grande difficulté du protocole élaboré pour cette deuxième itération, mais elle est interprétée comme une perte de compétence moyenne des élèves de CM2. La DEPP, sollicitée pour donner une mesure de l'impact de cette absence de contrôle de l'élaboration des protocoles, utilise une procédure d'*equating* (mise à niveau des métriques) pour permettre la comparabilité entre les deux années [Rocher, 2012]. Mais la suspicion à l'égard de ces évaluations est telle que l'ajustement des résultats de cette deuxième évaluation est dénoncé par beaucoup comme un « bidouillage » destiné à masquer l'impéritie du Ministère.
Celles-ci seront menées durant quatre années (de janvier 2009 à juin 2012) et ne seront pas reconduites après le changement de gouvernement de mai 2012. Le ministre Vincent Peillon indique qu'elles pourront être utilisées localement mais décide l'arrêt des « remontées » des informations à l'administration centrale, puisque « *les outils qui sont actuellement utilisés ne permettent pas une évaluation scientifiquement incontestable du système éducatif national* »[14]. Leur utilisation sera rendue facultative en 2013[15], puis abandonnée en 2014 (voir *infra*).

LES ÉVALUATIONS « BILANS » STANDARDISÉES À GRANDE ÉCHELLE

Présentes au Ministère dès la fin des années 1970, les enquêtes portant sur l'évaluation des compétences, appelées aussi évaluations standardisées à grande échelle, se sont multipliées depuis le début des années 2000 (voir plus loin la présentation de ces dispositifs). Leur objectif principal est de rendre compte de résultats au-delà du niveau individuel, en l'occurrence au niveau national et international. Le fait de leur assigner un objectif de représentativité implique des contraintes spécifiques dans leur élaboration.
Ce type d'évaluations occupe une place importante dans le débat sur l'éducation, notamment *via* la médiatisation – voire l'instrumentalisation politique – de leurs résultats [Mons, 2008]. La mise en œuvre de politiques éducatives se réfère aujourd'hui systématiquement à ces évaluations, en particulier aux évaluations internationales qui, derrière la diffusion de palmarès globalisants, fournissent un éclairage important sur les forces et les faiblesses des systèmes éducatifs [voir par exemple : Mons, 2007 ; Rocher, 2008b ; Baudelot et Establet, 2009 ; Rocher et Le Donné, 2012a].

13. Rapport conjoint IGEN-IGAENR : deuxième note de synthèse sur l'évaluation des élèves de CM2, n° 2009-028 du 31 mars 2009 et note interne DEPP non publiée.
14. Communiqué de presse du 21 mai 2012.
15. Circulaire n° 2013-060 du 10 avril 2013, *Bulletin officiel*, n° 15 du 11 avril 2013.

La DEPP, consciente de l'importance politique et scientifique de ces évaluations, obtient dès la décennie 1990 d'en être l'opérateur en France pour le compte des institutions qui les organisent (OCDE, IEA, Union européenne)[16].

Ces évaluations-bilans vouent une attention particulière aux comparaisons temporelles, afin de pouvoir juger des progrès réalisés par les systèmes éducatifs et de les rapprocher de caractéristiques structurelles, sociales, éducatives, etc. La question de la mesure de l'évolution du niveau des élèves dans le temps est donc centrale. Pourtant, dans l'histoire de l'évaluation et des tests, les études qui visent à comparer les compétences des sujets à différentes époques sont relativement rares.

Dans le domaine de l'intelligence cependant, des études comparatives assez anciennes ont révélé le fameux « effet Flynn », c'est-à-dire l'élévation des performances à des tests d'intelligence [Flieller, 2001]. Comme le notent Flieller, Manciaux et Kop [1995], ces enquêtes méritent qu'on leur prête attention pour deux raisons majeures. La première est d'ordre théorique : il s'agit de se prononcer sur le caractère absolu de certaines lois psychologiques, en appréciant leur permanence à travers différentes périodes de l'histoire. La seconde est d'ordre « pratique » : ces enquêtes permettent de répondre de manière objective à la demande sociale récurrente qui concerne l'évolution du niveau d'intelligence, de connaissances ou de compétences de la population. Pour répondre à cette nécessaire objectivation, ces enquêtes s'appuient sur des principes méthodologiques bien établis [Rocher, dans ce numéro, p. 37].

En France, l'intérêt pour ces évaluations à visée comparative trouve son origine dans le débat sur la baisse supposée du niveau scolaire des élèves, débat qui semble être particulièrement vif en France à la fin des années 1980. En 1992, dans un rapport au ministre de l'Éducation nationale, Claude Thélot, alors directeur de la DEPP, s'interroge sur les raisons de ce sentiment d'inquiétude et dégage trois pistes d'explication [Thélot, 1992]. Premièrement, après une période de massification du système éducatif, l'attention est portée sur la qualité et donc le niveau de compétence des élèves. Un système « de masse » peut-il être performant ? Deuxièmement, dans la lignée du rapport alarmiste américain *A Nation at Risk* de 1983 [National Commission on Excellence in Education, 1983], avec le renforcement de la compétition économique au niveau international, l'élévation du niveau des élèves apparaît comme un levier indispensable. Enfin, le sentiment de déclin du système éducatif porterait moins sur les mathématiques et les sciences, disciplines valorisées socialement, que sur les lettres et les humanités. Selon Thélot [1992], ce distinguo traduirait une appréhension profonde quant à l'avenir du pays, de sa langue et de son identité.

Néanmoins, Thélot [1992], tout comme Baudelot et Establet [1989], dans leur ouvrage *Le niveau monte*, soulignent le manque de mesures directes et objectives de l'évolution des acquis des élèves, alors même que cette question fait

16. C'est également dans cette perspective qu'à l'initiative de la DEPP, se met en place en 1997 le « Consortium université de Bourgogne », pour répondre à l'appel d'offres lancé par l'OCDE pour la mise au point et l'organisation de l'enquête PISA 2000 [Bottani et Vrignaud, 2005, p. 159].

l'objet d'une forte demande politique et sociale[17]. À l'époque, les seules données disponibles permettant une comparaison temporelle moins subjective sont celles issues des tests passés pendant les « trois jours » organisés par le ministère de la Défense [Baudelot et Establet, 1988]. Même si ces évaluations ne concernaient que les garçons, elles donnaient une bonne approximation de l'évolution du « niveau », ne serait-ce qu'en raison du nombre important de ceux qui les passaient et de leur proximité avec une génération entière de garçons.

Forte de ce constat, dans les années 1990, la DEPP conduit plusieurs études visant à mesurer l'évolution des acquis des élèves, comme la comparaison des compétences en français et en calcul des élèves des années 1920 à celles des élèves de 1995 qui avait clairement pour objectif de répondre aux tenants de la faillite du système éducatif, souvent nostalgiques d'un modèle scolaire révolu [Dejonghe, Levasseur *et alii*, 1996 ; Pons, 1996]. D'autres enquêtes ont concerné les élèves de troisième [Dessus, Jouvanceau, Murat, 1996] ou les élèves les plus performants scolairement [Peretti, Petrone, Thélot, 1996].

Cependant, les méthodes psychométriques permettant de construire des comparaisons diachroniques fiables apparaissent alors insuffisamment connues à la DEPP et plus généralement, à l'Insee ou dans l'enseignement des statistiques en France. Pourtant, des travaux reposant sur une méthodologie psychométrique adaptée à la comparaison diachronique existent à cette époque, dans le champ de la psychologie différentielle, comme en témoignent les enquêtes sur le niveau intellectuel des jeunes enfants [Flieller, Santigny, Schaeffer, 1986 ; Flieller, Manciaux, Kop, 1995].
Les premières enquêtes comparatives menées par la DEPP montrent, quant à elles, quelques faiblesses méthodologiques. Le recours à l'expertise de l'Inetop (Institut national d'étude du travail et d'orientation professionnelle) sur la comparabilité d'évaluations ayant eu lieu en sixième [Bonora et Vrignaud, 1996] et en troisième [Bonora et Vrignaud, 1997] permet de pointer les problèmes de comparabilité. Ces rapports sont aussi l'occasion d'introduire à la DEPP une connaissance des modèles de réponse à l'item (MRI), suite aux polémiques autour de l'enquête IALS[18] [Blum et Guérin-Pace, 2000 ; Murat et Rocher, dans ce numéro, p. 83]. En 1997, la reprise de l'évaluation LEC (lire, écrire, compter) de 1987 fait alors l'objet d'analyses psychométriques appropriées.
Comme on l'a vu plus haut, la France a pourtant une longue expérience des évaluations standardisées des élèves, à travers la mise en place des évaluations nationales diagnostiques de CE2 et de sixième. Malheureusement, aucun ajustement de la difficulté des épreuves n'a été entrepris afin de distinguer ce qui relève de la difficulté des épreuves de ce qui relève du niveau des élèves. En effet, nous l'avons rappelé, l'objectif premier de ces évaluations n'était pas de rendre compte de l'évolution du niveau des élèves dans le temps, mais de servir d'outils de repérage des difficultés pour les enseignants.

17. Paradoxalement, les évaluations des élèves sont très présentes dans le système scolaire français, à travers les contrôles continus fréquents conduits par les enseignants. Des études docimologiques, menées depuis près d'un siècle, notamment dans le cadre des travaux de la commission Carnegie sur le baccalauréat en 1936, montrent pourtant que le jugement des élèves par les enseignants est en partie empreint de subjectivité et peut dépendre de facteurs étrangers au niveau de compétence des élèves [voir par exemple : Piéron, 1963]. La notation des élèves est ainsi susceptible de varier sensiblement selon les caractéristiques des enseignants, des contextes scolaires, ainsi que des élèves eux-mêmes. Ces observations se retrouvent également aujourd'hui dans l'analyse des attestations de maîtrise des compétences du « socle commun de connaissances et de compétences » [Daussin, Rocher, Trosseille, 2010].
18. *International Adult Literacy Survey.*

L'essor des évaluations à visée diachronique

En 2001, l'avis du HCéé n° 2 [HCéé, 2001] pointe à son tour le manque d'informations objectives sur ce sujet et recommande la mise en place d'un dispositif *ad hoc* de suivi de l'évolution des acquis des élèves dans le temps [Salines et Vrignaud, 2001], comme cela existe dans d'autres pays, par exemple aux États-Unis où le dispositif NAEP (*National Assessment for Educational Progress*) fournit des séries de résultats comparables depuis la fin des années 1960 [Zwick, 1992 ; Jones et Olkin, 2004].
À la suite des recommandations du rapport de Salines et Vrignaud [2001], la DEPP donne naissance en 2003 au cycle des évaluations Cedre qui évalue les acquis des élèves de CM2 et de troisième, au regard de ce qui est attendu par les programmes scolaires. Chaque année, le domaine évalué est différent et à partir de 2009, des comparaisons temporelles concernent la maîtrise de la langue française [Colmant, Daussin, Bessonneau, 2011 ; Bourny, Bessonneau *et alii,* 2010), les langues étrangères [Bessonneau, Beuzon, Boucé *et alii.* 2012 ; Bessonneau, Beuzon, Daussin *et alii*, 2012], l'histoire-géographie [Garcia et Pastor, 2013 ; Garcia et Krop, 2013] et les sciences [Andreu, Étève, Garcia, 2014 ; Bret, Garcia, Roussel, 2014].
Depuis, d'autres dispositifs d'évaluations construits pour permettre des comparaisons diachroniques se sont développés en France ▸ **Tableau 1 p. 28-29**.

Plusieurs phénomènes relativement récents expliquent l'essor important de ces évaluations et leur multiplicité actuelle. Tout d'abord, au-delà de la demande du HCéé [Haut Conseil de l'évaluation de l'école, 2003], le souci de construire des indicateurs de suivi, pour le pilotage du système, est devenu de plus en plus prégnant, notamment dans le cadre de la LOLF, qui implique la construction d'indicateurs annuels de résultats [Rocher, 2008c]. Parallèlement, les évaluations internationales, telles que PISA [OCDE, 2013], PIRLS [Mullis, Martin *et alii,* 2012 ; Colmant et Le Cam, 2012] et TIMSS[19] [Mullis, Martin *et alii,* 2012], ont largement contribué à l'importance accordée aux comparaisons temporelles, en organisant les enquêtes de manière cyclique (voir tableau 1).

Alors que ces évaluations développées récemment ont été conçues de manière à assurer la comparaison diachronique, certains dispositifs proposent des comparaisons *ex post* de plus long terme, comme l'enquête SPEC6 [Rocher et Le Donné, 2012b] et l'enquête LEC [Rocher, 2008a]. Pour des synthèses concernant l'évolution des acquis des élèves, on se reportera à Rocher [2010] et à Daussin, Keskpaik, Rocher [2011].
Les évaluations réalisées dans le cadre des suivis longitudinaux de la DEPP (panels) permettent aussi de procéder à des comparaisons diachroniques, bien que ce ne soit pas leur objet premier qui est de décrire et d'expliquer les carrières et performances scolaires des élèves en rapprochant parcours scolaires, résultats aux évaluations et éléments de contexte. C'est le cas de la comparaison des acquis des élèves en début de CP de 1997 à 2011 [Le Cam, Rocher, Verlet, 2013].

Enfin, certains dispositifs concernent des sujets plus âgés. Ainsi, l'évaluation de la lecture, conçue par la DEPP et passée par tous les jeunes d'environ 17 ans lors de la JDC (Journée défense et citoyenneté), produit des indicateurs annuels de

19. *Trends in International Mathematics and Science Study.*

Tableau 1 Dispositifs d'évaluations standardisées en France permettant des comparaisons diachroniques

Test	Nom	Années
Évaluations nationales		
Cedre	Cycle des évaluations disciplinaires réalisées sur échantillons	Annuel, depuis 2003
LOLF	Évaluations pour les indicateurs de la LOLF	Annuel, depuis 2007
LEC	Lire, écrire, compter	1987, 1997, 2007
SPEC6	Étude spécifique des difficultés de lecture	1997, 2007
Panel CP	Évaluations standardisées des acquis des élèves du panel CP	1997, 2011
JDC	Journée défense et citoyenneté	Annuel, depuis 1998
IVQ	Information et vie quotidienne	2004, 2011
Évaluations internationales		
ESLC	*European Survey on Language Competences*	2011
PIRLS	*Progress in International Reading Literacy Study*	2001, 2006, 2011, 2016
PISA	*Programme for International Student Assessment*	Tous les trois ans depuis 2000
TIMSS	*Trends in International Mathematics and Science Study*	1995, 2015
Piaac	*Programme for the International Assessment of Adult Competencies*	2012

suivi de performance [voir par exemple, Vourc'h, Rivière *et alii,* 2014]. L'enquête IVQ (Information et vie quotidienne) évalue quant à elle un large échantillon d'adultes, et a permis une comparaison entre 2004 et 2011 [Jonas, 2012].

PERSPECTIVES

Comme nous l'avons indiqué en fin de partie 1, l'arrêt porté aux évaluations de CE1 et de CM2 était motivé par la clarification des objectifs des évaluations et par le rétablissement de la confiance du monde enseignant. Dès le printemps 2012, alors que se met en place la concertation pour la « Refondation de l'École », le ministre rétablit la distinction entre évaluations diagnostiques, outils professionnels des enseignants dans le sens mentionné plus haut, et évaluations bilans standardisées dont l'objectif est l'évaluation des acquis des élèves, indicateurs participant à l'évaluation du système éducatif[20].

20. Communiqué de presse du 21 mai 2012 et discours de Vincent Peillon lors de la conférence de presse de rentrée du 29 août 2012.

Population	Domaines
CM2 et 3e	Maîtrise de la langue (MDL) en CM2, compétences générales (CG) en troisième ; langues vivantes ; attitudes à l'égard de la vie en société ; histoire, géographie et éducation civique ; sciences expérimentales ; mathématiques.
CE1, CM2 et 3e	Compétences de base en français et en mathématiques, compétences du socle commun.
CM2	Compréhension de l'écrit, orthographe, calcul.
Début sixième	Automatismes, lexique, compréhension.
Début CP	Pré-lecture, écriture, numération, compréhension orale.
Environ 17 ans	Compréhension de l'écrit, lexique, automatismes.
16-65 ans	Littératie, numératie.
Troisième	Anglais, espagnol (compréhension de l'écrit et de l'oral).
CM1	Compréhension de l'écrit.
15 ans révolus	Compréhension de l'écrit, culture mathématique, culture scientifique.
CM1, TS	Mathématiques et sciences physiques.
16-65 ans	Littératie et numératie.

Aujourd'hui, d'une façon générale, l'évaluation des acquis des élèves peut répondre à trois objectifs :
– fournir aux enseignants des outils afin d'enrichir leurs pratiques pédagogiques en évaluant mieux les acquis de leurs élèves ;
– disposer d'indicateurs permettant de mesurer, au niveau national, les performances de notre système (évolutions temporelles et comparaisons internationales) ;
– doter les « pilotes de proximité » (recteurs, DASEN, IEN) d'indicateurs leur permettant de mieux connaître les résultats des écoles et d'effectuer une vraie régulation.
Quelle que soit la façon de répondre à ces trois objectifs, il est essentiel de tirer des leçons du passé qui a vu des évaluations nationales prétendre remplir conjointement plusieurs fonctions. Il apparaît important de réaffirmer, à la suite du Haut Conseil de l'éducation [2011] dans son bilan des résultats de l'école, que *« il n'est pas de bonne méthode de confondre deux types d'évaluations : d'une part les évaluations dans la classe dont l'enseignant a régulièrement besoin pour adapter son enseignement en fonction des acquis de ses élèves, d'autre part une évaluation nationale destinée au pilotage du système éducatif »*.
Le premier objectif devrait pouvoir être réalisé au travers d'outils pédagogiques proposés au niveau national ou académique, du type des anciennes évaluations CE2 et sixième, possiblement au début des nouveaux cycles, soit début de CM1 et début de cinquième, si l'on souhaite des évaluations exhaustives, grâce à des banques d'outils si l'on veut que les enseignants disposent d'outils utilisables toute l'année.

Ce type d'évaluations serait légitimement à la charge des instances pédagogiques du Ministère ou des académies.
Pour remplir le deuxième objectif, le Ministère dispose des divers outils évoqués plus haut dans cet article (dispositif Cedre, LOLF, panels, enquêtes internationales). Un effort de rationalisation de la complémentarité de ces outils a été récemment entrepris. Le dispositif Cedre a vu son cycle se réduire à cinq ans et les évaluations pour les indicateurs de la LOLF sont désormais organisées selon une périodicité de trois ans (chaque année un palier est évalué en commençant par le CE1 en 2014) et ne portent plus que sur les compétences 1 et 3 du socle commun. En outre, l'engagement renouvelé de la DEPP dans sa participation aux évaluations internationales (notamment avec la reprise de l'évaluation TIMSS en 2015) témoigne de la conscience d'une nécessaire complémentarité des points de vue sur l'état et l'évolution du système éducatif. Ainsi, PISA [OCDE, 2013] a non seulement confirmé les inégalités socio-scolaires et leur accroissement établis dans Cedre mais a en outre révélé leur aggravation comparativement aux autres pays de l'OCDE.

Le troisième objectif est sans doute le plus ardu à atteindre si l'on souhaite éviter de mélanger les deux premiers objectifs en dotant les acteurs locaux d'outils de pilotage. La DEPP a engagé un partenariat avec quelques académies pour contribuer à l'élaboration d'outils d'évaluation des acquis des élèves ou des performances des établissements. Il s'agit de fournir aux cadres territoriaux des éléments leur permettant d'envisager des mesures de régulation. Les évaluations LOLF feront ainsi l'objet tous les trois ans (en début de sixième) d'échantillons académiques représentatifs permettant d'apprécier l'évolution des performances de chaque académie. Une réflexion est actuellement menée entre la DEPP et quelques responsables académiques pour trouver une solution qui permette l'exhaustivité de ces évaluations dans leur académie en surmontant d'importants problèmes conceptuels, techniques, voire politiques.

Par ailleurs, la DEPP expérimente depuis quelques années la possibilité d'une administration informatisée des évaluations, en parallèle avec des études sur la comparabilité des supports utilisés : version papier-crayon *vs* version sur écran[21]. Le développement de ces évaluations progresse mais de redoutables défis restent encore à relever, notamment ceux de la qualité des infrastructures informatiques, en particulier dans les écoles, de l'accès des établissements au haut débit, de la capacité des acteurs locaux à assurer la maintenance et le renouvellement de leur parc informatique, de l'accompagnement des établissements scolaires dans la mise en œuvre de ces évaluations, etc. Cette nouvelle génération d'évaluations devrait permettre de faciliter la mise en œuvre pratique d'évaluations standardisées, qu'elles soient à visée bilan ou diagnostique, sur échantillons ou exhaustives. En outre, la nature même des compétences qui pourront être évaluées au moyen du support numérique devra être envisagée de façon renouvelée et innovante.

21. Ces études sont indispensables pour assurer la comparabilité temporelle d'évaluations passées sur des supports de différentes natures. Ce type d'études est également à l'ordre du jour des enquêtes internationales. Ainsi, pour PISA 2015, l'enquête se déroulera entièrement sur ordinateurs dans la majorité des 70 pays participants.

BIBLIOGRAPHIE

Andreu S., Étève Y., Garcia E., 2014, « Cedre 2013 – Grande stabilité des acquis en sciences en fin d'école depuis 2007 », *Note d'information,* n° 27, MENESR-DEPP.

Baudelot C., Establet R., 2009, *L'élitisme républicain – L'école française à l'épreuve des comparaisons internationales*, Paris, Le Seuil.

Baudelot C., Establet R., 1989, *Le niveau monte – Réfutation d'une vieille idée concernant la prétendue décadence de nos écoles*, Paris, Le Seuil.

Baudelot C., Establet R., 1988, « Le niveau intellectuel des jeunes conscrits ne cesse de s'élever », *Économie et Statistique*, n° 207, Insee, p. 31-39.

Bessonneau P., Beuzon S., Boucé S., Daussin J.-M., Garcia E., Lévy M., Marchois C., Trosseille B., 2012, « L'évolution des compétences en langues des élèves en fin de collège de 2004 à 2010 », *Note d'information*, n° 12.05, MENJVA-DEPP.

Bessonneau P., Beuzon S., Daussin J.-M., Garcia E., Lévy M., Marchois C., Trosseille B., 2012, « L'évolution des compétences en langues des élèves en fin d'école de 2004 à 2010 », *Note d'information*, n° 12.04, MENJVA-DEPP.

Blum A., Guérin-Pace F., 2000, *Des lettres et des chiffres – Des tests d'intelligence à l'évaluation du « savoir lire », un siècle de polémiques*, Paris, Fayard.

Bonora D., Vrignaud P., 1997, *Analyse interne utilisant les modèles MRI (ou IRT) des épreuves de mathématiques dans les dispositifs troisième de 1984, 1990 et 1995 : difficulté des items et évolution des compétences*, Rapport de convention MENESR-DEP, CNAM-INETOP.

Bonora D., Vrignaud P., 1996, *Étude de l'évolution des connaissances des élèves en début de sixième, perspective psychométrique classique et perspective* MRI (ou IRT), Rapport de convention, MENESR-DEP, CNAM-INETOP.

Bottani N., Vrignaud P., 2005, *La France et les évaluations internationales*, Les rapports établis à la demande du Haut Conseil de l'évaluation de l'école, Rapport n° 16. www.ladocumentationfrancaise.fr/rapports-publics/054000359/

Bourny G., Bessonneau P., Daussin J.-M., Keskpaik S., 2010, « L'évolution des compétences générales des élèves en fin de collège de 2003 à 2009 », *Note d'information*, n° 10.22, MEN-DEPP.

Bret A., Garcia E., Roussel L., 2014, « Cedre 2013 – Sciences en fin de collège : stabilité des acquis depuis six ans », *Note d'information*, n° 28, MENESR-DEPP.

Breton X., Marc A., 2009, *Les évaluations dans l'enseignement primaire au service de la réussite scolaire – Les propositions du Groupe UMP*, Assemblée nationale, p. 8.

Brézillon G., Chollet-Remvikos P., Rebmeister B., Zelty C., 2003, « Évaluations CE2 - sixième - cinquième – Repères nationaux septembre 2002 », *Les Dossiers Enseignement scolaire*, n° 141, MJENR-DEP.

Colmant M., Daussin J.-M., Bessonneau P., 2011, « Compréhension de l'écrit en fin d'école – Évolution de 2003 à 2009 », *Note d'information*, n° 11.16, MEN-DEPP.

Colmant M., Le Cam M., 2012, « Pirls 2011 – Étude internationale sur la lecture des élèves au CM1 – Évolution des performances à dix ans », *Note d'information*, n° 12.21, MEN-DEPP.

Daussin J.-M., Keskpaik S., Rocher T., 2011, « L'évolution du nombre d'élèves en difficulté face à l'écrit depuis une dizaine d'années », *France, portrait social*, Insee, p. 137-152.

Daussin J.-M., Rocher T., Trosseille B., 2010, « L'attestation de la maîtrise du socle commun est-elle soluble dans le jugement des enseignants ? » *Éducation & formations*, n° 79, MENJVA-DEPP, p. 45-58.

Dejonghe V., Levasseur J., Alinaudm B., Peretti C., Petrone J.-C., Pons C., Thélot C., 1996, « Connaissances en français et en mathématiques des élèves des années 20 et d'aujourd'hui », *Les dossiers d'Éducation et Formations*, n° 62, MENESR-DEP.

Dessus N., Jouvanceau P., Murat F., 1996, « Les connaissances des élèves en fin de troisième générale – évolution 1984-1990-1995 », *Note d'information*, n° 96.36, MENESR-DEP.

Dickes P., Tournois J., Flieller A., Kop J.-L., 1994, *La psychométrie – Théorie et pratique de la mesure en psychologie*, Paris, PUF.

Éducation & formations, 1990, « *Évaluation CE2-sixième – Résultats nationaux septembre 1989* », n° hors-série, Paris, MENJS-DEP, 57 p.

Flieller A., 2001, « Problèmes et stratégies dans l'explication de l'effet Flynn », *in* Huteau M., *Les figures de l'intelligence* , Paris, Éditions et applications psychologiques, p. 43-66.

Flieller A., Manciaux M., Kop J.-L., 1995, *Comparaison des compétences cognitives de deux cohortes d'écoliers de 7 ans observées à vingt ans d'intervalle* (1973-1992), Rapport final, ADEPS-Nancy 2, École de Santé publique Henri-Poincaré.

Flieller A., Santigny N., Schaeffer R., 1986, « L'évolution du niveau intellectuel des enfants de 8 ans sur une période de 40 ans (1944-1984) », *L'Orientation scolaire et professionnelle*, n° 15, p. 61-83.

Garcia É., Krop J., 2013, « Cedre 2012 histoire-géographie et éducation civique : baisse des acquis des élèves de fin de collège depuis six ans », *Note d'information*, n° 13.11, MEN-DEPP.

Garcia É., Pastor J.-M., 2013, « Cedre 2012 histoire-géographie et éducation civique en fin d'école primaire : grande stabilité des acquis depuis six ans », *Note d'information*, n° 13.10, MEN-DEPP.

Haut Conseil de l'Éducation, 2011, *Les indicateurs relatifs aux acquis des élèves – Bilan des résultats de l'école-2011*.
http://www.ladocumentationfrancaise.fr/var/storage/rapports-publics/114000565/0000.pdf

Haut Conseil de l'évaluation de l'école, 2001, « Apprécier et certifier les acquis des élèves en fin de collège : diplôme et évaluations-bilans », avis n°2, MEN-DEP.

Haut Conseil de l'évaluation de l'école, 2003, « Éléments de diagnostic sur le système scolaire français », Avis n° 9, MEN-DEP.

IGEN 2007, *Note sur le suivi de la mise en œuvre de l'évaluation des élèves à l'entrée de la première année du cours élémentaire* (CE1), MENESR, rapport n° 2007-030, 39 p.

IGEN-IGAENR 2009, *Troisième note de synthèse sur la mise en œuvre de la réforme de l'enseignement primaire*, Paris, MENESR, Note 2009-072, juillet 2009, 28 p.

IGEN-IGAENR, 2005, *Les acquis des élèves, pierre de touche de la valeur de l'école ?* Paris, MENESR, rapport n° 2005-079, 83 p.

Jonas N., 2012, « Pour les générations les plus récentes, les difficultés des adultes diminuent à l'écrit, mais augmentent en calcul », *Insee Première*, n° 1426, Insee.

Jones L. V., Olkin I., 2004, *The nation's report card: Evolution and perspectives*, Bloomington, Phi Delta Kappa Educational Foundation.

Laveault D., Grégoire J., 2002, *Introduction aux théories des tests en psychologie et en sciences de l'éducation* (2e édition), Bruxelles, De Boeck.

Le Cam M., Rocher T., Verlet I., 2013, « Forte augmentation du niveau des acquis des élèves à l'entrée au CP entre 1997 et 2011 », *Note d'information*, n° 13.19, MEN-DEPP.

Le Guen M., 1991, « L'évaluation des acquis des élèves : caractéristiques et évolution du dispositif national », *L'orientation scolaire et professionnelle*, vol. 20, n° 1, p. 39-69.

Levasseur J., 1996, « L'évaluation nationale des acquis des élèves », *Revue internationale d'éducation de Sèvres*, n° 11, CIEP, p. 101-114.

Mons N., 2008, « Évaluation des politiques éducatives et comparaisons internationales », *Revue française de pédagogie*, n° 164, ENS Éditions, p. 5-13.

Mons N., 2007, *Les nouvelles politiques éducatives – La France fait-elle les bons choix ?* Paris, PUF.

Mullis I., Martin M., Foy P., Arora A., 2012, *TIMSS 2011 international results in reading*, Chestnut Hill, MA, TIMSS & PIRLS International Study Center, Boston College.

National Commission on Excellence in Education, 1983, *A Nation at risk: the imperative for educational reform*, Washington D.C.

OCDE, 2013, *Résultats du PISA 2012* : *savoirs et savoir-faire des élèves*, vol. 1 à 5, Paris.

Pair C., 2001, *Forces et faiblesses de l'évaluation du système éducatif en France*, Rapport n° 3, rapport établi à la demande du Haut Conseil de l'évaluation de l'école. http://www.ladocumentationfrancaise.fr/rapports-publics/024000206/

Peretti C., Petrone J.-C., Thélot C., 1996, « L'évolution des compétences scolaires des meilleurs élèves depuis 40 ans », *Les dossiers d'Éducation & formations*, n° 69, MENESR-DEP.

Piéron H., 1963, *Examens et docimologi*e, Paris, PUF.

Pons C., 1996, « Connaissances en français et en calcul des élèves des années 20 et d'aujourd'hui », *Note d'information*, n° 96.19, MENESR-DEP.

Rocher T., 2012, « Comment assurer la comparabilité des scores issus d'évaluations nationales annuelles et exhaustives ? Comparaison de différentes méthodes d'ajustement des métriques (*equating*) », *Actes du 24e colloque international de l'ADMÉE-Europe*, Luxembourg, papier présenté au 24e colloque international de l'ADMÉE-Europe (admee2012.uni.lu).

Rocher T., 2010, « La performance de l'école primaire : quelques résultats récents tirés de l'évaluation des acquis des élèves », *Administration et Éducation*, n° 125, AFAE, p. 43-50.

Rocher, T., 2008a, « Lire, écrire, compter : les performances des élèves de CM2 à vingt ans d'intervalle (1987-2007) », *Note d'information*, n° 08.38, MEN-DEPP.

Rocher T., 2008b, « Que nous apprennent les évaluations internationales sur le fonctionnement des systèmes éducatifs ? Une illustration avec la question du redoublement », *Éducation & formations*, n° 78, p. 63-68, MEN-DEPP.

Rocher T. 2008c, « La détermination de standards minimaux dans le cadre d'indicateurs de résultats : méthodologie, intérêt, utilité », *Mesure et évaluation en éducation*, vol. 31, n° 2, ADMÉE Canada, p. 75-91.

Rocher T., Le Donné N., 2012a, « Les aspirations professionnelles des élèves de 15 ans dans 57 pays : ambition et réalisme », *L'orientation scolaire et professionnelle*, vol. 41, n° 3, p. 439-468.

Rocher T., Le Donné N. 2012b, « Les difficultés de lecture en début de sixième – Évolution à dix ans d'intervalle (1997-2007) », *Éducation & formations*, n° 82, MEN-DEPP, p. 31-37.

Salines M., Vrignaud P., 2001, *Apprécier et certifier les acquis des élèves en fin de collège : diplôme et évaluations-bilans*, Rapport n° 2, rapport établi à la demande du Haut Conseil de l'évaluation de l'école.
http://www.ladocumentationfrancaise.fr/rapports-publics/024000205/index.shtml

Thélot C., 1992, « Que sait-on des connaissances des élèves ? », *Les dossiers d'Éducation & formations*, n° 17, MEN-DEP.

Vourc'h R., Rivière J.-P., De La Haye, F., Gombert J.-E., 2014, « Journée défense et citoyenneté 2013 : des difficultés en lecture pour un jeune français sur dix », *Note d'information*, n° 12, MENESR-DEPP.

Zwick R., 1992, "Statistical and psychometric issues in the measurement of educational achievement trends: examples from the National Assessment of Educational Progress", *Journal of Educational Statistics*, vol. 17, No. 2, AREA, p. 205-218.

MESURE DES COMPÉTENCES

Méthodes psychométriques utilisées dans le cadre des évaluations des élèves

Thierry Rocher
MENESR-DEPP, bureau de l'évaluation des élèves

Cet article présente les méthodes psychométriques qui sont généralement employées dans les programmes d'évaluations standardisées des compétences des élèves, au niveau national et au niveau international. Nous proposons un panorama de ces méthodes, de façon pédagogique, mais également technique. Leurs fondements théoriques ainsi que leurs hypothèses sous-jacentes sont présentés. Nous montrons leur intérêt d'un point de vue pratique, mais également leurs limites. Enfin, une description des analyses psychométriques réalisées dans le cadre d'une évaluation du cycle Cedre est proposée.

Les programmes d'évaluations standardisées réalisés à la DEPP ont pour objectif de mesurer le niveau des acquis des élèves, à différents moments de la scolarité. Ces évaluations s'intéressent aux élèves comme éléments d'une population ; elles n'ont pas vocation à rendre compte de leurs résultats au niveau individuel. Elles se situent donc à un niveau global et doivent permettre d'apprécier les résultats du système éducatif et leur évolution dans le temps [Salines et Vrignaud, 2001 ; Bottani et Vrignaud, 2005 ; Trosseille et Rocher, dans ce numéro, p. 15]. D'un point de vue méthodologique, elles reposent sur des échantillons représentatifs [Garcia, Le Cam et Rocher, dans ce numéro, p. 101] et suivent des procédures standardisées afin de limiter l'erreur de mesure à tous les niveaux (passation, correction, etc.). Ces évaluations se distinguent d'autres enquêtes notamment à travers l'emploi d'un ensemble de méthodes relevant du domaine de la psychométrie, c'est-à-dire de la mesure de dimensions psychologiques, et qui a donné naissance au domaine de l'édumétrie dans le champ de l'éducation. Ces méthodes restent relativement méconnues

en France. Largement employées dans les évaluations nationales ou internationales, elles sont peu diffusées, que ce soit dans le monde académique, le monde éducatif ou encore celui de la statistique publique. Cet article a pour objectif de dresser un panorama des méthodes psychométriques employées dans les programmes d'évaluations standardisées et de présenter de manière pédagogique leurs fondements théoriques et leurs aspects pratiques. Nous présentons tout d'abord le cadre conceptuel de la mesure des compétences des élèves, qui consiste à considérer que les performances observées aux items d'une évaluation sont les manifestations d'une variable latente, non observable directement. Après avoir introduit quelques éléments descriptifs, nous présentons les modélisations habituellement employées, à savoir les modèles de réponse à l'item. Nous montrons l'intérêt de ces modèles, à la fois sur le plan théorique et sur le plan pratique, et nous étudions les hypothèses fondamentales sur lesquelles ils reposent. Enfin, nous décrivons le déroulement des analyses psychométriques qui sont réalisées dans le cadre d'une évaluation Cedre (cycle des évaluations disciplinaires réalisées sur échantillons).

CADRE GÉNÉRAL

Mesurer une variable latente

Les programmes d'évaluation des acquis des élèves, tels que PISA ou Cedre, se situent au carrefour de deux traditions méthodologiques : celle de la psychométrie, pour ce qui relève de la mesure de dimensions psychologiques, en l'occurrence des acquis cognitifs ; celle des enquêtes statistiques pour ce qui a trait aux procédures de recueil des données.

C'est la nature de la variable mesurée qui distingue principalement les programmes d'évaluation d'autres enquêtes statistiques. En effet, il est convenu que les compétences des élèves ne s'observent pas directement. Seules les manifestations de ces compétences sont observables, par exemple à travers les résultats obtenus à un test standardisé. L'existence supposée de la compétence visée est alors matérialisée dans la réussite au test. D'une certaine manière, on pourrait avancer que c'est l'opération de mesure elle-même qui définit concrètement l'objet de la mesure, d'où le célèbre pied de nez d'Alfred Binet, en réponse à la question *« qu'est-ce que l'intelligence ? »* : *« c'est ce que mesure mon test »*. Ainsi, le terme de « construit » est souvent employé pour désigner l'objet de la mesure.

Bien entendu, toute statistique peut être considérée comme un construit, pas seulement celles ayant trait à l'évaluation. Cependant, des degrés sont sans doute à distinguer, en lien avec le caractère tangible de la variable visée. Par exemple, la réussite scolaire peut-être appréhendée par la variable « réussite au baccalauréat » qui est mesurable directement, car elle est sanctionnée par un diplôme, donnant lieu à un acte administratif que l'on peut comptabiliser. Le « décrochage scolaire », quant à lui, est un concept qui doit reposer sur une définition précise, choisie parmi un ensemble de définitions possibles, ce choix faisant acte de construction. Une fois la définition établie, le calcul repose le plus souvent sur l'observation de variables administratives, telles que la non-réinscription dans un établissement scolaire.

En comparaison, la mesure des compétences se présente comme une démarche

de construction assez particulière. L'idée sous-jacente de la psychométrie consiste à postuler qu'un test mesure des performances qui sont la manifestation d'un niveau de compétence, non observable directement. Ainsi, l'objet de la mesure est une variable latente. Notons que cette approche n'est pas propre au domaine de la cognition. On retrouve ce type de variable en économie avec par exemple la notion de propension, en sciences politiques avec la notion d'opinion ou encore en médecine avec la notion de qualité de vie [voir par exemple : Falissard, 2008].
Envisager les résultats à une évaluation comme résultant d'un processus de mesure d'une variable latente ne s'impose pas de lui-même. En effet, il est tout à fait possible de considérer uniquement le nombre de points obtenus à un test et de ne pas donner plus de significations à cette statistique qu'un score observé à un test. Mais cette démarche est très fruste d'un point de vue théorique et trouve vite des limites en pratique, notamment en termes de comparabilité entre différentes populations ou entre différentes épreuves. Le cadre conceptuel de la mesure d'une variable latente est plus adapté à la problématique de l'évaluation des acquis des élèves, comme nous le verrons dans cet article.

Un exemple introductif

Avant d'entrer dans des considérations plus techniques, nous présentons tout d'abord un exemple d'application qui a pour seul objectif d'illustrer de façon pédagogique les grandes notions de psychométrie.
Cet exemple porte sur la taille des individus. La situation est la suivante : nous n'avons aucun moyen de mesurer directement la taille des individus d'un échantillon donné. Mais nous avons la possibilité de proposer un questionnaire, composé de questions appelant une réponse binaire (oui/non) et n'évoquant pas directement la taille. Nous nous plaçons ainsi artificiellement dans le cas de la mesure d'une variable latente que nous cherchons à approcher à l'aide d'un questionnaire, soit un dispositif de mesure apparemment comparable à celui d'une évaluation standardisée.
Ce cas d'école est depuis longtemps utilisé aux Pays-Bas dans les cours de psychométrie : Glas [2008] en donne quelques illustrations. Dans cet esprit, nous avons de notre côté élaboré un questionnaire de 24 items, nécessitant simplement d'indiquer l'accord ou le désaccord avec une série d'affirmations. Voici un extrait de ce questionnaire :

1. Je dois souvent faire attention à ne pas me cogner la tête
2. Pour les photos de groupe, on me demande souvent d'être au premier rang
3. On me demande souvent si je fais du basket-ball
4. Dans la plupart des voitures, je suis mal assis(e)
5. Je dois souvent faire faire les ourlets quand j'achète un pantalon
6. Je dois souvent me baisser pour faire la bise
7. Au supermarché, je dois souvent demander de l'aide pour attraper des produits en haut des gondoles
8. À deux sous un parapluie, c'est souvent moi qui le tiens

Ce questionnaire a été proposé via Internet à un échantillon composé de 276 adultes dans un réseau à la fois professionnel et personnel. L'échantillon est plutôt jeune (55 % sont âgés de moins de 30 ans) et féminin (65 % de femmes), mais la question de la représentativité n'est pas importante au regard de notre propos qui concerne les problématiques de mesure.

Une notion fondamentale en psychométrie est celle de la **validité** : le test mesure-t-il bien ce qu'il est censé mesurer ?
Dans le cadre de notre exemple, nous pouvons approcher la validité assez directement puisque la dernière question demande aux enquêtés d'indiquer leur taille[1]. Nous avons calculé un score de façon très simple à partir des 24 questions en attribuant 1 point pour chacune d'entre elles, en fonction de la modalité associée à une taille plus élevée : par exemple, les individus obtiennent un point s'ils répondent oui à la première question, 0 sinon ; et inversement, pour la deuxième question. Il est alors possible d'analyser la relation entre ce score et la taille déclarée : le coefficient de corrélation linéaire de 0,85 indique un lien positif et fort entre le score construit et la taille. De ce point de vue, nous pouvons conclure à la validité de notre questionnaire, même si l'ampleur de la corrélation observée peut être largement discutée.

En matière d'évaluation standardisée, nous ne disposons évidemment pas d'une variable de référence, telle que la taille réelle, puisque précisément les compétences sont inobservables directement. La question de la validité d'une évaluation devient alors une question complexe. La littérature abonde de références dans ce domaine [voir par exemple Newton et Shaw, 2014 ; en français, Laveault et Grégoire, 2002]. En résumé, différents types de validité sont généralement distingués : validité de contenu, de construit, critériée, etc. Dans le cas de Cedre par exemple, la validité est principalement assurée à travers une validité dite de contenu : un groupe de concepteurs composé d'enseignants, d'inspecteurs, de formateurs est garant, sur la base de leur propre expertise, de l'adéquation du contenu de l'évaluation avec les programmes scolaires, les instructions officielles et les pratiques de classes. Ainsi, un niveau de performance observé à l'évaluation de mathématiques est censé traduire un niveau de compétence, au regard des attendus en mathématiques.
Au-delà de la validité, une question centrale de psychométrie est celle de la **dimensionnalité** d'un ensemble d'items. Nous calculons un score, mais cela n'a de sens que sous l'hypothèse que les items mesurent la même dimension, que le test est unidimensionnel. Cependant, il est clair que les items présentés ici ne mesurent pas purement la dimension taille, mais interrogent chacun une multiplicité de dimensions. L'idée est qu'un facteur commun prépondérant relie ces items, facteur lié à la taille. Ainsi, la majorité des évaluations rend compte des résultats à travers un score global, selon un cadre unidimensionnel.

L'exemple nous permet également d'illustrer la notion de fonctionnements différentiels d'items ou FDI, qui est liée à la question de la dimensionnalité. Un FDI apparaît entre des groupes d'individus dès lors qu'à niveau égal sur la variable latente mesurée, la probabilité de réussir un item donné n'est pas la même selon le groupe considéré. Cela signifie qu'une autre variable, liée au groupe, est intervenue, au-delà de la dimension visée. Un fonctionnement différentiel se traduit souvent par une différence de réponse entre les groupes plus importante à l'item considéré qu'en moyenne sur l'ensemble des items. Par exemple, à la question *« À deux sous un parapluie, c'est souvent moi qui le tiens »*, 89 % des hommes répondent oui contre

1. Il ne s'agit donc pas de la taille exacte mais de la taille déclarée, ce qui peut introduire un décalage, par le jeu des arrondis que les personnes font naturellement concernant leur taille : par exemple, on observe certaines concentrations, autour de 165 cm, mais peu de valeurs telles que 163 cm… Nous supposerons cependant ici que la taille est déclarée sans erreur.

52 % des femmes, soit un écart de 37 points, alors qu'en moyenne sur l'ensemble des items, la différence entre les hommes et les femmes est de 20 points. Cet écart de 20 points renvoie à ce qu'on appelle l'impact, c'est-à-dire la différence entre les deux groupes sur la variable latente, en l'occurrence la différence de taille entre hommes et femmes. Un écart additionnel renvoie à un fonctionnement différentiel. À taille égale, les hommes disent tenir le parapluie plus souvent que les femmes. Une autre dimension que la taille, liée au genre, a joué dans la réponse. La question est alors dite « biaisée » selon le genre[2]. L'étude des FDI est fondamentale en matière de comparaison temporelle ou internationale des acquis des élèves. Nous revenons plus en détail sur cette notion par la suite.

De manière pratique, un concept important est celui de la fidélité du test. Le score calculé comporte une part d'erreur de mesure. En effet, on peut considérer que les items d'un test ont été échantillonnés dans l'« univers » possible des items censés mesurer la dimension visée par le test. Dès lors, un autre ensemble d'items n'aurait pas conduit exactement aux mêmes scores. Le test est dit fidèle lorsque l'erreur de mesure est réduite. Le coefficient α de Cronbach, présenté plus loin, est un indicateur de fidélité du test. En l'occurrence, pour le questionnaire sur la taille, il a pour valeur 0,80, ce qui est satisfaisant.

Au-delà de cet indice global, il est intéressant d'étudier les items eux-mêmes. Les taux de réponse observés aux différentes modalités proposées – ici, oui ou non – sont bien entendu des indicateurs essentiels. Par exemple, dans le cas d'une évaluation, les items peuvent être comparés en termes de difficulté, qui est appréciée par le pourcentage de bonnes réponses. Une autre notion importante est celle de pouvoir discriminant de chaque item, qui renvoie au lien avec les résultats obtenus à l'ensemble du test. En effet, si l'item mesure bien la dimension qu'il est censé mesurer, alors il discriminera bien les personnes selon cette dimension. Une manière de vérifier qu'il mesure bien la dimension supposée est d'examiner les corrélations de l'item avec d'autres items censés mesurer la même dimension. Concernant le questionnaire sur la taille, les corrélations items-test, c'est-à-dire les corrélations entre la réussite à un item donné et le score aux autres items, sont assez élevées, à l'exception d'un item dont la corrélation item-test est nulle. Il s'agit d'un item repris de l'article de Glas [2008] : « Dans un lit, j'ai souvent froid aux pieds. » Utilisé aux Pays-Bas, cet item doit donc être discriminant selon la taille des Néerlandais, mais ce n'est pas le cas sur notre échantillon français. Nous supposons qu'il s'agit d'une différence culturelle liée aux habitudes de border les draps ou la couette, forte en France et absente aux Pays-Bas où le problème d'avoir froid aux pieds la nuit se pose sans doute pour les personnes de grande taille. Ainsi, cet item ne mesure pas la dimension taille en France, mais plutôt une autre dimension décorrélée, telle que la frilosité…

Pour finir avec le cas d'école, nous abordons la notion d'échelle. Avant tout, notons que le questionnaire ne nous permet pas de connaître la taille des individus. Il nous permet simplement de classer avec plus ou moins de fiabilité les individus

2. Une autre question présente un fonctionnement différentiel du même ordre : « Au supermarché, je dois souvent demander de l'aide pour attraper des produits en haut des gondoles ». Aucun homme ne répond oui à cette question, alors qu'un tiers des femmes répond positivement, en lien avec leur taille… Nous laissons ici au lecteur le soin de formuler sa propre interprétation.

selon leur taille, et d'introduire une métrique. Ainsi, le score simple que nous avons calculé, compris entre 0 et 24, de moyenne 11,0 et d'écart-type 4,3, est une échelle de mesure, sur laquelle il est possible d'établir un classement des individus ainsi que des distances entre eux. Il s'agit d'une échelle dite d'intervalle, qui autorise la comparaison des intervalles de scores entre individus. Autrement dit, les rapports entre intervalles ne sont pas modifiés par transformation linéaire[3]. L'origine et l'unité peuvent donc être transformées, et ce de manière arbitraire. Dans notre exemple, nous pouvons rendre compte des résultats sur l'échelle des scores observés, de moyenne 11,0 et d'écart-type 4,3, mais également sur une échelle standardisée, de moyenne 0 et d'écart-type 1, ou de moyenne 250 et d'écart-type 50 comme dans Cedre, ou encore de moyenne 500 et d'écart-type 100 comme dans PISA. Autrement dit, les valeurs elles-mêmes n'ont pas de significations, au-delà du classement et de la distance entre individus.

APPROCHE CLASSIQUE

Dans un premier temps, nous posons quelques notations et nous présentons les principales statistiques descriptives utilisées pour décrire un test, issues de la « théorie classique des tests » que nous évoquons rapidement.

Réussite et score

On note n le nombre d'élèves ayant passé une évaluation composée de J items. On note Y_i^j la réponse de l'élève i (i = 1, ..., n) à l'item j (j = 1, ..., J). Dans notre cas, les items sont dichotomiques, c'est-à-dire qu'ils ne prennent que deux modalités (la réussite ou l'échec) :

$$Y_i^j = \begin{cases} 1 & \text{si l'élève } i \text{ réussit l'item } j \\ 0 & \text{si l'élève } i \text{ échoue à l'item } j \end{cases} \quad (1)$$

Le taux de réussite à l'item j est la proportion d'élèves ayant réussi l'item j. Il est noté p_j :

$$p_j = \frac{1}{n}\sum_{i=1}^{n} Y_i^j \quad (2)$$

Le taux de réussite d'un item renvoie à son niveau de difficulté. C'est certainement la caractéristique la plus importante, qui permet de construire un test de niveau adapté à l'objectif de l'évaluation, en s'assurant que les différents niveaux de difficulté sont balayés.

3. C'est le cas par exemple des échelles de température. S'il fait 20°C à Paris, 30°C à Grenoble et 40°C à Rome, l'écart de température entre Rome et Paris est deux fois plus grand que celui entre Grenoble et Paris. C'est également vrai en Fahrenheit, après transformation linéaire. En revanche, on ne peut pas dire qu'il fait deux fois plus chaud à Rome qu'à Paris, cela dépend de l'échelle utilisée. Seules les échelles dites de rapport (poids, taille, revenu, etc.) permettent des comparaisons de rapports.

Le score observé à l'évaluation pour l'élève i, noté S_i, correspond au nombre d'items réussis par l'individu i :

$$S_i = \sum_{j=1}^{J} Y_i^j \qquad (3)$$

La théorie classique des tests a précisément pour objet d'étude le score S_i obtenu par un élève à un test. Elle postule notamment que ce score observé résulte de la somme d'un score « vrai » inobservé et d'une erreur de mesure. Un certain nombre d'hypothèses portent alors sur le terme d'erreur [pour plus d'informations, voir par exemple Laveault et Grégoire, 2002].

Fidélité

Dans le cadre de la théorie classique des tests, la fidélité (*reliability*) est définie comme la corrélation entre le score observé et le score vrai : le test est fidèle, lorsque l'erreur de mesure est réduite. Une manière d'estimer cette erreur de mesure consiste par exemple à calculer les corrélations entre les différents sous-scores possibles : plus ces corrélations sont élevées, plus le test est dit fidèle[4].
Le coefficient α de Cronbach est un indice destiné à mesurer la fidélité de l'épreuve. Il est compris entre 0 et 1. Sa version « standardisée » s'écrit :

$$\alpha = \frac{J\,\overline{r}}{1 + (J-1)\overline{r}} \qquad (4)$$

où $\overline{r}$ est la moyenne des corrélations inter-items.
De ce point de vue, cet indicateur renseigne sur la consistance interne du test. En pratique, une valeur supérieure à 0,8 témoigne d'une bonne fidélité[5].

Indices de discrimination

Des indices importants concernent le pouvoir discriminant des items. Nous présentons ici l'indice « r-bis point » ou coefficient point-bisérial qui est le coefficient de corrélation linéaire entre la variable indicatrice de réussite à l'item Y^j et le score S. Appelé également « corrélation item-test », il indique dans quelle mesure l'item s'inscrit dans la dimension générale. Une autre manière de l'envisager consiste à le formuler en fonction de la différence de performance constatée entre les élèves qui réussissent l'item et ceux qui échouent. En effet, on peut montrer que :

$$r_{bis-point}(j) = corr(Y^j, S) = \frac{\overline{S}_{(j1)} - \overline{S}_{(j0)}}{\sigma_S}\sqrt{p_j(1-p_j)} \qquad (5)$$

où $\overline{S}_{(j1)}$ est le score moyen sur l'ensemble de l'évaluation des élèves ayant réussi l'item j, $\overline{S}_{(j0)}$ celui des élèves ayant échoué à l'item et σ_S est l'écart-type des scores.

4. Notons au passage que la naissance des analyses factorielles est en lien avec ce sujet : Charles Spearman cherchait précisément à dégager un facteur général à partir de l'analyse des corrélations entre des scores obtenus à différents tests.
5. La littérature indique plutôt un seuil de 0,70 [Peterson, 1994]. Cependant, comme le montre la formule ci-dessus, le coefficient α est lié au nombre d'items, qui est important dans les évaluations conduites par la DEPP afin de couvrir les nombreux éléments des programmes scolaires. Des facteurs de correction existent néanmoins et permettent de comparer des tests de longueurs différentes.

C'est donc bien un indice de discrimination, entre les élèves qui réussissent et ceux qui échouent à l'item. En pratique, on préfère s'appuyer sur les $r_{bis-point}$ corrigés, c'est à dire calculés par rapport au score à l'évaluation privée de l'item considéré. Une valeur inférieure à 0,2 indique un item peu discriminant [Laveault et Grégoire, 2002].

MODÈLES DE RÉPONSE À L'ITEM (MRI)

Dans la pratique, l'approche classique comporte certaines limites. En se concentrant sur l'analyse du score observé, c'est-à-dire du nombre de bonnes réponses aux items d'un test donné, les résultats dépendent de l'ensemble des items considérés. L'approche classique permet donc difficilement de distinguer ce qui relève de la difficulté du test de ce qui relève du niveau de compétence des élèves. Le recours à une modélisation plus adaptée, qui se situe au niveau des items eux-mêmes et non au niveau du score agrégé, est apparu nécessaire. En particulier, les modèles de réponse à l'item (MRI), nés dans les années 1960, se sont imposés dans le champ des évaluations standardisées à grande échelle. Nous présentons quelques-uns de ces modèles.

Présentation générale

Les MRI sont une classe de modèles probabilistes. Ils modélisent la probabilité qu'un élève donne une certaine réponse à un item, en fonction de paramètres concernant l'élève et l'item. De manière très générale, les MRI peuvent être présentés de la manière suivante :

$$P(Y = k|\boldsymbol{\theta}, \boldsymbol{\xi}) = F(\boldsymbol{\theta}, \boldsymbol{\xi}, k) \quad (6)$$

La probabilité qu'un élève donne la réponse k à l'item Y dépend de caractéristiques $\boldsymbol{\theta}$ concernant l'élève et de caractéristiques $\boldsymbol{\xi}$ concernant l'item Y. La fonction F est typiquement une fonction de répartition, à valeur dans]0, 1[.

En comparaison de la théorie classique des tests, ces modèles ont l'avantage de séparer ce qui relève des élèves de ce qui relève des items, la réponse résultant d'une interaction entre ces deux composantes. Les MRI ont un intérêt pratique pour la construction de tests et que nous détaillons par la suite : si le modèle est bien spécifié sur un échantillon donné, les paramètres des items – en particulier leurs difficultés – peuvent être considérés comme fixes et applicables à d'autres échantillons dont il sera alors possible de déduire les paramètres relatifs aux élèves – en particulier, leur niveau de compétence. Les modèles de réponse à l'item ont donné lieu à une littérature extrêmement fournie. Le lecteur intéressé est invité à consulter, par exemple, Embretson et Reise [2000] ou bien, en français, Bertrand et Blais [2004].

Notre attention va se concentrer sur le cas où $\boldsymbol{\theta}$ est un scalaire (un nombre réel), c'est-à-dire que le MRI est dit unidimensionnel. En outre, nous nous restreignons ici au cas d'items dichotomiques ($k \in \{0,1\}$). Des extensions existent, mais leur présentation sort du cadre de cet article.

Modèle de Rasch (1PL)

Proposé par Rasch [1960], le modèle le plus simple, appelé aussi MRI « à un paramètre » (1PL pour *One-Parameter Logistic*) s'écrit de la manière suivante :

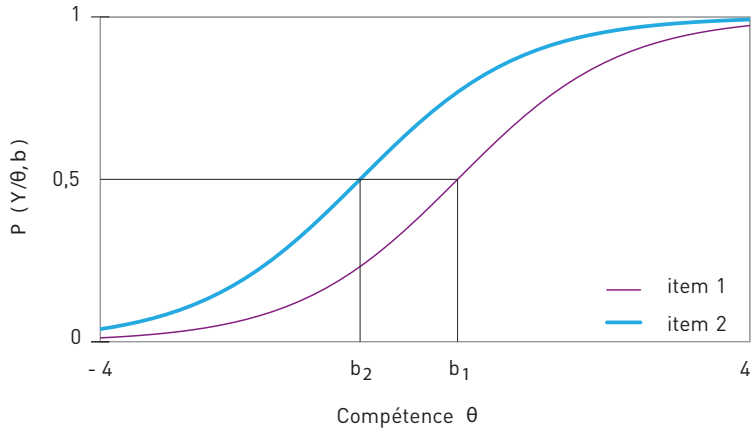
$$P_{ij} = P(Y_i^j = 1|\theta_i, b_j) = \frac{e^{\theta_i - b_j}}{1 + e^{\theta_i - b_j}} \quad (7)$$

i.e. la probabilité P_{ij} que l'élève i réussisse l'item j est une fonction sigmoïde[6] du niveau de compétence de l'élève i et du niveau de difficulté b_j de l'item j.

La fonction sigmoïde étant une fonction croissante, il ressort que la probabilité de réussite augmente lorsque le niveau de compétence de l'élève augmente et diminue lorsque le niveau de difficulté de l'item augmente, ce qui traduit à l'évidence les relations attendues entre réussite, difficulté et niveau de compétence. L'intérêt de ce type de modélisation, et ce qui explique son succès, c'est de séparer deux concepts-clé, à savoir la difficulté de l'item et le niveau de compétence de l'élève.

Autre avantage : le niveau de compétence des élèves et la difficulté des items sont placés sur la même échelle, par le simple fait de la soustraction ($\theta_i - b_j$). Cette propriété permet d'interpréter le niveau de difficulté des items par rapprochement avec le continuum de compétence. Ainsi, les élèves situés à un niveau de compétence égal à b_j auront 50 % de chances de réussir l'item, ce que traduit visuellement la représentation des courbes caractéristiques des items (CCI) selon ce modèle ▸ **Figure 1**.

▸ Figure 1 Modèle de réponse à l'item – 1 paramètre



Note de lecture : la probabilité de réussir l'item (en ordonnées) dépend du niveau de compétence (en abscisse). Par définition, le paramètre de difficulté d'un item correspond au niveau de compétence ayant 50 % de chances de réussir l'item. Ainsi, l'item 1 en trait fin est plus difficile que l'item 2 en trait plein. La probabilité de le réussir est plus élevée quel que soit le niveau de compétence.

6. La fonction sigmoïde est définie par : $\forall x, f(x) = \frac{e^x}{1 + e^x}$, à valeur dans]0, 1[.

Modèle à deux paramètres (2PL)

Birnbaum [1968] a proposé d'introduire un deuxième paramètre, dit de discrimination :

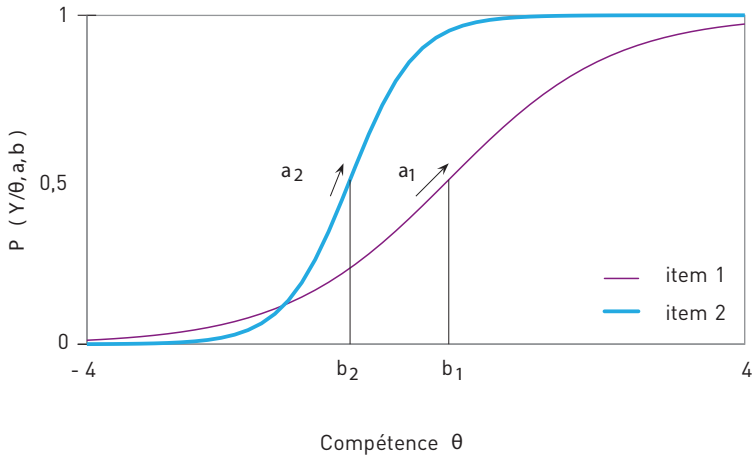
$$P_{ij} = P(Y_i^j = 1|\theta_i, a_j, b_j) = \frac{e^{1,7a_j(\theta_i - b_j)}}{1 + e^{1,7a_j(\theta_i - b_j)}} \quad (8)$$

où a_j ($a_j > 0$) représente la pente au point d'inflexion de la courbe caractéristique de l'item j qui varie d'un item à l'autre et la constante 1,7 est introduite pour rapprocher la fonction sigmoïde de la fonction de répartition de la loi normale ▸ **Figure 2**.

Pour un item très discriminant, la probabilité de le réussir sera très faible en deçà d'un certain niveau de compétence et très élevée au-delà de ce même niveau. Ainsi, une faible différence de niveau de compétence peut conduire à des probabilités de réussite très différentes. C'est le cas de l'item 2 sur la figure 2. De son côté, un item peu discriminant pourra conduire à de faibles différences de probabilité de réussite, pour un écart de niveau de compétence important.

Cet indice de discrimination peut ainsi être interprété en termes de quantité d'information portée par l'item. Nous ne développons pas cette notion statistique ici, mais l'idée est la suivante : un élève qui réussit un item très discriminant se situe très certainement au-dessus du niveau de difficulté de l'item sur l'échelle de compétence, alors que pour un item de discrimination faible, l'incertitude est plus grande quant à la position de l'élève sur l'échelle. De ce point de vue, un item discriminant apporte de l'information.

Figure 2 **Modèle de réponse à l'item – 2 paramètres**



Note de lecture : la probabilité de réussir l'item (en ordonnées) dépend du niveau de compétence (en abscisse). L'item 1 en trait fin est plus difficile que l'item 2 en trait plein ($b_1 > b_2$), et il est moins discriminant ($a_1 < a_2$).

Par rapport au cadre du modèle de Rasch, l'estimation des paramètres est plus complexe (voir annexe p. 57). Mais au-delà des aspects techniques, certaines propriétés ne sont plus valables dans le cas du modèle à deux paramètres. C'est le cas de la propriété dite d'« objectivité spécifique », qui pourrait se résumer au fait que dans le modèle de Rasch, la probabilité de réussir un item d'un certain niveau de difficulté est toujours inférieure à la probabilité de réussir un item plus difficile. Mais dans le cas du modèle à deux paramètres, les courbes caractéristiques peuvent se croiser : un item peut alors apparaître plus facile ou plus difficile selon le niveau de compétence considéré. Selon certains auteurs, la propriété d'objectivité spécifique conférerait à l'opération de mesure en sciences sociales des propriétés équivalentes à celles prévalant en sciences physiques [Andrich, 2004]. Les tenants de cette vision sont donc partisans de construire l'instrument de mesure en fonction des propriétés du modèle de Rasch : les caractéristiques du test doivent satisfaire les exigences du modèle, d'où l'appellation de théorie de réponse à l'item (*Item Response Theory*)[7]. Cependant, en pratique, l'égalité des discriminations imposée par le modèle de Rasch est une contrainte très exigeante : elle revient à éliminer de nombreux items après avoir estimé leur adéquation au modèle. En effet, la prise en compte de la discrimination permet de mieux modéliser le fonctionnement des items et revient finalement à donner un poids plus important aux items les plus discriminants.

Notons enfin que les modèles présentés ne sont pas identifiables (voir annexe p. 57). Il est nécessaire de fixer des valeurs arbitraires concernant la moyenne et l'écart-type des θ. Ainsi que nous l'avons évoqué dans l'exemple introductif, le continuum θ peut s'apparenter à une échelle de température sur laquelle il est possible d'opérer des transformations.

Autres MRI

Ces modèles admettent de nombreuses variantes qui sortent du cadre de cet article. Ainsi, une extension « naturelle » des modèles présentés précédemment vers une représentation multidimensionnelle consiste à supposer que θ n'est plus une seule variable latente, mais un vecteur de dimension D. Des distinctions concernent alors la nature compensatoire ou non des dimensions.
Une autre approche assez courante consiste à introduire un troisième paramètre, dit de pseudo-chances (*guessing*). Il s'agit d'une asymptote horizontale non nulle à la courbe caractéristique de l'item : la probabilité de réussite pour les faibles niveaux de compétence ne tend plus vers 0, mais vers une certaine valeur positive, qui dépend de l'item, et qui représente la chance de réussite « au hasard ».
Enfin, nous ne développons pas le cas d'items polytomiques mais des extensions existent selon que l'on considère que les réponses possibles sont un nombre de points attribués du fait de la maîtrise de différents aspects – *partial credit models* – ou bien que l'on considère que les réponses sont hiérarchisées en niveaux plus ou moins corrects imbriqués les uns dans les autres – *graded response models*.

7. C'est d'ailleurs la terminologie la plus répandue dans la littérature en langue anglaise, bien qu'elle soit critiquée, notamment par Goldstein [1989] qui considère qu'il ne s'agit pas d'une théorie mais bien d'une modélisation.

APPLICATIONS

Les MRI ont de nombreux avantages pratiques. Nous donnons un aperçu de quelques applications montrant l'intérêt d'avoir recours à ces modèles. Bien entendu, leur mise en œuvre est soumise au respect de certaines hypothèses que nous décrivons dans la section suivante.

Assurer la comparabilité

Les MRI sont très utiles dès lors qu'il s'agit de comparer les niveaux de compétence de différents groupes d'élèves. Par exemple, dans le cadre de comparaisons temporelles, la reprise à l'identique de l'ensemble des items passés lors de la précédente enquête n'est pas forcément pertinente, au regard de l'évolution des programmes scolaires, des pratiques, de l'environnement, etc. Certains items doivent être retirés, d'autres ajoutés. Par conséquent, les élèves des deux cohortes passent une épreuve en partie différente. Dès lors, comment assurer la comparabilité des résultats ?
Cette problématique renvoie à la notion d'ajustement des métriques ou de parallélisation des épreuves (en anglais : *equating*). Il s'agit de positionner sur la même échelle de compétence les élèves de différentes cohortes, à partir de leurs résultats observés à des évaluations différentes. De nombreuses techniques existent et sont couramment employées dans les programmes d'évaluations standardisées. Typiquement, les comparaisons sont établies à partir d'items communs, repris à l'identique d'un moment de mesure à l'autre. Les modèles de réponse à l'item fournissent alors un cadre approprié, dans la mesure où ils distinguent les paramètres des items, qui sont considérés comme fixes, des paramètres des élèves, considérés comme variables.
Plusieurs stratégies d'estimation sont possibles. La première vise à estimer les paramètres des items – la difficulté β_j et les discriminations a_j pour un MRI à deux paramètres – à partir des données de la première cohorte, en fixant la moyenne et l'écart-type des niveaux de compétence θ_i, par exemple à 0 et à 1 respectivement. Les valeurs des paramètres des items communs sont considérées comme fixes et elles sont utilisées pour estimer les θ_i de la deuxième cohorte.
Une autre possibilité consiste à estimer les paramètres des items sur chacun des groupes pris séparément. Les paramètres des items communs sont alors « alignés » de manière à en déduire les différences de compétences entre groupes. En effet, dans le cas du modèle à deux paramètres par exemple, les modifications

$$\theta^* = A\theta + B,\ b_j^* = Ab_j + B \text{ et } a_j^* = a_j \ / \ A \qquad (9)$$

ne modifient pas la probabilité de réussite. L'*equating* consiste alors à déterminer les coefficients A et B tels que les paramètres des items communs aux deux évaluations soient proches, selon qu'ils sont estimés sur un groupe ou sur un autre. Puis, l'ajustement des métriques se déduit en appliquant $\theta^* = A\theta + B$, c'est-à-dire une transformation linéaire, de la même manière que l'on passe des degrés Celsius aux degrés Fahrenheit. De nombreuses méthodes ont été proposées pour estimer A et B [Kolen et Brennan, 2004][8].

8. Par exemple, une méthode très simple, dite *mean/mean*, consiste tout simplement à remplacer les a et les a_j^* par leurs moyennes respectives $\overline{a}$ et $\overline{a^*}$ pour calculer A, puis à remplacer les b_j et les b_j^* par leurs moyennes respectives $\overline{b}$ et $\overline{b^*}$ pour calculer B. De leur côté, Stocking et Lord [1983] ont proposé une procédure plus complexe qui consiste à minimiser une fonction de perte pour trouver A et B.

Une perspective différente, appelée « estimation concourante », envisage toutes les données de manière simultanée en autorisant des différences de niveau de compétence entre groupes. Les réponses des élèves aux items qu'ils n'ont pas vus sont traitées comme des valeurs manquantes par l'algorithme d'estimation (voir annexe). C'est la stratégie qui est retenue dans Cedre, dans la mesure où elle conduit aux résultats les plus stables, ainsi que le rapporte le plus souvent la littérature scientifique sur le sujet.
Dans tous les cas, l'hypothèse est faite que les items communs « fonctionnent » de la même manière, quel que soit le groupe d'élèves considéré. Cela signifie que leur difficulté n'est pas altérée d'un groupe d'élèves à l'autre. Cette hypothèse est fondamentale et renvoie à la notion de fonctionnements différentiels d'items que nous développons plus loin.

Quelques variantes

Nous avons présenté le principe de l'*equating* entre deux groupes d'élèves à partir d'items communs. La même démarche peut être employée pour estimer des niveaux de compétence comparables, pour les mêmes élèves, à différents moments de mesure.
Par exemple, dans le cadre de l'évaluation de l'enseignement intégré de science et technologie (EIST), les élèves ont été suivis de la sixième à la troisième et ont passé cinq évaluations [Le Cam et Cosnefroy, dans ce numéro, p. 283]. Deux évaluations successives comportent des items communs, ce qui a permis de calculer des scores sur une échelle commune aux cinq temps de mesure, facilitant ainsi l'analyse des progressions des élèves au cours du collège. La même méthode a été appliquée aux données du panel d'élèves de sixième pour suivre l'évolution des acquis des élèves entre la sixième et la troisième, à partir d'items repris entre les deux évaluations [Ben Ali et Vourc'h, dans ce numéro, p. 211].

Une autre application du principe d'*equating* consiste à ajuster les métriques *via* les élèves et non plus *via* les items. Par exemple, pour la reprise de l'évaluation Lire, écrire, compter en CM2 [Rocher, 2008], l'épreuve de calcul de 2007 a repris des items de l'évaluation de 1987, mais également des items d'une autre évaluation sur échantillon, datant de 1999. Il n'y avait aucun item commun entre l'épreuve de 1987 et celle de 1999, mais grâce à la reprise des deux épreuves en 2007, il est possible d'estimer les paramètres de l'ensemble des items et d'en déduire les niveaux de compétences positionnés sur la même échelle, quelle que soit l'année considérée.
Un autre dispositif, courant en matière de tests de langues, consiste à estimer au fur et à mesure les paramètres des items, afin de constituer une banque d'items dans laquelle il sera possible de piocher pour proposer aux candidats des épreuves différentes, selon les moments, selon les pays, afin d'éviter les risques d'exposition et de triches tout en garantissant l'établissement de scores comparables quelle que soit l'épreuve passée. Cette mécanique repose parfois sur la gestion de « flux » d'items. Par exemple, un candidat passe une épreuve dont une partie est composée d'items sélectionnés dans une banque dont on connaît les paramètres et l'autre partie est constituée d'items non calibrés, qui ne sont pas pris en compte dans le calcul du score pour ce candidat, mais les données recueillies serviront à estimer les paramètres de ces nouveaux items.

Ce même principe – dit de pré-test/post-test – a été appliqué pour les évaluations nationales exhaustives de CE1 et de CM2 ayant eu lieu entre 2009 et 2012 afin de comparer l'évolution des scores d'une année sur l'autre, alors que les épreuves étaient entièrement renouvelées chaque année, afin d'éviter les risques de bachotage[9].

Évaluations adaptatives

Un cadre d'application très important des MRI est celui des évaluations dites « adaptatives ». Le principe est le suivant : chaque élève passe une première épreuve ; s'il échoue, une épreuve plus facile lui est proposée ; s'il réussit, il passera une épreuve plus difficile. Ce processus itératif conduit à une estimation plus précise et plus rapide du niveau de compétence de chaque élève. En outre, proposer aux élèves des items de difficulté adaptée à leur niveau peut apparaître comme un levier pour favoriser la motivation des élèves par rapport à la situation d'évaluation [Keskpaik et Rocher, dans ce numéro, p. 119]. Avec le développement de l'informatique, cette procédure s'est répandue dans le domaine de l'évaluation [Wainer, 2000]. À chaque item, selon la réponse de l'élève, son niveau de compétence est réestimé et l'ordinateur propose un nouvel item dont la difficulté correspond à ce niveau. En positionnant sur la même échelle les paramètres de difficulté des items et les niveaux de compétences des élèves, les modèles de réponse à l'item sont particulièrement prisés dans le domaine des tests adaptatifs.

La principale contrainte de ce type de procédure est qu'il est nécessaire d'avoir estimé au préalable le niveau de difficulté d'un grand nombre d'items[10]. Cela suppose que chaque item ait été passé par un échantillon représentatif de la population visée, que sa difficulté ait été estimée et enregistrée dans une banque d'items, dans laquelle il sera possible de choisir le plus approprié lors de la procédure de test adaptatif. La constitution d'une telle banque implique un coût financier très important, qui limite la mise en pratique des tests adaptatifs[11].

Il existe d'autres stratégies d'adaptation, moins exigeantes. C'est le cas par exemple des procédures d'orientation (*multi-stage testing*) utilisées dans les enquêtes auprès des adultes IVQ et Piaac [Murat et Rocher, dans ce numéro, p. 83]. L'adaptation des items n'est pas faite individuellement mais pour des groupes de sujets déterminés en fonction de leurs résultats à un test d'orientation. Cette procédure est moins contraignante en pratique. Le recours à l'ordinateur n'est pas requis. Elle a l'avantage de pouvoir être appliquée pour une passation collective de tests papier-crayon, comme ce fut le cas par exemple avec les anciens tests de la Journée d'appel de

9. Pour cette approche, plusieurs approches ont été mises en concurrence, dont les modèles de réponse à l'item. D'ailleurs, après analyse, et pour des raisons pratiques, ce ne sont finalement pas les modèles de réponse à l'item qui ont été retenus mais une approche non paramétrique [Rocher, 2011]. En effet, les comparaisons de résultats entre les années pouvaient être réalisées directement après la passation, dans ces écoles, sur la base des scores observés (nombre de bonnes réponses). L'approche non paramétrique a ainsi permis d'établir des règles simples de passage entre les scores, permettant ainsi à chaque école d'assurer la comparabilité temporelle des résultats. Cela montre que les MRI, bien que très adaptés à ces problématiques, ne sont pas nécessairement incontournables et que d'autres méthodes sont envisageables, selon les contraintes des évaluations.

10. En faisant l'hypothèse sans doute assez forte que le niveau de difficulté de l'item existe indépendamment du test dans lequel il se situe.

11. Autre difficulté, il faut aussi que la réponse du sujet soit corrigée immédiatement, ce qui rend difficile le recours à un codage manuel et impose une procédure d'estimation des compétences intégrée à l'outil de collecte.

préparation à la défense [Rivière, De La Haye *et alii*, 2010]. Elle ne nécessite pas d'estimer au préalable la difficulté des items et donne potentiellement des résultats plus précis que ceux obtenus par un seul test, dans le cas où les niveaux de compétence sont très dispersés (cf. une application aux données d'IVQ : Murat et Rocher, 2009]. Au-delà des aspects pratiques, cette procédure se justifie également sur le plan théorique. Les dimensions cognitives intéressantes à évaluer ne sont pas forcément les mêmes selon les niveaux de compétences. Pour les personnes en difficulté face à l'écrit, il peut être intéressant d'insister sur les processus de bas niveaux comme le décodage des mots, alors que pour les autres personnes, différents aspects de la compréhension pourront être plus finement évalués. Ainsi, ce n'est pas seulement la difficulté du test qui est adaptée, mais la nature même de ce qu'il est censé mesurer.

Cahiers tournants

Nous présentons un autre cas pratique d'utilisation des MRI avec la méthode dite des « cahiers tournants ». Cette méthode est utilisée pour évaluer un nombre important d'items sans allonger le temps de passation. Elle consiste à répartir les items dans des cahiers différents qui comportent des items communs. Cette répartition doit répondre à certaines contraintes[12].

Par exemple, pour l'évaluation Cedre en sciences expérimentales de 2013 en troisième, l'équivalent de six heures et demie d'évaluation ont été créées. En effet, Cedre a pour objectif d'évaluer les acquis des élèves au regard des programmes scolaires. L'« univers » des items est donc très large. Le matériel a été réparti dans 13 blocs d'une demi-heure chacun. Ces blocs ont été ensuite répartis dans 13 cahiers différents, chaque cahier contenant 4 blocs. Ainsi, les élèves sont soumis à deux heures d'évaluation, ce qui est raisonnable.

La manière d'agencer les 13 blocs dans les 13 cahiers « tournants » répond à plusieurs contraintes :

- chaque bloc se retrouve le même nombre de fois au total, afin d'équilibrer le « poids » de chaque bloc ;
- chaque association de blocs (chaque paire) se trouve au moins une fois dans un cahier, afin de pouvoir calculer toutes les corrélations inter-items ;
- un bloc se retrouve à chacune des dispositions possibles : le bloc 1 apparaît en première position dans un des cahiers, en deuxième position dans un autre cahier, etc.

Le **tableau 1 p. 52** donne la répartition des blocs dans les cahiers, pour l'évaluation Cedre de troisième en sciences expérimentales en 2013. Le plan de rotation respecte les principes énoncés ci-dessus. Par ailleurs, cette évaluation est composée pour près de la moitié de blocs d'items repris de l'évaluation de 2007 afin d'établir des comparaisons. Les procédures d'estimation des MRI permettent facilement de gérer les valeurs manquantes aléatoires induites par la méthode des cahiers tournants. En outre, l'objectif est bien de rendre compte de la distribution des niveaux de compétences de manière globale, et non pas de manière individuelle, pour chaque élève, qui n'a pas passé les mêmes items que son voisin.

12. Cette méthode est en réalité une adaptation de procédures d'analyse de variance dans le cas de plans d'expérience incomplets [Cochran et Cox, 1950].

Tableau 1 Répartition des blocs dans les cahiers pour l'évaluation Cedre sciences expérimentales 2013

Cahiers	Séquence 1	Séquence 2	Séquence 3	Séquence 4
1	SVT 1*	SVT 3	SVT 4*	PHY B*
2	SVT 2	SVT 4*	SVT 5	PHY C
3	SVT 3	SVT 5	SVT 6*	PHY D*
4	SVT 4*	SVT 6*	PHY A	PHY E
5	SVT 5	PHY A	PHY B*	PHY F*
6	SVT 6*	PHY B*	PHY C	MIX*
7	PHY A	PHY C	PHY D*	SVT 1*
8	PHY B*	PHY D*	PHY E	SVT 2
9	PHY C	PHY E	PHY F*	SVT 3
10	PHY D*	PHY F*	MIX*	SVT 4*
11	PHY E	MIX*	SVT 1*	SVT 5
12	PHY F*	SVT 1*	SVT 2	SVT 6*
13	MIX*	SVT 2	SVT 3	PHY A

Note de lecture : le cahier 1 est composé de quatre blocs : SVT 1*, SVT 3, SVT 4* et PHY B*.
Les blocs étoilés sont les blocs repris de 2007.

HYPOTHÈSES

L'hypothèse d'unidimensionnalité

L'unidimensionnalité est une hypothèse fondamentale des modèles présentés précédemment. Seul le niveau de compétence θ explique la réussite à un item de difficulté et de discrimination données. Le respect de cette hypothèse est une condition préalable à la mise en œuvre de ces modèles. Si d'autres facteurs entrent en ligne de compte dans la probabilité de réussite aux items – par exemple une compétence différente de celle visée –, l'hypothèse d'unidimensionnalité doit être rejetée et le modèle ne peut être appliqué.

Bien que fondamentale, cette hypothèse est rarement testée statistiquement. Pour cause, la notion d'unidimensionnalité a longtemps souffert d'une absence de définition formelle. Ainsi, une quantité impressionnante d'indices ont été mis au point et visent à évaluer l'importance d'une dimension principale. Mais la plupart d'entre eux souffrent d'un manque de fondement théorique ainsi que de faiblesses techniques [Hattie, 1985]. Il faut attendre Stout [1987] pour poser une définition plus formelle de l'unidimensionnalité, à partir de la notion d'indépendance locale, c'est-à-dire l'indépendance des réussites entre deux items, conditionnellement à la dimension visée. En effet, là encore, si une corrélation est constatée entre items, après avoir contrôlé du niveau à l'ensemble du test, c'est qu'une deuxième dimension est intervenue dans la réussite à ces deux items. Notons que l'unidimensionnalité stricte n'existe probablement pas. Les processus mis en œuvre pour réussir un ensemble d'items sont complexes et varient selon les élèves et les contextes. Dès lors, il est difficilement concevable que ces processus se réduisent rigoureusement à une seule et même dimension [Goldstein, 1980]. C'est pourquoi, en pratique, évaluer l'unidimensionnalité revient en fait à évaluer l'existence d'une dimension dominante [Blais et Laurier, 1997][13].

13. Cela rejoint la démarche en analyse factorielle exploratoire qui consiste à comparer les valeurs propres des différents facteurs. D'ailleurs, les MRI peuvent être vus comme des analyses en facteurs communs [Rocher, 2013].

Les fonctionnements différentiels d'items

Nous l'avons évoqué avec le questionnaire sur la taille : un fonctionnement différentiel d'item (FDI) apparaît entre des groupes d'individus dès lors qu'à niveau égal sur la variable latente mesurée, la probabilité de réussir un item donné n'est pas la même selon le groupe considéré. La question des FDI est importante, car elle renvoie à la notion d'équité entre les groupes : un test ne doit pas risquer de favoriser un groupe par rapport à un autre. Ainsi, aux États-Unis, quantité de tests sont passés au crible dans le but de déterminer la présence d'éventuels biais d'items (*« Male/Female »*, *« Black/White »,* etc.) surtout si les résultats ont des conséquences sur le devenir des individus, comme pour les tests de sélection d'entrée à l'université, les tests de recrutement, etc. Les évaluations standardisées à grande échelle sont également concernées, en particulier les évaluations internationales qui doivent assurer la comparabilité des difficultés des items d'un pays à l'autre [Vrignaud, 2002]. C'est en effet l'hypothèse forte qui est faite dans le cadre des évaluations internationales : l'opération de traduction ne modifie pas la difficulté de l'item. Or, des analyses montrent que la hiérarchie de difficulté des questions posées est à peu près conservée pour des pays partageant la même langue, mais qu'elle peut être bouleversée entre deux pays ne parlant pas la même langue [Rocher, 2003].

Une définition formelle du FDI peut s'envisager à travers la propriété d'invariance conditionnelle : à niveau égal sur la compétence visée, la probabilité de réussir un item donné est la même quel que soit le groupe de sujets considéré. Formellement, un fonctionnement différentiel se traduit donc par :

$$P(Y|Z,G) \neq P(Y|Z) \qquad (10)$$

où Y est le résultat d'une mesure de la compétence visée, typiquement la réponse à un item ; Z est un indicateur du niveau de compétence des sujets ; G est un indicateur de groupes de sujets.

La probabilité de réussite, conditionnellement au niveau mesuré, est identique pour tous les groupes de sujets. En réalité, deux conditions sont nécessaires et suffisantes pour qu'un FDI se manifeste : l'item est sensible à une seconde dimension distincte de la dimension principale visée par le test et les groupes se différencient sur cette seconde dimension conditionnellement à la dimension principale. En guise d'illustration, considérons un item, dans une épreuve de mathématiques, qui nécessite la lecture d'un texte. Cet item est donc sensible à une dimension parasite. En outre, les filles ont de meilleures performances en lecture, et ce à niveau égal en mathématiques. L'item est fortement susceptible de présenter un fonctionnement différentiel selon le genre. Ce simple exemple permet d'entrevoir le lien entre dimensionnalité et fonctionnement différentiel, lien qui peut être formellement démontré [Rocher, 2013] et qui doit conduire à envisager les FDI de manière plus large que des indicateurs de biais.

Ainsi, une analyse de FDI qui intègre des éléments d'interprétation apporte des renseignements précieux au chercheur qui s'interroge sur les différences entre groupes de sujets, sur la dimensionnalité ou sur le caractère universel de certains concepts [Vrignaud, 2002]. Les biais ne sont alors plus envisagés comme des nuisances dans le processus de mesure, mais comme des éléments explicatifs, au service d'une démarche heuristique.

En pratique, de très nombreuses méthodes ont été proposées afin d'identifier les FDI. Ces méthodes ont chacune des avantages en matière d'investigation des différents

éléments pouvant conduire à l'apparition de ces FDI [Rocher, 2013]. Dans le cas des évaluations standardisées menées à la DEPP, il s'agit avant tout d'identifier les fonctionnements différentiels pouvant apparaître entre deux moments de mesure, s'agissant des items repris à l'identique. Dans ce cas, les différentes méthodes d'identification donnent des résultats relativement proches. Une stratégie très simple, employée dans Cedre, consiste donc à comparer les paramètres de difficulté des items repris, estimés de façon séparée pour les deux années. Si la difficulté d'un item a évolué, comparativement aux autres items, c'est le signe d'un fonctionnement différentiel, qui peut être lié par exemple à un changement de programmes ou de pratiques, comme nous le montrons dans l'illustration présentée dans la section suivante.

MÉTHODOLOGIE SUIVIE POUR LES ÉVALUATIONS CEDRE

En écho aux éléments théoriques exposés, nous présentons concrètement dans cette dernière partie la méthodologie suivie, en matière d'analyse psychométrique, par les évaluations Cedre. Cedre a pour objet de mesurer les acquis des élèves au regard des programmes scolaires, à partir d'évaluations réalisées par des échantillons représentatifs d'élèves, en CM2 et en troisième [voir Trosseille et Rocher, dans ce numéro, p. 15]. Chaque année, une discipline différente est évaluée et des comparaisons sont effectuées tous les cinq ou six ans.
L'exemple retenu est celui de l'évaluation des compétences des élèves de troisième en sciences expérimentales qui a établi une comparaison à six ans d'intervalle, entre 2007 et 2013. Les grandes lignes de la méthodologie employée sur les aspects psychométriques sont présentées. Pour plus de détails, le lecteur est invité à consulter le rapport technique disponible sur Internet [Bret, Garcia *et alii*, 2015].

Le matériel d'évaluation

En 2007, les élèves avaient passé 207 items au total dont 103 ont été repris pour l'évaluation de 2013 et 104 non repris. Cette sélection repose sur des critères statistiques ainsi que pédagogiques. En particulier, des items peuvent ne pas être retenus pour des raisons liées à l'évolution des programmes ou des pratiques.
En 2012, lors de l'expérimentation[14], 106 items ont été testés sur un échantillon d'environ 3 500 élèves. Après analyse, 72 items ont été retenus pour l'évaluation de 2013. Cette sélection repose principalement sur l'examen de statistiques descriptives concernant les items tels que la répartition des réponses données, le taux de réussite, le taux de non-réponse, le pouvoir discriminant (le « r-bis point »). Une vérification est faite quant à la précision des items sélectionnés selon le niveau de compétence[15], afin de s'assurer que le continuum est bien couvert.
Au final, en 2013, les élèves ont passé 175 items, dont 103 étaient des items repris de 2007 et 72 des items nouveaux. Ces items ont été répartis en 13 blocs, ventilés dans 13 cahiers selon le schéma présenté dans le tableau 1 : 7 blocs ont été repris à l'identique

14. Chaque évaluation est précédée d'une phase expérimentale l'année n-1.
15. D'un point de vue technique, la précision d'un item est l'inverse de la racine carrée de l'information de Fisher.

de l'évaluation de 2007 et 6 blocs nouveaux ont été intégrés en 2013.
Notons enfin que sur les 72 nouveaux items introduits en 2013, 32 items sont des questions ouvertes appelant une réponse rédigée et nécessitant la mise en œuvre de procédures standardisées de correction (supervision, corrections multiples, etc.). Chacun des deux formats de questions – QCM et questions ouvertes – présentent des avantages et des inconvénients : les premières forcent les choix de réponse mais garantissent l'objectivité du codage, tandis que les secondes permettent l'authenticité des réponses mais leur correction nécessite d'être très contrôlée [Vrignaud, 2003].

Les étapes

Les principales étapes de l'analyse psychométrique sont les suivantes :

1. Analyse « classique » des items
2. Étude de la dimensionnalité
3. Détection des fonctionnements différentiels d'items (avec le cycle précédent)
4. Étude de la qualité d'ajustement des items au modèle de réponse à l'item (MRI)
5. Application du MRI
6. *Equating* : ancrage avec le cycle précédent pour assurer la comparabilité des scores.

Suite à l'analyse « classique » menée sur l'ensemble des élèves (de 2007 et de 2013), 33 items ont été supprimés pour cause de mauvaise discrimination (r-bis < 0,2) : 19 items de 2007, 13 items communs et 1 de 2013. Il apparaît que cette suppression concerne pour l'essentiel des items construits en 2007, ce qui renvoie en effet à des niveaux de discrimination moins robustes pour cette évaluation, déjà observés en 2007 mais au-dessus du seuil de 0,3 à l'époque. En revanche, nous pouvons observer que l'expérimentation de 2012 a bien joué son rôle puisqu'un seul item présente une mauvaise discrimination. Au final, les analyses portent donc sur une évaluation composée de 85 items de 2007 non repris en 2013, de 90 items de 2007 repris en 2013 et de 71 items nouveaux en 2013.
L'étude dimensionnelle a montré une forte unidimensionnalité. Ainsi, sur les items passés en 2013, l'analyse factorielle des items sur la base des coefficients de corrélations tétrachoriques[16] a révélé une première valeur propre de 32,9 contre 3,6 pour la deuxième, ce qui témoigne de la présence d'une dimension principale prépondérante. En particulier, les items repris de 2007 et les items nouveaux de 2013 peuvent être considérés comme relevant d'une même dimension.
L'analyse des FDI a permis de détecter 5 items (la règle retenue est celle d'un écart de paramètres de difficulté β d'au moins 0,5) : 3 items en faveur de 2007, 2 items en faveur de 2013. Tous ces items sont des items de physique-chimie. Ils ont été éliminés des calculs. L'évolution des programmes est susceptible de produire des FDI. Ainsi, les 3 items présentant un FDI en défaveur des élèves de 2013 sont des items de physique-chimie portant sur la combustion. Or, par le biais de changements de programmes, il se trouve que la combustion n'est plus abordée en troisième. Si

16. Le coefficient de corrélation tétrachorique entre deux items est le coefficient de corrélation estimé entre les deux variables normales latentes qui conditionnent la réussite à chacun des items. Il est moins sensible aux effets seuil et plafond que le coefficient de corrélation linéaire, ou Φ , dans le cas d'items dichotomiques [Rocher, 1999].

ce type d'analyse peut souvent se révéler pertinent[17], il arrive qu'aucune explication ne soit trouvée à l'apparition de FDI.

Le calcul des scores

L'estimation des paramètres des items et des scores a été réalisée sur l'ensemble des élèves des deux années 2007 et 2013. Un modèle de réponse à l'item à deux paramètres a été employé. Ce choix se justifie par la variabilité des items en matière de pouvoir discriminant. Le modèle présente de bons critères d'ajustement aux données. D'ailleurs, les items présentent tous un indice dit de « FIT » acceptable, c'est-à-dire que leurs paramètres estimés permettent de rendre compte correctement des données.

Les scores estimés sont alors standardisés de sorte que les élèves de 2007 aient une moyenne de 250 et un écart-type de 50. Puis, la distribution des scores est « découpée » en six groupes de la manière suivante : nous déterminons le score-seuil en deçà duquel se situent 15 % des élèves (groupes 0 et 1), nous déterminons le score-seuil au-delà duquel se situent 10 % des élèves (groupe 5). Entre ces deux niveaux, l'échelle a été scindée en trois parties d'amplitudes de scores égales correspondant à trois groupes intermédiaires. Ces choix sont arbitraires et ont pour objectif de décrire plus précisément le continuum de compétence.

En effet, les modèles de réponse à l'item ont l'avantage de positionner sur la même échelle les scores des élèves et les difficultés des items. Ainsi, chaque item est associé à un des six groupes, en fonction des probabilités estimées de réussite selon les groupes. Un item est dit « maîtrisé » par un groupe dès lors que l'élève ayant le score le plus faible du groupe a au moins 50 % de chance de réussir l'item. Les élèves du groupe ont alors plus de 50 % de chance de réussir cet item.

À partir de cette correspondance entre les items et les groupes, une description qualitative et synthétique des compétences maîtrisées par les élèves des différents groupes est proposée. Ces principaux résultats sont présentés dans une *Note d'information* [Bret, Garcia, Roussel, 2014].

Perspectives

Les principes méthodologiques présentés sont aujourd'hui prédominants dans le domaine des évaluations standardisées. Ce type d'approche comporte cependant des limites. Par exemple, les modèles de réponse à l'item sont des outils puissants, d'un point de vue pratique, mais ils reposent sur des hypothèses fortes. En particulier, l'hypothèse d'unidimensionnalité est évidemment contestable lorsqu'on sait la multiplicité des compétences mises en jeu lors de la résolution d'une tâche.

Comme nous l'avons évoqué, des modélisations permettent de prendre en compte la multidimensionnalité, mais le plus souvent ces modèles sont multi-unidimensionnels, chaque item se rapportant à une seule dimension. C'est cette structure simple qui est le plus souvent considérée alors que c'est sans doute une structure complexe

17. Un autre exemple tiré de Cedre histoire-géographique et éducation civique en troisième entre 2006 et 2012 : les items proposés ayant trait à la connaissance des règles électorales ont présenté des FDI en faveur des élèves de 2012 par rapport aux élèves de 2006. En effet, l'évaluation de 2012 s'est déroulée au mois de mai, en pleine période d'élections.

qui prévaut. Des modèles existent et prennent en considération ces aspects : les modèles dits de classification diagnostique qui permettent d'établir des profils d'élèves à partir de leurs réponses et d'une analyse *a priori* des items selon un cadre théorique autorisant une structure complexe (chaque item est relié à un ensemble d'attributs que les élèves sont censés maîtriser pour réussir l'item).
Du point de vue des perspectives, notons enfin que l'avènement du numérique dans le domaine des évaluations standardisées amènera sans doute progressivement à reconsidérer les modélisations en cours, afin d'intégrer les « traces » laissées par les élèves lors de leur activité pendant l'évaluation.

Annexe – Procédures d'estimation des MRI

D'un point de vue statistique, les modèles de réponse à l'item peuvent être formulés de manière plus générale comme des analyses factorielles d'items ou encore comme des modèles multiniveaux, avec les items comme effets fixes et le niveau de compétence comme effet aléatoire [Goldstein, Bonnet et Rocher, 2007 ; Rocher, 2013]. Plusieurs méthodes ont cependant été spécifiquement développées pour estimer les paramètres du modèle. Baker et Kim [2004] les décrivent de manière précise. L'estimation est généralement conduite en deux temps : l'estimation des paramètres des items puis l'estimation des θ en considérant les paramètres des items comme fixes. Nous donnons des éléments concernant quelques méthodes, qui reposent sur la maximisation de vraisemblance, pour le modèle de Rasch et pour le modèle à deux paramètres.
Nous reprenons les notations des équations (7) et (8) p. 45 et 46 qui formulent la probabilité P_{ij} d'un élève i de répondre correctement à un item j, respectivement dans le cadre d'un modèle de Rasch et dans le cadre d'un modèle de réponse à l'item à deux paramètres, pour un item dichotomique.

Notons tout d'abord que les modèles présentés ne sont pas identifiables. Par exemple, dans le modèle à deux paramètres, les transformations $\theta_i^* = A\theta_i + B$, $b_j^* = Ab_j + B$ et $a_j^* = a_j / A$ avec A et B deux constantes ($A > 0$), conduisent aux mêmes valeurs des probabilités. Généralement, l'indétermination est levée en standardisant la distribution des θ (moyenne de 0 et écart-type de 1), ou bien dans le cadre du modèle de Rasch en fixant leur difficulté moyenne des items à 0.
Sous l'hypothèse d'indépendance locale des items, la fonction de vraisemblance s'écrit :

$$L(\mathrm{y},\boldsymbol{\xi},\theta) = \prod_{i=1}^{n}\prod_{j=1}^{J} P_{ij}^{y_{ij}}[1-P_{ij}]^{1-y_{ij}} \qquad (11)$$

où y est le vecteur des réponses aux items (*pattern*), $\boldsymbol{\xi}$ est le vecteur des paramètres des items.

Modèle de Rasch

Pour estimer les paramètres du modèle de Rasch, la procédure CML (*Conditional Maximum Likelihood*) peut être employée. Cette procédure consiste à conditionner la vraisemblance par le score observé à l'ensemble des items S_i (nombre de bonnes réponses), qui entretient une relation bijective avec θ_i. En effet, l'intérêt du modèle de Rasch, en matière d'estimation, résulte de ce qu'il définit un modèle exponentiel, au sens statistique du terme. Or, un modèle exponentiel admet une statistique exhaustive[18]. En l'occurrence, le score S_i est une statistique exhaustive pour le modèle de Rasch.

La connaissance d'une statistique exhaustive simplifie grandement la procédure d'estimation des paramètres : conditionner les pattern y_i par S_i permet en effet d'obtenir une vraisemblance indépendante de θ. La densité conditionnelle se calcule alors en utilisant le fait que S_i suit une loi multinomiale. Il s'agit alors d'un problème classique de maximisation de vraisemblance pour estimer les paramètres b_j.

18. Une statistique S(X), fonction de la variable (ou du vecteur) aléatoire X, est dite exhaustive si la loi de X conditionnellement à S(X) est indépendante des paramètres d'estimation.

Modèle à deux paramètres

Dans le cadre d'un modèle MRI à deux paramètres, la propriété d'exhaustivité du score observé n'est plus satisfaite. La procédure CML ne peut être appliquée. D'autres techniques d'estimation, plus coûteuses d'un point de vue algorithmique, doivent être employées.

Une première approche « naturelle » consisterait à annuler les dérivées de L par rapport aux paramètres du modèle, puis résoudre un système de $2J + n$ équations, par exemple avec une méthode itérative de type Newton-Raphson. Cette procédure appelée JML pour *Joint Maximum Likelihood* conduit cependant à des estimateurs biaisés. En effet, le nombre de paramètres augmente avec le nombre d'observations – un θ_i pour chaque observation –, ce qui ne correspond pas au cadre habituel des résultats sur les estimateurs sans biais convergents.

La procédure de maximisation de la vraisemblance marginale MML (*Marginal Maximum Likelihood*) permet de lever cette difficulté.

Estimation des paramètres des items (procédure MML)

La procédure MML consiste à estimer les paramètres des items en supposant que les paramètres des individus sont issus d'une distribution fixée *a priori* (le plus souvent normale). La maximisation de vraisemblance est *marginale* dans le sens où les paramètres concernant les individus n'apparaissent plus dans la formule de vraisemblance.

Si θ est considérée comme une variable aléatoire de distribution connue, la probabilité inconditionnelle d'observer un *pattern* y_i donné peut s'écrire :

$$P(\mathrm{y} = \mathrm{y}_i) = \int_{-\infty}^{+\infty} P(\mathrm{y} = \mathrm{y}_i|\theta_i)g(\theta_i)d\theta_i \qquad (12)$$

avec g la densité de θ.

L'objectif est alors de maximiser la fonction de vraisemblance :

$$L = \prod_{i=1}^{n} P(\mathrm{y} = \mathrm{y}_i) \qquad (13)$$

Cependant, l'annulation des dérivées de L par rapport aux a_j et aux b_j conduit à résoudre un système d'équations relativement complexe et à procéder à des calculs d'intégrales qui peuvent s'avérer très coûteux en termes de temps de calcul.

La résolution de ces équations est classiquement réalisée grâce à l'algorithme EM (*Expectation-Maximization*) impliquant des approximations d'intégrales par points de quadrature. L'algorithme EM est théoriquement adapté dans le cas de valeurs manquantes. Le principe général est de calculer l'espérance conditionnelle de la vraisemblance des données complètes (incluant les valeurs manquantes) avec les valeurs des paramètres estimées à l'étape précédente, puis de maximiser cette espérance conditionnelle pour trouver les nouvelles valeurs des paramètres. Le calcul de l'espérance conditionnelle nécessite cependant de connaître (ou de supposer) la loi jointe des données complètes. Une version modifiée de l'algorithme considère dans notre cas le paramètre θ lui-même comme une donnée manquante.

En outre, ce cadre d'estimation permet aisément de traiter des valeurs manquantes structurelles, par exemple dans le cas de cahiers tournants ou bien dans le cas de reprise partielle d'une évaluation.

Une fois les paramètres des items estimés, ils sont considérés comme fixes et il est possible d'estimer les θ_i, par exemple *via* la maximisation de la vraisemblance donnée par l'équation (10) p. 53. Les enquêtes internationales proposent quant à elles une méthode d'imputation multiple pour estimer les θ_i. Elles fournissent pour chaque élève un jeu de « valeurs plausibles », calculées selon une logique bayésienne, c'est-à-dire qui tient compte de l'information disponible par ailleurs, en l'occurrence celle des questionnaires de contexte [OCDE, 2012].

Pour finir, notons que les logiciels disponibles pour mener à bien ces calculs sont majoritairement des logiciels commerciaux dont le fonctionnement exact n'est pas très explicite. C'est pourquoi nous implémentons actuellement en interne ces procédures avec le logiciel libre R.

BIBLIOGRAPHIE

Andrich D., 2004, "Controversy and the Rasch model", *Medical Care*, vol. 42, No. 1, p. 1-7.

Baker F. B., Kim S.-H., 2004, *Item response theory – Parameter estimation techniques,* 2e ed., New York, Marcel Dekker.

Bertrand R., Blais J.-G., 2004, *Modèles de mesure – L'apport de la théorie des réponses aux items*, Sainte-Foy, Presses de l'Université du Québec.

Bret A., Garcia É., Rocher T., Roussel L., Vourc'h R., 2015, *Cedre, 2013 – Cycle des évaluations disciplinaires réalisées sur échantillons*, Rapport technique, MENESR-DEPP. Consultable en ligne : www.education.gouv.fr/methodologie-cedre.html

Bret A., Garcia E., Roussel L., 2014, « Cedre 2013 – Sciences en fin de collège : stabilité des acquis des élèves depuis six ans », *Note d'information*, n° 14-28, MENESR-DEPP.

Cochran W. G., Cox G. M., 1950, *Experimental designs,* New York, John Wiley and Sons.

Embretson S. E., Reise, S. P., 2000, *Item response theory for psychologists*, New Jersey, Lawrence Erlbaum Associates inc. publishers.

Falissard B., 2008, *Mesurer la subjectivité en santé – Perspective méthodologique et statistique,* 2e éd., Issy-les-Moulineaux, Elsevier-Masson.

Glas C. A. W., 2008, "Item response theory in educational assessment and evaluation", *Mesure et Evaluation en Education*, vol. 31, No. 2, p. 19-34.

Goldstein H., 1980, "Dimensionality, bias, independance and measurement scale problems in latent trait score models", *British Journal of Mathematical and Statistical Psychology*, vol. 33, No. 2, p. 234-246.

Goldstein H., Bonnet, G., Rocher T., 2007, "Multilevel structural equation models for the analysis of comparative data on educational performance", *Journal of Educational and Behavioral Statistics*, vol. 32, No. 3, p. 252-286.

Goldstein H., Wood R., 1989, "Five decades of item response modelling", *The British Journal of Mathematical and Statistical Psychology*, vol. 42, No. 2, p. 139-167.

Hattie J., 1985, "Methodology review : assessing unidimensionality of tests and items", *Applied Psychological Measurement*, vol. 9, No. 2, p. 139-164.

Kolen M. J., Brennan R. L., 2004, *Test equating, scaling and linking,* New York, Springer.

Laveault D., Grégoire J., 2002, *Introduction aux théories des tests en psychologie et en sciences de l'éducation,* Bruxelles, De Boeck.

MURAT F., ROCHER T., 2009, « Création d'un score global dans le cadre d'une épreuve adaptative », *Économie et Statistique*, n° 424-425, Insee, p. 149-178.

NEWTON P., SHAW S., 2014, *Validity in Educational and Psychological Assessment,* London, Sage Publications Ltd.

OCDE, 2012, *PISA 2009 – Technical Report*, Paris, OCDE.

PETERSON R. A., 1994, "Cronbach's Alpha Coefficient : A Meta-Analysis", *Journal of Consumer Research*, No. 21, p. 381-391.

RIVIÈRE J.-P., DE LA HAYE F., GOMBERT J.-E., ROCHER T., 2010, « Les jeunes Français face à la lecture : nouvelles pistes méthodologiques pour l'évaluation massive des performances cognitives », *Revue Française de Linguistique Appliquée*, n° 15, p. 121-144.

ROCHER T., 2013, *Mesure des compétences : les méthodes se valent-elles ? Questions de psychométrie dans le cadre de l'évaluation de la compréhension de l'écrit*, thèse de doctorat, Université Paris Ouest Nanterre La Défense.

ROCHER T., 2011, « Ajustement des évaluations nationales de CM2 (janvier 2009 - janvier 2010) », *Document de travail, série « Méthodes »*, n° 2011-M04, MEN-DEPP.

ROCHER T., 2008, « Lire, écrire, compter : les performances des élèves de CM2 à vingt ans d'intervalle (1987-2007) », *Note d'information*, n° 08.38, MEN-DEPP.

ROCHER T., 2003, « La méthodologie des évaluations internationales de compétences », *Psychologie* et *Psychométrie*, vol. 24, n° 2-3, p. 117-146.

STOCKING M. L., LORD, F. M., 1983, " Developing a common metric in item response theory", *Applied Psychological Measurement*, vol. 7, No. 2, p. 201-210.

STOUT W., 1987, "A non parametric approach for assessing latent trait unidimensionality", *Psychometrika*, vol. 52, No. 4, p. 293-325.

VRIGNAUD P., 2008, « La mesure de la littératie dans PISA : la méthodologie est la réponse, mais quelle était la question ? », *Éducation & formations,* n° 78, MEN-DEPP, p. 69-84.

VRIGNAUD P., 2003, « Objectivité et authenticité dans l'évaluation – Avantages et inconvénients des questions à choix multiples et des questions à réponses complexes : importance du format de réponse pour l'évaluation des compétences verbales », *Psychologie et Psychométrie*, vol. 24, n° 2-3, p. 147-188.

VRIGNAUD P., 2002, « Les biais de mesure : savoir les identifier pour y remédier », *Bulletin de Psychologie*, vol. 55, n° 6, p. 625-634.

LES ÉPREUVES STANDARDISÉES

Élément-clé du pilotage du système éducatif luxembourgeois

Christophe Dierendonck
Université du Luxembourg,
Unité de recherche *Education, Culture, Cognition and Society* (ECCS),
Institut *Lifelong Learning and Guidance* (LLLG)

Amina Kafaï
Ministère de l'Éducation nationale, de l'Enfance et de la Jeunesse (Luxembourg),
Service de Coordination de la Recherche et de l'Innovation pédagogiques et technologiques (SCRIPT),
Agence pour le développement de la qualité scolaire (ADQS)

Antoine Fischbach, Romain Martin et Sonja Ugen
Université du Luxembourg, *Luxembourg Centre for Educational Testing* (LUCET)

Jusqu'au début des années 2000, le pilotage de l'école luxembourgeoise s'opérait uniquement en référence aux *inputs* investis dans le système. Depuis lors, on assiste à une transformation progressive vers un pilotage par les *outputs* atteints par le système, basé sur la conduite d'évaluations externes des acquis des élèves et sur la mise en projet de développement de la qualité scolaire de tous les établissements scolaires. Dans ce contexte, depuis l'année scolaire 2008-2009, des évaluations externes des acquis des élèves, appelées « Épreuves Standardisées » (ÉpStan), sont conduites dans toutes les classes de grade 3 (CE2 en France) et de grade 9 (troisième en France), et le dispositif tend à s'élargir.
Ces épreuves sont commanditées par le ministère de l'Éducation nationale, de l'Enfance et de la Jeunesse (MENJE) et élaborées par le *Luxembourg Centre for Educational Testing* (LUCET) de l'université du Luxembourg.
Par ailleurs, l'Agence pour le Développement de la Qualité Scolaire (ADQS) a été créée en 2009 au sein du MENJE pour accompagner les établissements scolaires dans la définition et la mise en œuvre de leur projet de développement scolaire. Dans cet article, l'accent est mis sur la présentation des objectifs et de la méthodologie des ÉpStan ainsi que sur l'utilisation, par le LUCET et par l'ADQS, des résultats de ces épreuves à des fins de pilotage du système éducatif et de développement de la qualité scolaire. L'article se termine sur une présentation des défis futurs du LUCET et de l'ADQS.

Au Luxembourg, comme dans d'autres pays, la comparaison des systèmes éducatifs effectuée par l'enquête PISA 2000 a fait office d'électrochoc[1] et a permis une véritable prise de conscience quant à l'état de l'École. Progressivement, l'assurance et le développement de la qualité scolaire ont été placés au centre des préoccupations du ministère de l'Éducation nationale, de l'Enfance et de la Jeunesse (MENJE) et de son Service de Coordination de la Recherche et de l'Innovation pédagogiques et technologiques (SCRIPT). Le pilotage de l'école luxembourgeoise, qui s'opérait jusqu'alors uniquement en référence aux *inputs* investis dans le système (formation des enseignants, infrastructures à disposition, ressources financières, matériels, horaires et programmes, etc.), a ainsi initié sa transformation en direction d'un pilotage par les *outputs* atteints par le système (mesure des performances scolaires, vérification de l'atteinte des socles de compétence, publication des taux de redoublement, de *diplômation*, de décrochage scolaire, etc.).

En 2007, le MENJE a officialisé sa volonté d'élaborer des instruments d'évaluation externe des acquis des élèves, avec l'objectif de produire, de manière récurrente, des données fiables sur les compétences des élèves aux moments-clés du parcours scolaire [SCRIPT, 2007]. C'est dans ce contexte que depuis l'année scolaire 2008-2009, des évaluations externes[2] des acquis des élèves sont organisées. Appelées « Épreuves Standardisées » (ÉpStan), elles sont commanditées par le Ministère et élaborées par le *Luxembourg Centre for Educational Testing* (LUCET) de l'université du Luxembourg. On signalera, comme l'indique le **tableau 1**, qu'à côté des ÉpStan conduites actuellement aux grades 1, 3 et 9, existent deux autres types d'épreuves nationales d'évaluation des acquis des élèves qui sont élaborées au sein même du MENJE sans la collaboration de l'université : les épreuves communes de cycle 4.2. (grade 6) qui servent à orienter les élèves en fin d'école fondamentale et les épreuves communes de V^{e}ES /9^{e}EST (grade 9)[3] qui remplissent une fonction de bilan sommatif.

Tableau 1 Épreuves nationales d'évaluation des acquis des élèves organisées à l'heure actuelle au Luxembourg

Épreuves nationales	Niveaux d'études testés	
	Selon la nomenclature internationale	Selon la nomenclature française
ÉpStan de cycle 2.1.	Grade 1	CP
ÉpStan de cycle 3.1.	Grade 3	CE2
Épreuves commune de cycle 4.2.	Grade 6	Sixième
ÉpStan de V^{e}ES /9^{e}EST	Grade 9	Troisième
Épreuves commune de V^{e}ES /9^{e}EST		

1. Dans les trois domaines évalués par PISA, le Luxembourg s'est classé à l'antépénultième position du classement des pays.
2. Au sens strict, on pourrait argumenter que seules les ÉpStan de grade 9 (administrées sur ordinateur) sont véritablement des épreuves externes puisqu'elles sont élaborées, administrées et corrigées de manière externe. Ce statut d'« épreuve externe» peut cependant également être conféré aux ÉpStan de grade 1 et 3, car même si l'enseignant se voit confier la responsabilité d'administrer les épreuves et de les corriger, il a pu être montré empiriquement que des formats de question fermés corrigés à partir d'une grille de correction extrêmement détaillée (voir partie « la passation » et note de bas de page n° 12 p. 68) assurent une fidélité inter-correcteurs extrêmement élevée qui garantit au maximum la standardisation de la correction.
3. La nomenclature des années d'étude varie selon le type d'enseignement fréquenté (ES = enseignement secondaire de type général ; EST = Enseignement secondaire technique).

En 2009, une loi a ancré l'assurance de la qualité de l'enseignement dans les écoles fondamentales[4] et secondaires comme une des trois missions du SCRIPT. Ce texte légal précise que l'assurance de la qualité de l'enseignement doit être fondée, d'une part, sur des évaluations internes menées par le Ministère et, d'autre part, sur des évaluations externes du système éducatif conduites par des instituts universitaires (ex. : PISA, ÉpStan). Pour poursuivre cette mission et coordonner les actions en la matière, l'Agence pour le développement de la qualité scolaire[5] (ADQS) a été créée au sein du SCRIPT. Cette agence a pour tâche principale d'accompagner les écoles fondamentales (appelées écoles) et les écoles secondaires (appelées lycées) dans les processus de développement de la qualité scolaire qu'elles sont obligées (au fondamental) ou invitées (au secondaire) à mettre en place et à renouveler tous les trois ans. Concrètement, il est demandé à chaque établissement de rédiger un plan de réussite scolaire (PRS au fondamental) ou un plan de développement scolaire (PDS au secondaire) qui comprend quatre étapes : 1. analyse de la situation de l'école, 2. définition des objectifs à atteindre ; 3. mise en œuvre des actions et 4. suivi et bilan des actions menées. L'ADQS offre aux écoles un soutien scientifique pour mieux comprendre leur environnement scolaire (compilation de données relatives aux écoles du pays et à chaque école en particulier, préparation de questionnaires, aide à l'interprétation de données issues d'enquêtes ou d'évaluations nationales ou internationales, etc.) ainsi que des outils méthodologiques (cadre de référence pour le développement de la qualité scolaire, etc.) dans leur démarche de définition d'objectifs, de mise au point d'actions concrètes et d'évaluation de ces actions.

Dans cet article, l'accent est mis sur la présentation des objectifs et de la méthodologie d'un seul des trois types d'épreuves nationales (les ÉpStan) ainsi que sur l'utilisation, par le LUCET et par l'ADQS, des résultats ÉpStan à des fins de pilotage du système éducatif et de développement de la qualité scolaire. L'article se termine sur une présentation des défis futurs du LUCET et de l'ADQS.

LE DISPOSITIF DES ÉPSTAN

Objectif principal, objectifs secondaires et description générale du dispositif

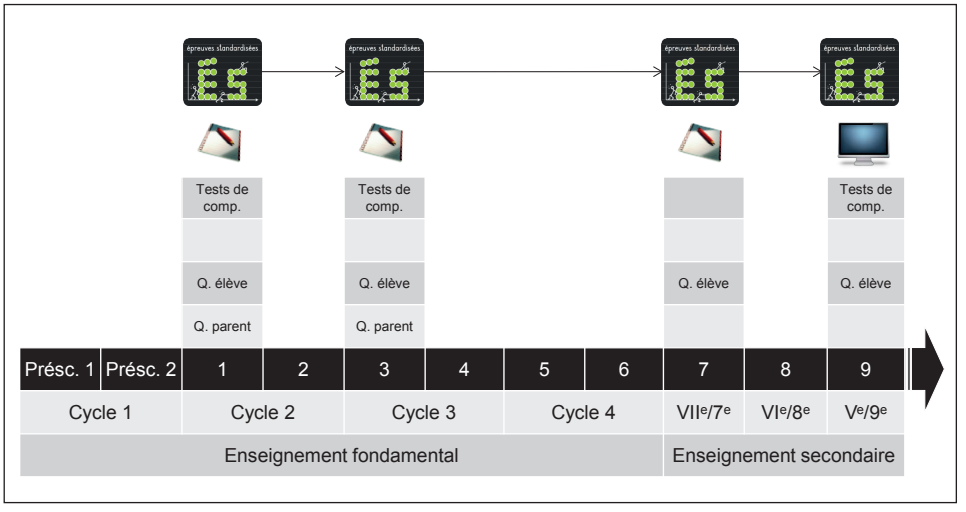
L'objectif principal du dispositif des ÉpStan est le *monitoring* du système scolaire et non son pilotage, qui relève de la responsabilité du MENJE. Les termes « monitoring » et « pilotage » ne sont en effet pas synonymes [voir à ce sujet, Dierendonck et Martin, 2008]. Le terme anglais suppose de collecter des informations fiables sur le fonctionnement et les résultats du système scolaire alors que le terme français implique, outre la prise d'informations, une action en vue de modifier une situation donnée en direction d'une situation souhaitée. Les prises d'information annuelles conduites au travers du dispositif ÉpStan autorisent non seulement des analyses

4. Les écoles fondamentales accueillent les enfants entre 4 et 12 ans (âges théoriques).
5. L'ADQS (https://portal.education.lu/qualitescolaire/Accueil.aspx), composée de 14 personnes (statisticienne, pédagogues, enseignants, psychologues), est une des trois divisions du SCRIPT. Les deux autres divisions du SCRIPT sont la cellule de compétence pour l'innovation pédagogique et technologique (INNO) et l'institut de formation continue (IFC).

transversales fondées sur des données fiables, mais également une comparaison des performances scolaires des élèves à travers le temps. Ces analyses, reprises dans un rapport national, sont censées fonder les décisions en matière de pilotage du système éducatif.

Actuellement, le dispositif concerne l'ensemble des élèves de l'enseignement public et privé (subventionné) en début des grades 1, 3, 7 et 9 ▸ **Figure 1**. Il comprend, selon l'année d'études considérée, des **tests de compétence** en compréhension de l'oral et de l'écrit (luxembourgeois au grade 1, allemand au grade 3, allemand et français au grade 9) et en mathématiques[6] présentés en version papier-crayon ou sur ordinateur, un **questionnaire bilingue (allemand/français) destiné aux élèves** portant sur les caractéristiques familiales (langue parlée à la maison, origine sociale, etc.) et sur des dimensions motivationnelles spécifiques (concept de soi, anxiété, intérêt, climat de classe, etc.) et un **questionnaire trilingue[7] (allemand, français, portugais) complété par les parents** (niveau d'études, profession, pays de naissance des parents). L'élaboration des tests de compétence et des questionnaires sont sous la responsabilité des chercheurs du LUCET. Les données récoltées sont ensuite mises à la disposition des acteurs concernés. Les informations portant sur les caractéristiques familiales sont principalement utilisées pour calculer les intervalles de performance attendue, mais elles peuvent également être utilisées à des fins de recherche. Le **tableau 2** donne une vue exhaustive des composantes et des particularités du dispositif. Les résultats obtenus par les élèves aux tests de compétence n'entrent pas en ligne de compte pour déterminer le parcours scolaire. À ce titre, ces tests sont à considérer comme des évaluations à faibles enjeux pour les élèves.
Les objectifs secondaires du dispositif, qui ont émergé progressivement au fil du temps, notamment pour répondre à certains besoins exprimés par les acteurs de

▸ **Figure 1** **Vue synthétique du dispositif ÉpStan actuel**



6. Dans le cadre du test de mathématiques au grade 9, les élèves ont en permanence la possibilité de voir les items en version allemande ou en version française.
7. À côté des trois langues officielles du pays (luxembourgeois, allemand et français), d'autres langues sont très utilisées au Luxembourg, notamment le portugais qui est parlé par plus de 20 % de la population.

terrain, sont de l'ordre du pilotage et du développement de la qualité au niveau des écoles, au niveau des classes et au niveau des élèves. Il s'agit notamment de mettre à la disposition des différents acteurs des rapports synthétisant les données issues des ÉpStan. Ces rapports présentent, pour chaque niveau d'agrégation possible des données (national, type d'enseignement, école, classe, élève), une comparaison normée qui tient compte des caractéristiques sociodémographiques des élèves. L'objectif de cette démarche multi-niveaux est triple : transmettre des signaux aux acteurs pour qu'ils amorcent des processus de développement scolaire au sein des écoles et des classes, renforcer la capacité diagnostique des enseignants en cernant les besoins individuels des élèves en termes d'apprentissage et amener les élèves et

Tableau 2 Vue exhaustive des composantes et des particularités du dispositif ÉpStan actuel

	Enseignement fondamental		Enseignement secondaire	
Grades	1	3	7	9
Nombre approximatif d'élèves concernés	5 100	5 100	5 500	6 600
Support	Papier-crayon	Papier-crayon	Papier-crayon	Ordinateur
Tests de compétence	– Mathématiques – Compréhension de l'oral (luxembourgeois) – Précurseurs de la lecture (par ex. conscience phonologique)	– Mathématiques – Compréhension de l'oral (allemand) – Compréhension de l'écrit (allemand)		– Mathématiques – Compréhension de l'écrit (allemand) – Compréhension de l'écrit (français)
Questionnaire élève	– Langues parlées à la maison – Concept de soi global – Concepts de soi dans les branches évaluées – Intérêt scolaire global – Intérêt pour les domaines scolaires évalués – Anxiété scolaire globale – Anxiété pour les branches évaluées – Besoin de cognition – Climat de classe – Attitude envers l'école	– Langues parlées à la maison – Concept de soi global – Concepts de soi dans les branches évaluées – Intérêt scolaire global – Intérêt pour les domaines scolaires évalués – Anxiété scolaire globale – Anxiété pour les branches évaluées – Besoin de cognition – Climat de classe – Attitude envers l'école	– Langues parlées à la maison – Pays de naissance – Concept de soi global – Concepts de soi dans les branches évaluées – Intérêt scolaire global – Intérêt pour les domaines scolaires évalués – Anxiété scolaire globale – Anxiété pour les branches évaluées – Besoin de cognition – Climat de classe – Attitude envers l'école – Profession des parents	– Langues parlées à la maison – Pays de naissance – Concept de soi global – Concepts de soi dans les branches évaluées – Intérêt scolaire global – Intérêt pour les domaines scolaires évalués – Anxiété scolaire globale – Anxiété pour les branches évaluées – Besoin de cognition – Climat de classe – Attitude envers l'école – Profession des parents
Questionnaire parents	– Profession – Niveau d'études – Pays de naissance	– Profession – Niveau d'études – Pays de naissance		

leurs parents à réfléchir et, le cas échéant, réagir aux résultats individuels obtenus. Dans des pays plus grands que le Luxembourg, ces différents objectifs sont souvent poursuivis dans le contexte de prises de données distinctes, avec notamment des évaluations par échantillon pour le niveau système. Or, au Luxembourg, étant donné la taille du pays, la participation de l'ensemble des élèves pour les différents niveaux est requise pour des raisons méthodologiques. Il a donc été décidé que les différents objectifs évoqués précédemment s'appuient sur une seule et même collecte de données.

À long terme, le dispositif devrait concerner les grades 1, 3, 5, 7 et 9 pour ce qui est des tests de compétence et des questionnaires. Ceci devrait permettre d'opérer un suivi longitudinal de cohortes d'élèves et d'étudier ainsi les trajectoires scolaires. Les effets à moyen et à long terme de projets pilotes et d'interventions spécifiques pourront également être évalués à partir de ce dispositif.

Communication autour du dispositif

La communication mise en place autour du dispositif est primordiale et se fait en étroite collaboration entre le LUCET et le ministère de l'Éducation pour faciliter l'acceptation des ÉpStan et la compréhension de leurs objectifs par les acteurs sur le terrain. En début d'année scolaire, le ministère de l'Éducation annonce de manière officielle les ÉpStan auprès des responsables des établissements scolaires, des enseignants et des parents d'élèves des grades concernés. Des réunions régionales d'information sont ensuite organisées par le Ministère et le LUCET pour permettre aux enseignants concernés d'interagir avec les mandataires et les concepteurs du dispositif. Le LUCET s'occupe de l'élaboration (et prochainement de la diffusion[8]) des informations concrètes autour de la passation des ÉpStan (documents explicatifs, manuels de test et de correction, tests de compétence, questionnaires, rapports, etc.). Pour garantir une transparence maximale, toutes les communications ponctuelles sont reprises sur un site web bilingue (www.epstan.lu) qui comprend en outre une série de vidéos explicatives, des exemples de questions et les rapports des années précédentes.

Élaboration des épreuves et des documents d'accompagnement

Les tests de compétence, composés d'une trentaine d'items par domaine scolaire évalué, sont élaborés par des groupes de travail dirigés par des chercheurs du LUCET spécialisés en psychologie, psychométrie, éducation ou didactique. Un groupe de travail compte, outre le chercheur responsable, de cinq à sept membres, en majorité des enseignants actifs[9] dans les grades et les branches scolaires concernés par le dispositif ÉpStan. Les réunions de travail se font en moyenne toutes les deux semaines. Une plateforme informatique permet aux membres des groupes de travail de partager, dans un environnement sécurisé, les documents utilisés et les items produits.

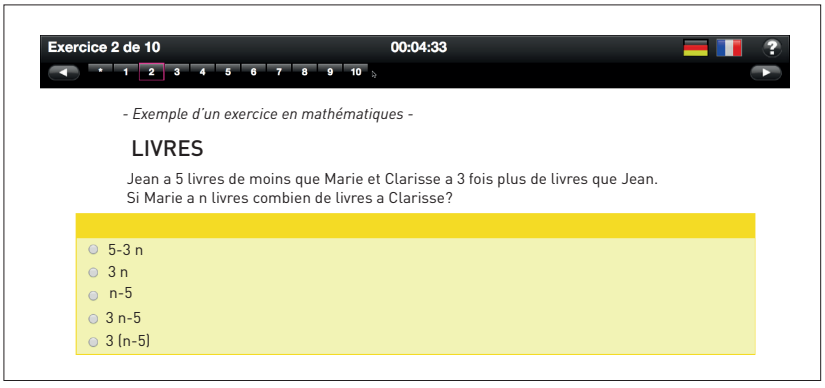
8. La diffusion des lettres d'information et du matériel de *testing* incombe à l'ADQS depuis 2008-2009, mais il est prévu qu'à terme, le LUCET soit responsable de cette tâche également.
9. Chaque année, le ministère de l'Éducation et le LUCET lancent un appel pour recruter des enseignants intéressés par le développement des épreuves. Ce travail est rémunéré en tant qu'heures supplémentaires ou par une décharge d'enseignement. Les enseignants reçoivent une formation portant sur le développement d'items (par ex. format de questions) et sur le cadre conceptuel défini pour élaborer les ÉpStan.

Sur la base des documents de référence officiels spécifiant les standards à atteindre à la fin de chaque cycle[10], un cadre conceptuel d'élaboration (*test framework*) a été développé pour chaque test de compétence. Ce document spécifie le nombre d'items nécessaires par types de compétences et par niveau de difficulté. Les différents groupes de travail développent et classifient les items élaborés en fonction de ce cadre conceptuel. Les items sont modifiés jusqu'à ce que les membres d'un groupe de travail tombent d'accord sur le contenu et les caractéristiques de chacun des items. La finalisation d'un item au niveau fondamental inclut l'illustration et la présentation dans le carnet de test ainsi que l'enregistrement des textes, consignes et items pour le test de compréhension de l'oral. Au niveau secondaire, les items finalisés sont transcrits sur la plateforme informatique OASYS (*Online Assessment SYStem*) développée par le LUCET ▸ **Figure 2**.

L'élaboration des items prend place durant toute l'année scolaire et aboutit, généralement en mai, à la conduite d'un pré-test du matériel. À l'issue de ce pré-test, les groupes de travail discutent à nouveau des items et les modifient si nécessaire en fonction des données empiriques récoltées. Les analyses statistiques utilisées lors du pré-test sont les mêmes que lors de la collecte de données principale.

Au niveau des épreuves de grade 9, trois versions sont élaborées pour chaque test de compétence. Ces versions présentent des degrés de difficulté différents qui correspondent aux différences de niveau scolaire existant entre les trois types d'enseignement au Luxembourg[11]. Pour que les résultats restent comparables entre les types d'enseignement, les versions de chaque test comprennent des items communs et des items spécifiques.
La compilation finale du test principal s'opère non seulement à partir du cadre conceptuel, mais également en tenant compte d'autres critères comme l'ordre des items, l'alternance des formats de réponse ou la place des items d'ancrage longitudinal et

▸ **Figure 2** **Capture d'écran d'un item de mathématiques présenté aux élèves de grade 9 sur la plateforme de *testing* OASYS**



10. Au niveau du secondaire, ces documents de référence restent au niveau descriptif des contenus à aborder (standards de contenus), mais ils ne définissent pas encore les critères minimaux à atteindre (standards de performance).
11. Chaque test se décline en une version pour l'enseignement secondaire (ES), une version pour l'enseignement secondaire technique orientation technique et polyvalente (EST) et une version pour l'enseignement secondaire technique orientation pratique et modulaire (EST-PR).

d'ancrage transversal. Pour garantir une comparabilité longitudinale des résultats, une sélection d'items est reprise d'une année à l'autre selon des critères statistiques spécifiques. Cet objectif de suivi longitudinal empêche la publication d'un grand nombre d'items.
Les chercheurs responsables des différents groupes de travail ont des échanges réguliers au sein d'une instance nommée « Conseil des développeurs d'items », dont la finalité est de garantir que les différents tests de compétence se structurent de manière similaire et que les items élaborés sont de qualité élevée. Cette instance permet également aux chercheurs de réfléchir aux développements potentiels du dispositif (de nouveaux types d'items par exemple) et d'échanger à propos des avancées scientifiques dans le domaine de l'évaluation des compétences scolaires.

Les préparatifs logistiques

Les préparatifs logistiques du dispositif ÉpStan s'opèrent en étroite concertation entre l'ADQS et le LUCET. Ils comprennent l'impression, l'envoi, la réception et l'archivage de tout le matériel nécessaire. Au niveau fondamental, le matériel inclut pour chaque élève : des carnets de test (en général, un carnet par test de compétence), un questionnaire motivationnel, un questionnaire parent, la lettre d'informations pour les parents ainsi qu'une feuille sur laquelle l'enseignant reporte le codage des réponses données. Chaque enseignant au fondamental reçoit un manuel de passation pour les tests, un manuel de passation pour le questionnaire élève et un manuel de correction pour chaque test de compétence. À titre d'illustration, environ 5 100 élèves du grade 3 répartis en 380 classes ont participé en 2013-2014 aux ÉpStan. Un carnet de test contenant entre 24 à 28 pages, la quantité de matériel imprimée peut être estimée à environ 840 000 pages pour le grade 3. Au grade 9, ce nombre est réduit étant donné que la passation se fait par ordinateur. Les établissements secondaires reçoivent cependant eux aussi des manuels de passation, des lettres d'information pour les parents et des lettres d'information pour les enseignants des branches concernées (au total environ 16 000 pages).

La passation

Les ÉpStan sont administrées au début de l'année scolaire en novembre[12] et évaluent les compétences acquises au cycle précédent. Puisque le retour aux acteurs de terrain se fait relativement rapidement (fin janvier), ceci permet aux titulaires des classes d'utiliser le retour d'informations fourni afin d'ajuster leurs interventions pédagogiques. Au fondamental, la passation et la correction se font par l'enseignant de la classe au départ d'un protocole d'administration du test et d'une grille de correction très détaillés supposés garantir une standardisation élevée[13]. Les tests de compétence sont répartis sur quatre matinées à des dates fixes. La durée de chaque test est de 50 minutes au grade 3 et de 35 minutes au grade 1.

Au secondaire, les trois tests de compétence sont administrés sur ordinateur lors d'une seule matinée de testing. La passation de l'épreuve est organisée dans chaque

12. Ce n'est qu'à partir de début octobre que l'on dispose des bases de données ministérielles permettant de définir de manière précise les participants aux ÉpStan. Les tests sont donc organisables, au plus tôt, en novembre.
13. En 2010-2011, une étude de fidélité inter-correcteurs a montré une correspondance très élevée (Kappa = 0,95) entre la correction réalisée par les enseignants et la correction effectuée par des correcteurs spécifiquement formés.

établissement par un coordinateur qui envoie les horaires de *testing* de chaque classe au LUCET afin de permettre un bon déroulement au niveau technique. Les jours de passation peuvent être choisis par chaque établissement à condition de se dérouler au mois de novembre. Chaque test de compétence dure 50 minutes, avec une pause de 15 minutes entre chaque test. Le dernier test est suivi du questionnaire élève dont la passation prend 20 minutes. L'ordre de présentation des trois tests de compétence est aléatoire. Pour soutenir les établissements, des assistants formés par l'équipe des ÉpStan sont présents lors de la première passation. Durant toute la période de *testing*, le LUCET assure une assistance continue par courriel et *via* une ligne téléphonique spécifique. L'ADQS est également contactée en cas de questions ou de problèmes rencontrés par les acteurs concernés par le dispositif ÉpStan.

Durant la période de *testing*, l'ADQS procède à des observations et à des entretiens dans quelques écoles afin de contrôler le respect des procédures et des consignes définies et de recueillir des informations en lien avec la passation proprement dite (conditions de passation dans la salle de classe, durée des épreuves, comportement des élèves pendant les épreuves, attitudes et perceptions des enseignants par rapport aux ÉpStan). Ce recueil d'informations qualitatives a pour objectif de contribuer à l'amélioration continue du dispositif.

Après réception du matériel de *testing* à l'université, les feuilles de codage et les questionnaires pour le fondamental sont scannés (*via* le logiciel Teleform) afin de permettre le traitement informatisé des données. Les bases de données sont ensuite vérifiées et nettoyées.

Analyse des données[14]

Scaling et ancrage longitudinal

Afin de permettre la comparabilité des résultats dans le temps et la conduite d'analyses de tendance, il est nécessaire de rendre compte des résultats issus des différents tests sur une échelle de performance commune. Deux cohortes d'élèves sont considérées chaque année : la cohorte 1 des élèves de l'année du test et la cohorte 0 composée de l'ensemble des élèves testés les années précédentes. La procédure adoptée dans le cadre des ÉpStan s'inspire de celle décrite par Nagy et Neumann [2010]. Cette procédure comprend cinq étapes successives : conformité au modèle de Rasch, analyse descriptive du fonctionnement différentiel des items (FDI), catégorisation des FDI selon la typologie définie par l'*Educational Testing Service* (ETS), analyse de sensibilité et estimation finale des paramètres de compétence des sujets.

Lors de l'étape 1, seules les données de la cohorte 1 sont analysées. Les items sont calibrés de manière libre et la contrainte d'identification est placée sur les sujets [Wu, Adams *et alii*, 2007]. On procède ensuite à la sélection des items jugés suffisamment discriminants (items qui présentent une corrélation satisfaisante avec le score global et qui permettent de distinguer les élèves ayant un score global élevé de ceux ayant un score global faible [Ebel et Frisbie, 1986]. On s'intéresse également à l'ajustement

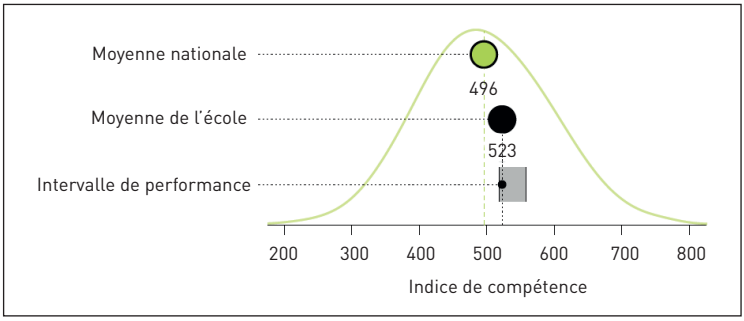
14. Pour plus de détails, se référer au rapport technique disponible sur http://www.epstan.lu.

des items au modèle de Rasch : seuls les items dont le comportement est bien prédit par le modèle sont conservés [Bond et Fox, 2010 ; Gustafsson, 1980 ; Martin-Löf, 1974 ; Wright, Linacre *et alii*, 1994].
L'étape 2 est une étape descriptive lors de laquelle les paramètres de difficulté des items d'ancrage longitudinal potentiels (c'est-à-dire les items communs aux cohortes 1 et 0) sont comparés graphiquement. Concrètement, les indices de difficulté estimés au départ de la cohorte 1 ainsi que les intervalles de confiance (à 95 %) qui y sont associés sont tracés en regard de ceux obtenus à partir de la cohorte 0. Ceci permet d'établir un diagnostic rapide de la robustesse des paramètres de difficulté des items au travers des cohortes et peut aider à identifier des items susceptibles de présenter un problème de fonctionnement différentiel.
Durant l'étape 3, on s'intéresse au fonctionnement différentiel des items (FDI) d'ancrage potentiels. Une classification de ces items est établie selon la magnitude du FDI en utilisant la typologie définie par l'*Educational Testing Service* (ETS) et reprise dans Nagy et Neumann [2010]. Trois types d'items sont ainsi distingués : les items de type ETS A (sans FDI), les items de type ETS B (avec un FDI modéré) et les items de type ETS C (avec un FDI important). Lors de cette étape, les données des deux cohortes sont mises en relation et l'on crée un tableau de données ancrées à partir duquel les compétences des élèves sont estimées [scores WLE, *Weighted Likelihood Estimation*, décrits dans Warm, 1989]. Les scores WLE ainsi qu'une variable indicatrice de cohorte (1 ou 0) sont ensuite introduites, pour chaque item, dans un modèle de régression logistique. Le coefficient de régression associé à la variable renseignant la cohorte permet d'apprécier le degré auquel la relation d'appartenance à une cohorte influence la probabilité de donner une bonne réponse à cet item, et par conséquent, est utilisé en tant que mesure d'intensité du FDI pour obtenir la classification mentionnée précédemment. Une autre classification, plus fine, basée également sur ce coefficient permet de distinguer 4 groupes d'items d'intensité croissante de FDI. En somme, ces classifications permettent de dégager 7 groupes d'items d'ancrage potentiels définissant chacun un scenario d'ancrage possible.
L'étape 4 consiste à mener une analyse de sensibilité sur la base des 7 scénarios d'ancrage définis lors de l'étape précédente, le but étant de conserver un maximum d'items d'ancrage tout en minimisant l'impact sur les compétences estimées des élèves de la cohorte 1. Au terme de cette étape, un ensemble d'items d'ancrage optimal est retenu.
Enfin, lors de l'étape 5, on procède à l'estimation finale de la compétence des élèves de la cohorte 1 en prenant soin de fixer les paramètres des items d'ancrage retenus à cet effet lors de l'étape précédente.

Calcul des intervalles de performance attendue pour disposer d'une comparaison équitable

Pour permettre une comparaison plus équitable des performances des écoles et des classes, des intervalles de performance attendue sont déterminés au départ de modèles de régression dans lesquels la performance agrégée (au niveau école ou au niveau classe) est prédite par les caractéristiques sociodémographiques agrégées des élèves (au niveau école ou au niveau classe) ▸ **Figure 3**. Les caractéristiques individuelles retenues sont la filière fréquentée (implicitement), le sexe, la ou les langues parlées à la maison, le statut socioéconomique (estimé au départ de l'indice de richesse à la maison, du nombre de livres dans le foyer et de l'occupation professionnelle des parents), l'année de naissance et le parcours scolaire antérieur effectué ou non au Luxembourg.

Figure 3 Exemple de graphique renseignant l'intervalle de performance attendue pour une école en particulier



Calcul des seuils de coupure pour déterminer les niveaux de compétence

L'analyse des données issues des ÉpStan ne se limite pas à offrir une comparaison normée qui tient compte des caractéristiques des populations scolaires. Elle fournit également une comparaison critériée[15] qui permet de situer chaque élève par rapport aux attentes du programme d'études. Ainsi, pour chaque test, les items sont répartis en différents niveaux de compétence. Cette répartition des items est réalisée d'abord de manière théorique au départ des référentiels, puis sur la base empirique suite à l'analyse des données du pré-test. À partir des paramètres de difficulté des items, un paramètre de difficulté médian est défini pour chaque niveau de compétence. C'est ce paramètre qui permet de calculer le score de coupure pour chaque niveau. Concrètement, si un élève est assigné à un niveau de compétence particulier, cela signifie qu'il y a une probabilité élevée que cet élève réussisse au moins la moitié des items classés dans ce niveau de compétence. Cette « probabilité élevée » a été opérationnalisée à 0,62, à l'instar de ce qui se fait dans le programme PISA [OCDE, 2009, p. 300].

L'UTILISATION DES DONNÉES ISSUES DES ÉPSTAN

Utilisation des données par l'université

Une fois que les données des ÉpStan ont été analysées, différents rapports sont mis à la disposition des acteurs concernés[16] sur une plateforme informatique du MENJE. Sur le site www.epstan.lu, outre des exemples de rapports, on peut trouver deux documents explicatifs : un document technique exhaustif qui détaille la méthodologie du dispositif d'évaluation externe et un document explicatif destiné aux directions et aux enseignants.

15. Décrire ce dont un élève ou un étudiant est capable, sans égard à la performance des autres, tel est le but visé avec la mesure à interprétation critériée [Scallon, 2004, p. 2].
16. Grade 1 et 3 : environ 15 300 rapports élève (1 rapport par domaine scolaire testé), 360 rapports classe et 160 rapports école. Grade 7 : résultats du questionnaire uniquement présentés dans le rapport national. Grade 9 : environ 19 000 rapports élève (un rapport par domaine scolaire testé), 1 000 rapports classe et 35 rapports école.

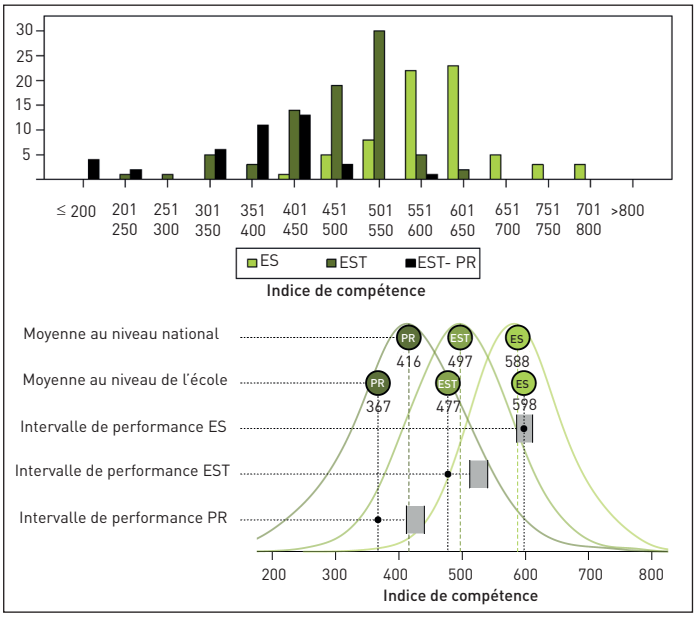
Le « rapport national » – pour les décideurs et pour le public

Le « rapport national », publié tous les trois ans, est rédigé par le LUCET et rendu public sur le site www.epstan.lu. Il présente les objectifs et la méthodologie des ÉpStan ainsi que plusieurs types d'analyse des données permettant la comparaison, mais de manière anonyme : une analyse globale des résultats observés, une analyse des résultats qui tient compte des facteurs socio-économique et migratoire et une analyse comparant les résultats obtenus dans les filières d'enseignement. Le « rapport national » se termine par une analyse synthétique de la situation actuelle et par la présentation de différentes pistes d'action. Le rapport 2011-2014 présente, pour la première fois, une analyse des tendances observées ces trois dernières années scolaires.

Le « rapport école » – pour les comités d'écoles fondamentales et les directions des écoles secondaires

Un « rapport école », envoyé aux comités d'écoles fondamentales, à chaque direction d'établissement secondaire et à l'ADQS, présente de manière globale les résultats des élèves de l'école concernée et autorise une comparaison anonyme avec les résultats observés ailleurs dans le pays et au sein des différentes filières d'enseignement (ES, EST, EST-PR) ▸ **Figure 4**. Un intervalle de performance attendue est calculé pour chaque école. Cet intervalle permet une comparaison qui tient compte des caractéristiques principales de la population scolaire accueillie dans l'école (niveau socio-économique et origine migratoire). Le « rapport école » renseigne aussi les indicateurs moyens observés au niveau école pour les différentes dimensions motivationnelles étudiées au travers du questionnaire (concept de soi, anxiété, intérêt, climat de classe, etc.) et permet une comparaison avec les moyennes observées sur le plan national.

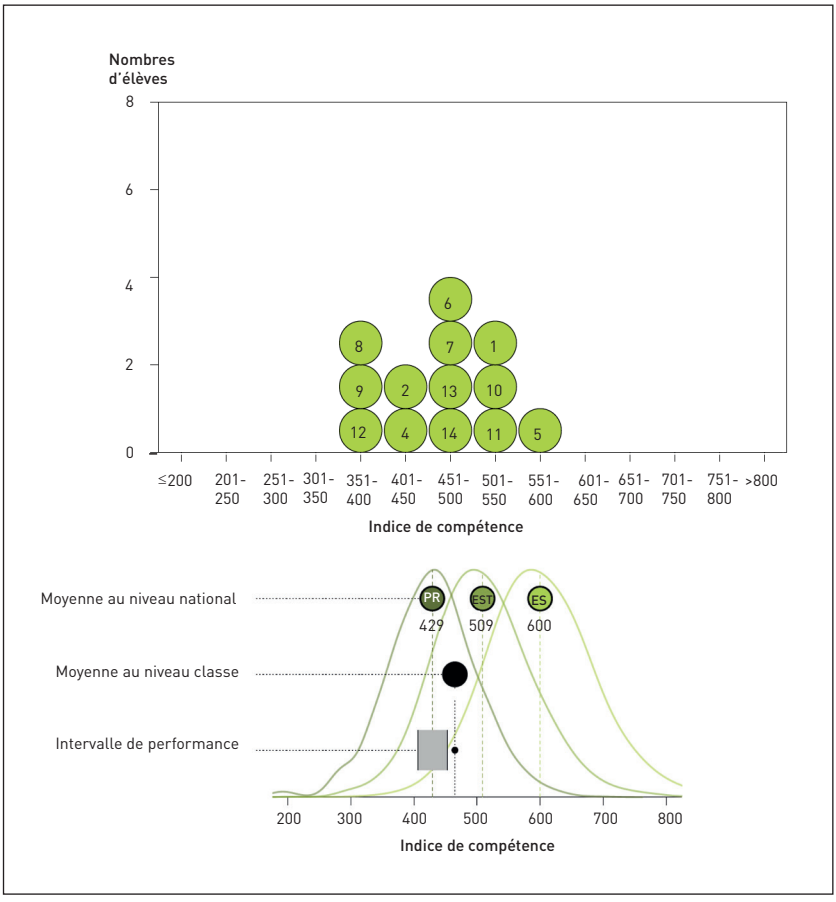
▸ **Figure 4** **Principaux graphiques issus du rapport adressé à un établissement secondaire en particulier**



Le « rapport classe » pour les enseignants

Un « rapport classe » est mis à la disposition des enseignants concernés par les ÉpStan sur une plateforme de téléchargement du MENJE avec code d'accès personnalisé ▸ **Figure 5**. Ce rapport présente, notamment, la distribution des résultats des élèves de la classe (l'enseignant dispose de la clé de codage permettant d'identifier précisément ses élèves) et permet de situer ceux-ci par rapport aux résultats observés au niveau du pays et au niveau des différentes filières d'enseignement. Un intervalle de performance attendue est calculé pour chaque classe. Cet intervalle permet une comparaison qui tient compte des caractéristiques principales de la population scolaire accueillie dans la classe (niveau socio-économique et origine migratoire). Le « rapport classe » renseigne également les moyennes observées au niveau classe pour les différentes dimensions motivationnelles étudiées au travers du questionnaire (concept de soi, anxiété, intérêt, climat de classe, etc.) et permet une comparaison avec les indicateurs moyens nationaux.

▸ **Figure 5** **Principaux graphiques issus du rapport adressé à une classe de secondaire en particulier**



Le « rapport élève » – pour les élèves et leurs parents

Un « rapport élève » est rédigé, pour chaque élève, dans les trois domaines de compétence évalués par les ÉpStan ▸ **Figure 6**. Les enseignants téléchargent ces rapports individualisés en même temps que le « rapport classe ». Ils ont la responsabilité de mettre ces rapports à disposition des élèves et des parents. Ces rapports individuels décrivent la performance individuelle des élèves et permettent une double comparaison (normative et critériée). La comparaison normative consiste à renseigner le score obtenu par l'élève en le situant sur un graphique par rapport à la moyenne de la classe et par rapport aux distributions et scores moyens observés dans chaque filière d'enseignement. La dimension critériée apparaît dans la partie droite du rapport élève. Le score obtenu par l'élève permet de situer celui-ci dans un niveau de compétence qui décrit le type de questions ou le type de tâches face auxquelles les élèves donnent généralement une réponse correcte. Enfin, le rapport renseigne le temps que les élèves ont effectivement passé à faire le test.

Utilisation des données par l'ADQS

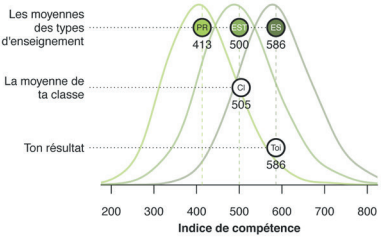
Pour rappel, l'ADQS est chargée d'accompagner les établissements scolaires dans la démarche de développement de la qualité scolaire. Le travail de l'ADQS se décline essentiellement en trois types d'actions : accompagner les écoles dans l'élaboration,

▸ **Figure 6** **Exemple de rapport élève (compréhension de l'écrit en français, enseignement secondaire)**

Français – Compréhension de l'Écrit

Résultats des Épreuves Standardisées 2013/2014
LTX, 9TE1, élève no. 12

Ton résultat : Tu as atteint un résultat de 586. Ton résultat est au-dessus de la moyenne de ta classe et au-dessus de la moyenne des élèves de l'enseignement secondaire technique théorique et polyvalent. Le graphique suivant illustre ces résultats.



ES : enseignement secondaire
EST : enseignement secondaire technique théorique et polyvalent
PR : enseignement secondaire technique pratique et préparatoire

Les courbes en arrière-fond du graphique montrent la distribution des indices de compétence dans les trois types d'enseignement.

Tu as travaillé pendant 45 minutes sur une durée maximale de 50 minutes.

Explications (d'après le modèle de compétences de Kühn, 2008)

Avec un résultat **supérieur à 663**, les élèves sont en général capables de comprendre des textes complexes, continus ou discontinus (p. ex. des articles extraits de magazines pour jeunes ou des textes littéraires, des textes factuels complexes ou des romans de la littérature pour la jeunesse). Ils sont aptes à travailler sur des textes de manière autonome, en se fixant des objectifs et en utilisant des stratégies de lecture variées dans le but d'explorer de nouvelles thématiques et d'acquérir des connaissances. Ils sont capables d'analyser le langage dans différents types de textes et de donner un avis personnel sur les contenus.

Avec un résultat **entre 612 et 663**, les élèves sont en général capables de comprendre des textes assez complexes de par leur structure, le langage utilisé et la thématique abordée (p. ex. des textes informatifs, descriptifs ou explicatifs dans un lexique ou une encyclopédie), des informations et des commentaires ou des textes littéraires plus courts (p. ex. des récits). Ils sont capables d'utiliser de manière autonome des techniques et des stratégies de lecture. Ils savent faire des déductions et les justifier. Ils savent replacer dans le contexte et commenter ce qu'ils ont lu.

Avec un résultat *entre 550 et 611*, les élèves sont en général capables de comprendre des textes informatifs simples, rédigés essentiellement dans un langage courant, relatifs aux domaines qui leur sont familiers ou qui les intéressent, tout comme des textes littéraires (p. ex. des contes, de courts articles ou reportages extraits de la presse, des publicités). Ils sont capables de repérer des informations et de les rendre avec leurs propres mots, d'établir des liens entre différentes parties d'un texte et entre les paragraphes, de faire des déductions simples et de commenter un texte de manière rudimentaire. Ils sont capables de catégoriser des textes, de reconnaître la fonction d'un texte et son objectif visé.

Avec un résultat **entre 448 et 549**, les élèves sont en général capables de comprendre des textes simples et familiers, relatifs à leur environnement immédiat (p. ex. des récits simples, des lettres personnelles, des textes factuels simples extraits de brochures d'information, de journaux ou de magazines, ainsi que des recettes ou des petites annonces). Ils sont capables de reconnaître les caractéristiques globales d'un texte (p. ex. repérer l'auteur, le type de texte, la thématique, les circonstances, les acteurs concernés), de localiser et de citer des informations explicitement contenues dans le texte (qui, quoi, où ?).

Avec un résultat **inférieur à 448**, les élèves échouent à la plupart des questions du test. Ceci ne signifie pas pour autant qu'ils n'ont aucune compétence en compréhension de textes en français, mais le document de référence utilisé et le test proposé ne permettent pas d'être plus précis.

www.epstan.lu

la conduite et l'évaluation de leur plan de réussite scolaire (PRS) ou de leur plan de développement scolaire (PDS), documenter les écoles en leur envoyant les données quantitatives et qualitatives qui les concernent et intervenir sur demande des établissements scolaires pour soutenir ou encadrer des projets spécifiques.
Seules les données produites au niveau école à l'issue des ÉpStan sont utilisées par l'ADQS dans le cadre de la documentation des écoles sur leurs situations respectives. Concrètement, chaque école compile – dans un « classeur école » ou dans un « classeur lycée » structuré en quatre parties (données démographiques, performances scolaires, processus scolaires, perceptions des acteurs) – les informations que l'ADQS lui envoie dès que celles-ci sont disponibles. Les résultats obtenus par l'établissement scolaire lors des ÉpStan sont insérés dans la partie du « classeur école » ou du « classeur lycée » consacrée aux performances scolaires de l'établissement.

LES DÉFIS POUR LE FUTUR

Défis pour l'université

Élargir le dispositif

Actuellement, les ÉpStan sont administrées au grade 1, au grade 3 et au grade 9. Au grade 7, seul le questionnaire motivationnel est soumis aux élèves. Un modèle « 1-3-5-7-9 » complet (tests de compétence et questionnaire) est actuellement proposé aux décideurs politiques pour disposer, à long terme, en début de chaque cycle d'apprentissage, d'une évaluation des acquis des élèves développés lors du cycle précédent et permettre un suivi longitudinal de cohortes d'élèves. Afin de disposer de profils de compétence plus larges, il est aussi question d'évaluer d'autres domaines disciplinaires (les sciences notamment) et d'autres processus (le raisonnement notamment). Face à ces ambitions partagées par le LUCET et le MENJE, la question de la faisabilité d'un tel dispositif se pose car le dispositif actuel, conduit annuellement rappelons-le, nécessite déjà une quinzaine de chercheurs engagés à temps plein. En outre, les données collectées actuellement pourraient être exploitées davantage (ex. : données comportementales recueillies lors des passations, etc.), voire partagées, sous certaines conditions, avec des chercheurs externes au LUCET.

Questionner la perception, la compréhension et l'utilisation réelle des rapports ÉpStan

Au fil du temps, les routines de communication, d'organisation, de collecte, d'analyse des données et de rédaction des rapports produits dans le cadre des ÉpStan ont progressivement été affinées et la plateforme informatique OASYS utilisée pour le *testing* dans les écoles secondaires apparaît très fiable et très flexible.
Quant aux rapports envoyés aux différents acteurs, ils connaissent chaque année des développements plus ou moins marqués, tant au niveau de la forme que des contenus produits. Ainsi, la perspective initiale quasi-exclusivement psychométrique ne permettant pas d'identifier les élèves et leurs lacunes (jusque 2009-2010) s'est transformée progressivement en une perspective mixte (depuis 2010-2011) recherchant un équilibre entre comparaisons normatives et informations plus

diagnostiques faisant davantage référence au curriculum scolaire et autorisant une identification des élèves en difficulté. C'est aussi en 2010-2011 que le rapport élève apparaît pour la première fois. Cette transformation progressive a été décidée à la demande du terrain pour augmenter l'utilité des épreuves, des données récoltées et des rapports envoyés aux enseignants et aux élèves[17]. Sur ce point, nous n'avons que très peu d'informations fiables sur la façon dont les acteurs se saisissent des rapports mis à leur disposition.

Ces questions liées à la perception du dispositif, à la compréhension et à l'utilisation des rapports par les acteurs sont des questions difficiles car, jusqu'ici, peu d'informations étaient disponibles à ce sujet. Il a donc été décidé de conduire, durant le premier semestre 2014, une enquête par questionnaire et des discussions en groupe auprès des différents acteurs concernés par les ÉpStan (enseignants, directions, coordinateurs, parents, élèves), ceci afin d'examiner dans quelle mesure les rapports mis à disposition sont lus, compris et utilisés par les différents publics-cibles. Les informations récoltées sont en cours d'analyse et seront discutées en 2015 dans le cadre d'une démarche qualité menée en interne. Plus largement, l'étude de l'impact réel du dispositif ÉpStan dans le cadre du pilotage du système scolaire luxembourgeois semble constituer un défi majeur pour le futur.

Développer de nouveaux types d'items d'évaluation

Les ÉpStan sont actuellement uniquement constituées d'items à choix multiple et de questions ouvertes à réponse courte, ceci afin de garantir la plus grande objectivité lors de la correction des réponses des élèves. Mais il est question de profiter des possibilités offertes par la plateforme informatique de *testing* OASYS pour expérimenter de nouveaux formats d'items mettant en œuvre des processus ou des comportements spécifiques (prise de notes durant la lecture, surligner des passages, glisser-déplacer avec la souris, etc.).

Assurer la validité des constats dressés

Les ÉpStan ont clairement une fonction informative puisque les performances des élèves à ces tests ne sont pas prises en compte pour influencer ou déterminer leur parcours scolaire. On cherche en outre à réduire au maximum l'anxiété que ces tests pourraient provoquer chez les élèves.

Face à ce genre d'évaluations « à faibles enjeux » pour les élèves, on formule généralement l'hypothèse implicite que les performances observées lors des tests ont été réalisées avec un niveau d'effort élevé [Wolf et Smith, 1995 ; Zerpa, Hachey *et alii*, 2011]. À ce sujet, les résultats de la recherche sont contradictoires : certaines études ont mis en évidence que les élèves sont raisonnablement motivés à donner le meilleur d'eux-mêmes quand le test est à faibles enjeux [Baumert et Demmrich, 2001] tandis que

17. On soulignera que cette transformation a également engendré un paradoxe puisque l'objectif principal du monitoring du système empêche d'aller au bout de la logique au niveau de l'élève et au niveau de l'enseignant, en ne rendant publics que quelques items utilisés dans les tests. Cela dit, le nombre d'items rendus publics s'accroît chaque année et une banque d'items illustrant les différents niveaux de compétence sera rapidement constituée.

d'autres ont conclu que les enjeux du test exerçaient effectivement une influence significative positive sur la motivation et la performance [Wolf et Smith, 1995]. Les élèves seraient donc plus ou moins motivés et pourraient dès lors s'impliquer différemment. Il est donc possible de questionner la validité [au sens de Messick, 1995] des constats dressés à l'issue des évaluations « à faibles enjeux », puisqu'on peut faire l'hypothèse que, lors de ce type d'épreuves dépourvues de conséquences scolaires pour les élèves, l'effort consenti par certains n'est pas maximal et que, *de facto*, leurs performances ne peuvent véritablement être considérées comme des indicateurs valides de leurs acquis. Plus globalement, les scores moyens observés aux évaluations externes à faibles enjeux seraient dès lors sous-estimés et ne reflèteraient donc pas ce dont les élèves sont réellement capables, mais plutôt « ce qu'ils démontrent avec un effort minimal » [O'Neil, Sugrue, Baker, 1996].
Pour examiner dans quelle mesure les ÉpStan sont touchées par ce problème éventuel de validité, des données de motivation initiale pour les tests et d'effort consenti durant ceux-ci [voir à ce sujet Keskpaik et Rocher, 2012 et Dierendonck, Sonnleitner *et alii*, 2013] sont recueillies depuis deux années. Ces données font actuellement l'objet d'une analyse minutieuse.

Défis pour l'ADQS

De la méfiance à la confiance, en passant par l'indifférence

La dynamisation des établissements scolaires *via* les PRS et les PDS et la création de l'ADQS en 2009 ont initialement été perçues par une partie du personnel enseignant comme des tentatives de contrôle du travail accompli dans les écoles ou d'évaluation des enseignants eux-mêmes[18]. Mais peu à peu, par un travail intense de simplification des procédures et de communication/explicitation des objectifs et des rôles de chacun mené auprès de chaque établissement scolaire, les tensions se désamorcent et les projets de développement se mettent en place, certes pas partout avec la même facilité, le même entrain ou les mêmes ambitions, loin de là, mais toutes les écoles au Luxembourg sont « en mouvement ». Le défi de l'ADQS consiste à présent à trouver les moyens d'un accompagnement différencié, chevillé aux besoins particuliers des écoles. L'idée est d'offrir de nouvelles formations, de nouveaux outils et de rechercher des personnes-relais (comme les inspecteurs et les instituteurs-ressources à l'école fondamentale) pouvant jouer un rôle démultiplicateur, car l'ADQS demeure une petite structure de 14 personnes qui est chargée d'accompagner plus de 150 écoles fondamentales et 35 établissements secondaires.

À la recherche d'une complémentarité entre toutes les évaluations (internes et externes, nationales et internationales)

À côté des évaluations faites par les enseignants (devoirs en classe, bilans intermédiaires) et des enquêtes internationales comme PIRLS et PISA, il y a actuellement, comme énoncé en introduction, trois types d'évaluations « externes » nationales des

18. Ce sentiment a été exacerbé par le contexte politique du projet de réforme de la fonction publique et de l'évaluation des fonctionnaires (dont les enseignants font partie).

acquis des élèves au Luxembourg coordonnées par l'ADQS : les épreuves communes de cycle 4.2 (grade 6)[19], les épreuves communes de Ve/9e (grade 9)[20] et les ÉpStan (grades 1, 3 et 9). Alors que les deux premiers types d'évaluations remplissent une fonction spécifique (orientation des élèves pour les épreuves communes de grade 6 et bilan des acquis pour les épreuves communes de grade 9), les ÉpStan tentent de conjuguer des fonctions différentes et de poursuivre des objectifs multiples, parfois difficilement conciliables. Toutes ces évaluations sont en outre construites par des acteurs différents avec des cadres de référence[21] et d'élaboration différents, ce qui engendre une certaine confusion chez les enseignants. Le défi de l'ADQS serait de clarifier la situation et de travailler à l'articulation et à la complémentarité des différentes évaluations des acquis des élèves présentes au sein du système scolaire, quitte à repenser les fonctions et les caractéristiques de certains dispositifs.

Amener les établissements scolaires à utiliser le « classeur école » et le « classeur lycée » pour définir leur PRS ou leur PDS

L'ADQS a pu constater que les informations qu'elle met à la disposition de chaque équipe pédagogique dans le cadre du « classeur école » ou du « classeur lycée » demeuraient sous-exploitées au moment de définir leur projet trisannuel de développement scolaire. En particulier, pour identifier d'éventuels problèmes et formuler des pistes de développement possible, il n'y aurait que très peu de tentatives d'analyse et de mise en relation des informations rassemblées au sein des quatre parties de l'outil (données démographiques, performances scolaires, processus scolaires, perceptions des acteurs). Il s'agit donc pour l'ADQS de mener une analyse de l'outil proposé et d'envisager des pistes pour faciliter la lecture et la mise en relation des données mises à disposition. L'idée est aussi de compléter l'outil actuel avec d'autres données (de contexte et de performance) susceptibles de mettre en évidence la « valeur ajoutée » de chaque établissement scolaire.

Ne pas laisser les enseignants seuls avec des constats

Si la mise à disposition de données comparatives fiables est un élément essentiel du pilotage et du développement de la qualité du système scolaire, il ne suffit pas d'établir des constats pour permettre aux acteurs de les dépasser. Sur le plan des pratiques pédagogiques par exemple, on peut supposer que des pistes didactiques, formulées par des experts disciplinaires (de la cellule de l'innovation du SCRIPT et/ou de l'université) au départ des résultats des évaluations externes, seraient bien accueillies par

19. Ces épreuves ont lieu au courant du deuxième trimestre de l'année scolaire, à des dates déterminées, en disposant d'un temps bien défini et avec un document reprenant des consignes de passation. Les types d'items sont variables d'une année à l'autre et sont corrigés par l'enseignant. Les résultats permettent de comparer chaque élève à l'ensemble des élèves ayant passé l'épreuve. Ils sont utilisés dans la procédure d'orientation des élèves vers l'enseignement secondaire.

20. Ces épreuves, élaborées par le MENJE, portent exclusivement sur le contenu des programmes des branches concernées et traitées en classe et n'exigent, dès lors, aucune préparation complémentaire par rapport à une évaluation en classe ordinaire. Les copies des élèves sont corrigées uniquement par le titulaire de la classe et les notes obtenues comptent au même titre qu'une évaluation en classe. Les résultats sont utilisés par les écoles pour situer leurs classes au sein du lycée et au niveau national.

21. À ce sujet, il faut rappeler que contrairement au référentiel de l'enseignement fondamental, les référentiels pour l'enseignement secondaire ne présentent, à l'heure actuelle, que des « standards de contenus » et pas encore de « standards de performance », ce qui rend encore très subjective l'appréciation des performances scolaires des élèves par les enseignants, mais également par les concepteurs des épreuves externes d'évaluation.

les enseignants, comme c'est le cas dans d'autres systèmes scolaires. Sans préjuger de leurs compétences et de leur liberté pédagogique, les enseignants placés devant le constat que leurs élèves font « moins bien qu'attendu » ou qu'une proportion élevée d'entre eux n'atteint pas les exigences du programme disposeraient ainsi d'une opportunité ou d'une aide pour réagir aux constats dressés.

Optimiser les ressources humaines et financières pour le développement scolaire

Les compressions budgétaires accrues qui touchent actuellement l'éducation ont conduit à une réduction des ressources humaines affectées aux écoles. Le ministère de l'Éducation est évidemment conscient de ce contexte particulier et le fait que certains établissements parviennent, à partir de leurs propres ressources, à entreprendre et à mener à bien des projets de développement ne peut justifier que rien ne soit fait pour les établissements qui n'y parviennent pas. C'est pourquoi le Ministère entend à l'avenir examiner les procédures d'allocation des ressources aux établissements dans le but d'encourager ceux-ci à s'engager pleinement dans le développement de leur qualité. À l'heure actuelle, les écoles fondamentales qui en ont fait la demande ont déjà reçu entre 2 et 12 heures par semaine afin de définir leur PRS. Les écoles secondaires ont quant à elles la possibilité de dégager 7 heures par semaine pour travailler à leur PDS. Mais cela reste insuffisant et l'effort doit à présent être dirigé vers l'accompagnement des équipes pédagogiques qui en font la demande.

BIBLIOGRAPHIE

Baumert J., Demmrich A., 2001, "Test motivation in the assessment of student skills – The effects of incentives on motivation and performance", *European Journal of Psychology of Education*, No. 16, Springer, p. 441-462.

Bond T., Fox C. , 2010, *Applying the Rasch Model – Fundamental measurement in the human sciences*, 2nd ed., New York, London, Routledge.

Burton R., Martin R., 2008, « L'orientation scolaire au Luxembourg : "Au-delà de l'égalité des chances... le gâchis d'un potentiel humain" », *in* Martin R., Dierendonck C., Meyers C., Noesen M., *La place de l'école dans la société luxembourgeoise de demain*, Bruxelles, De Boeck, p.165-186.

Dierendonck C., Sonnleitner P., Ugen S., Keller U., Fischbach A., Martin R., 2013, *La mesure de la motivation et de l'effort des élèves dans le cadre des épreuves standardisées au Luxembourg*, Actes du congrès de l'actualité de la recherche en éducation et formation (AREF-AECSE), Montpellier.

Dierendonck C., Martin R., 2008, « Le pilotage des systèmes éducatifs », *in* Martin R., Dierendonck C., Meyers C. Noesen M., *La place de l'école dans la société luxembourgeoise de demain,* Bruxelles, De Boeck, p. 429-476.

Ebel R., Frisbee D., 1986, *Essentials of educational measurement*, 4th ed., Toronto, Prentice Hall.

Gustafsson J.-E., 1980, "Testing and obtaining fit of data to the Rasch model", *British Journal of Mathematical and Statistical Psychology*, No. 33, vol. 2, The British Psychological society, p. 205-233.

Keskpaik S., Rocher T., 2012, *Les évaluations à faibles enjeux : quel rôle joue la motivation ? Une expérience à partir de PISA*, Communication dans le cadre du 24e colloque de l'Admée-Europe, Luxembourg.

Martin-Löf P., 1974, "The Notion of Redundancy and Its Use as a Quantitative Measure of the Discrepancy between a Statistical Hypothesis and a Set of Observational Data", *Scandinavian Journal of Statistics,* vol. 1, No. 1, p. 3-18.

Messick S., 1995, "Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning", *American Psychologist*, vol. 50, No. 9, p. 741-749.

Nagy G., Neumann M., 2010, „Psychometrische Aspekte des Tests zu den voruniversitären Mathematikleistungen in TOSCA-2002 und TOSCA-2006: Unterrichtsvalidität, Rasch-Homogenität und Messäquivalenz", *in* Trautwein U., Neumann M., Nagy G., Lüdtke O., Maaz K., *Schulleistungen von Abiturienten – Die neu geordnete gymnasiale Oberstufe auf dem Prüfstand*, Wiesbaden, Springer VS, p. 281-306.

OCDE, 2009, *PISA 2006*, *Technical report*, Paris, OECD Publishing.

O'Neil H. , Sugrue B., Abedi J., Baker E., Golan S., 1996, *Final Report of Experimental Studies on Motivation and NAEP Test Performance*, CSE Technical Report 427, Los Angeles, CRESST, University of California.

Scallon G., 2004, *L'évaluation des apprentissages dans une approche par compétences*, Bruxelles, De Boeck.

SCRIPT, 2007, « Courrier de l'Éducation Nationale – Die Steuerung des Luxemburger Schulwesens », n° spécial, MENFP.

Warm T., 1989, "Weighted Likelihood Estimation of Ability in Item Response Theory", *Psychometrika*, vol. 54, No. 3, Springer, p. 427-450.

Wolf L., Smith J., 1995, "The Consequence of Consequence: Motivation, Anxiety and Test Performance", *Applied Measurement in Education*, vol. 8, No. 3, p. 227-242.

Wright B., Linacre J. , Gustafsson J.-E., Martin-Löf P., 1994, "Reasonable mean-square fit values", *Rasch Measurement Transactions*, vol. 8, No. 3, p. 370.

Wu M., Adams R., Wilson M., Haldane S., 2007, *ACER ConQuest. Version 2.0: Generalised item response modelling software,* Camberwell, ACER Press.

Zerpa C., Hachey K., Van Barneveld C., Simon M., 2011, *Modeling Student Motivation and Students' Ability Estimates From a Large-Scale Assessment of Mathematics*, SAGE Open.

L'ÉVALUATION DES COMPÉTENCES DES ADULTES

Quelles contraintes ? Quelles spécificités ?

Fabrice Murat
MENESR-DEPP, bureau des études sur les établissements et l'éducation prioritaire

Thierry Rocher
MENESR-DEPP, bureau de l'évaluation des élèves

Évaluer les compétences des adultes est une opération bien plus complexe qu'évaluer les élèves. Les adultes sont sortis souvent depuis fort longtemps de l'école, ne sont plus habitués à la situation d'évaluation et en gardent parfois un mauvais souvenir. De plus, les conditions de passation, au domicile des personnes interrogées, sont aussi plus difficiles que les évaluations dans les salles de classe. Pour résoudre ces problèmes, des méthodologies spécifiques ont été développées ces dernières années, la demande de statistiques sur ce sujet ayant émergé au début des années 1990. Deux dispositifs d'enquêtes se dessinent : les enquêtes internationales (IALS, *International Adult Literacy Survey* ; ALLS, *Adult Literacy and Life Skills Survey* ; Piaac, *Program for the International Assesment of Adult*) sous l'égide de l'OCDE et les enquêtes françaises (IVQ, Information et vie quotidienne) pilotées par l'Insee. Cet article fait le point sur la méthodologie, les contraintes communes à ce type d'enquête et les points où les deux dispositifs se distinguent.

La mesure des compétences des adultes est à la jonction de deux traditions d'enquêtes : les « enquêtes-ménage », organisées notamment par l'Insee sur des thèmes aussi variés que l'emploi, le logement ou la santé, et les évaluations standardisées de compétences, plus généralement pratiquées sur des élèves. Ces deux modes de collecte d'informations statistiques ont leurs spécificités et l'enjeu d'une évaluation des compétences des adultes est d'en tenir compte, dans un protocole bien adapté. La qualité des résultats obtenus est très sensible aux choix faits pour répondre à cet objectif.

Jusqu'au début des années 2000, les enquêtes statistiques menées par l'Insee abordant le thème des compétences des adultes, et plus précisément celui de l'illettrisme, avaient une base déclarative : la personne enquêtée indiquait si elle éprouvait des

difficultés à lire des journaux, à remplir un chèque, etc. Une telle approche est très subjective et l'usage de tests pour vérifier ces déclarations est vite apparu nécessaire. Les évaluations de compétences des élèves, bien plus développées, ont pu servir d'exemple pour construire des outils de mesure, en tenant compte de la contrainte d'une enquête à domicile, dans un cadre moins standardisé que celui de la classe, avec des personnes moins disposées que des élèves à entrer dans une logique d'évaluation.

Le développement des évaluations de compétences des adultes ne s'est pas fait sans difficulté, comme le montre le bilan de la première opération d'envergure ayant eu les mêmes objectifs, l'enquête *International Adult Literacy Survey* (IALS) menée par *Statistic Canada* et *Educational Testing Service* (ETS, organisme responsable de nombreux tests d'évaluation aux États-Unis), dont les résultats ont été diffusés par l'Organisation de coopération pour le développement économique (OCDE). Les conclusions étaient pour la France très surprenantes [NCES, 1998 ; OCDE, 2000] : 40 % des Français entraient dans la catégorie des plus mauvais lecteurs (vite assimilée à celle des illettrés), ce qui situait la France bien loin derrière la plupart des pays participants (entre autres, l'Allemagne, la Suède, les États-Unis). Des investigations ont mis en évidence un certain nombre de problèmes méthodologiques (traductions non équivalentes des textes et des questions en termes de difficulté, plan de sondage et correction de la non-réponse peu satisfaisants, conditions de passation peu adaptées, etc.) qui ont justifié le retrait de la France de l'opération et la non-diffusion officielle des résultats [Blum et Guérin-Pace, 2000 ; Murat, 2008].

Depuis cette date, de gros progrès ont été faits dans les dispositifs nationaux et internationaux. La France a développé son propre dispositif d'enquêtes (les enquêtes Information et vie quotidienne – IVQ) en réponse aux problèmes identifiés dans l'enquête IALS et elle a également accepté de participer au programme Piaac (*Programme for the International Assessment of Adult Competencies*). Cet article décrit les principes et les caractéristiques des évaluations de compétences des adultes, par comparaison avec les évaluations des élèves, et fait un point sur les opérations existantes et à venir [voir aussi Degorre et Murat, 2009 pour une présentation générale sur ces questions].

POURQUOI ÉVALUER LES COMPÉTENCES DES ADULTES ?

L'intérêt d'évaluer les compétences des adultes est multiple et relève d'objectifs un peu différents de ceux des évaluations en milieu scolaire. En effet, on justifie celles-ci soit dans un but diagnostique pour repérer les élèves en difficulté et les aider, soit pour valider les acquis par un examen final, soit pour donner une image globale du système éducatif, en termes de niveau moyen, de disparité, de points forts et de points faibles [Trosseille et Rocher, dans ce numéro, p. 15]. Les finalités d'une évaluation sur population adulte sont un peu différentes et mettent plutôt l'accent sur l'importance de la lecture et du calcul dans la vie personnelle et professionnelle. Les personnes vivant en France ont-elles un degré de maîtrise suffisant pour faire face aux mutations du monde du travail et de la vie quotidienne, qui nécessitent un accès bien maîtrisé à une information de plus en plus complexe ?

Décrire le niveau de compétence de la population

Contrairement aux évaluations d'élèves, ces opérations peuvent assez difficilement être utilisées pour évaluer le fonctionnement du système éducatif (même si de nombreux journalistes ne se privent pas de le faire) : les personnes évaluées sont sorties parfois depuis très longtemps d'un système scolaire très différent de celui d'aujourd'hui. Cependant, pour faire un tel usage de ces évaluations, on peut se restreindre aux personnes les plus jeunes. L'étude du lien entre les caractéristiques sociodémographiques des individus et les compétences permet aussi de renouveler les connaissances sur les inégalités sociales de réussite scolaire [Place et Vincent, 2009]. De plus, on peut envisager une certaine circularité de cette relation : les parents les moins compétents risquent de ne pas être en mesure d'aider leurs enfants dans leurs études, provoquant chez eux un déficit de compétences [Murat, 2009] ▸ **Encadré**.

L'étude de la répartition de la population dans les différents niveaux de compétences fait l'objet de publications nationales [Murat, 2005 ; Jonas 2012 ; Jonas 2013], mais aussi dans les nombreuses régions, de métropole ou des DOM, qui ont augmenté la taille de l'échantillon pour pouvoir publier des chiffres locaux (tous les DOM ont procédé à ce type d'extension et c'est le cas aussi de l'Aquitaine, du Nord-Pas-de-Calais et des Pays de la Loire en 2004, de la Haute-Normandie, de l'Île-de-France, du Nord-Pas-de-Calais, de la Picardie et de la Provence-Alpes-Côte d'Azur en 2011). Des analyses plus approfondies ont aussi été menées selon le sexe [Djider et Murat, 2006] ou l'âge des individus [Micheaux et Murat, 2006].

LE LIEN ENTRE COMPÉTENCES DES PARENTS ET SCOLARITÉ DES ENFANTS

La sociologie de l'éducation a beaucoup mis en avant – à côté du milieu social mesuré par la profession du père – l'importance du « capital culturel », souvent appréhendé par le diplôme de la mère, pour étudier la réussite scolaire des enfants. L'enquête IVQ propose un indicateur différent de ce capital culturel : le niveau de compétence à l'écrit, en calcul et en compréhension orale. Une analyse menée sur les données de l'enquête IVQ de 2004 montrait l'importance de ces caractéristiques en utilisant le retard scolaire comme indicateur de performance scolaire [Murat, 2009] : la moitié des enfants de 7 ans à 18 ans dont les parents avaient de faibles compétences en lecture avaient pris au moins une année de retard scolaire contre un cinquième quand les résultats des parents aux tests d'IVQ étaient très satisfaisants. L'écart était à peu près équivalent avec les compétences en calcul, un peu moindre avec celles en compréhension orale. La prise en compte des autres caractéristiques du ménage (professions et diplômes des parents, revenus, etc.) réduisait ces écarts, mais ils restaient significatifs, indiquant une autre source d'inégalités à l'école que celles classiquement mesurées.

Il a paru utile de reproduire cette analyse sur l'édition de 2011 de l'enquête, car entre-temps, du fait d'une politique de réduction des redoublements, en particulier dans l'enseignement élémentaire, le taux de retard a sensiblement diminué : il est passé de 34 % pour IVQ 2004 à 26 % pour IVQ 2011. La corrélation entre les compétences des parents et le retard scolaire des enfants reste très nette : 40 % des enfants dont les parents ont eu de faibles performances en lecture ont pris du retard contre 14 % pour les enfants dont les parents sont parmi les plus compétents.

Tableau 1 Retard scolaire des enfants en fonction des compétences des parents (en %)

Quartile	Lecture	Calcul	Compréhension orale	Les trois compétences
1	40	40	33	41
2	26	25	26	23
3	22	21	23	23
4	14	16	20	15
Ensemble	26	26	26	26

Lecture : 40 % des enfants se trouvant dans le premier quartile de compétences parentales en lecture (c'est-à-dire les 25 % des enfants dont les parents ont les scores les plus bas dans ce domaine) ont au moins un an de retard scolaire.
Champ : enfants de 7 à 18 ans vivant chez leurs parents en France métropolitaine.
Source : enquête Information et vie quotidienne 2011, Insee.

Du fait de la baisse du taux de retard, la comparaison avec IVQ 2004 n'est pas immédiate : comme ce taux se rapproche de 0, les écarts sont mécaniquement moins marqués. Le recours à une modélisation logistique permet de résoudre ce problème et d'introduire des variables de contrôle dans l'analyse pour tenir compte des corrélations entre les deux variables et d'autres caractéristiques de l'enfant ou des parents (sexe, milieu social, revenus, etc.). La prise en compte de l'âge de l'enfant est ainsi indispensable, car il joue fortement sur la probabilité d'avoir redoublé par le passé. Un premier modèle a donc été construit en reliant le retard scolaire avec l'âge de l'enfant et le score moyen des parents dans les trois domaines. Le coefficient associé au score dans IVQ 2011 est de - 0,64. C'est légèrement inférieur à la valeur de 2004 (- 0,72).

Cependant, la modélisation peut être enrichie avec d'autres variables, à la fois liées aux compétences et au retard scolaire. C'est le cas par exemple du diplôme des parents, dont on sait l'importance pour l'étude de la scolarité, et qui est assez logiquement lié à leurs compétences. Cette variable a donc été ajoutée dans l'analyse, ainsi que l'âge, le sexe, le pays de naissance, la profession des parents, l'âge et le sexe de l'enfant et le revenu du ménage. Le coefficient associé au score dans les trois domaines passe à - 0,34, ce qui est encore une fois un peu inférieur à la valeur de 2004 (- 0,41). Ce coefficient reste toutefois statistiquement significatif et indique une corrélation avec le retard scolaire d'une ampleur à peu près équivalente à celle concernant le diplôme.

Ces résultats pourraient être approfondis comme en 2004 pour tester des effets plus subtils en fonction du sexe du parent par exemple. Rappelons qu'en 2004, une spécialisation était apparue selon la discipline : les compétences en calcul paraissaient plus importantes pour les pères, tandis que pour les mères, ce sont les compétences en lecture qui jouaient le plus. Ce résultat n'est pas apparu clairement dans les premières analyses sur 2011. Plus précisément, des coefficients ne sont significatifs que dans les régressions non pondérées. En 2004, une certaine sensibilité des résultats à l'usage ou non des pondérations avait déjà été mise en évidence, ce qui incite à la prudence et à poursuivre l'expertise méthodologique.

Guider la politique de lutte contre l'illettrisme

Mieux connaître la population des personnes le plus en difficulté face à l'écrit est aussi tout à fait crucial. Il faut distinguer cette population en fonction du degré de difficultés et de leur origine (absence de scolarité, maîtrise de la langue française insuffisante, etc.) afin de cibler l'action de formation sur les différents groupes.

C'est un des objectifs suivis par l'enquête IVQ qui a développé un module spécifique pour mesurer précisément les compétences des personnes en situation d'illettrisme [Besse, Luis *et alii*, 2009] et le parcours individuel de ces personnes a fait l'objet d'analyses spécifiques [Guérin-Pace, 2009]. L'Agence nationale de lutte contre l'illettrisme (ANLCI) utilise en effet l'enquête IVQ pour chiffrer l'illettrisme dans ses différents rapports. L'intérêt des institutions locales pour cette enquête s'explique aussi en partie par l'implication de ces institutions dans la lutte contre l'illettrisme et la nécessité de quantifier ce phénomène de façon fine. L'illettrisme a aussi été confronté à d'autres formes de difficultés sociales : la pauvreté [Murat, 2006] ou le fait d'habiter dans un quartier défavorisé [Murat, 2007].

Affiner l'analyse du marché du travail

Ces évaluations peuvent aussi servir, en s'inspirant de différentes théories économiques, à mieux comprendre le fonctionnement du marché du travail, en donnant des compétences des personnes interrogées un indicateur plus direct que le diplôme. On peut très rapidement rappeler les deux positions qui s'opposent dans les travaux concernant l'influence de l'éducation sur le marché du travail : d'une part, la théorie du capital humain, qui postule que les études permettent d'accroître les compétences des individus, ensuite valorisées sur le marché du travail, et d'autre part la théorie du signal, qui avance que les compétences sont une donnée préexistante, qui permettent d'atteindre un niveau d'études plus ou moins élevé, les employeurs utilisant alors justement ce niveau d'études comme indicateur de ces compétences. Ces deux théories impliquent une corrélation entre compétences et marché du travail et se distinguent surtout sur la question de la formation de ces compétences, en amont de ce que les évaluations d'adultes permettent d'étudier. Les enquêtes de l'OCDE, IALS et Piaac, ont ainsi donné lieu à de nombreuses analyses sur le lien entre littératie et marché du travail [Green et Riddle, 2001].

LES SPÉCIFICITÉS D'UNE ÉVALUATION D'ADULTES

Quoi évaluer ?

Dans la phase d'élaboration d'une évaluation de compétence, la question centrale est : que veut-on mesurer ? Les limitations dues aux temps d'enquête sont très fortes : une durée d'évaluation d'environ trois quarts d'heure semble difficile à dépasser si l'on souhaite disposer de réponses données dans de bonnes conditions. C'est pourquoi les évaluations des compétences des adultes se restreignent le plus souvent aux disciplines « fondamentales », la lecture et le calcul, en privilégiant souvent le premier domaine. La compréhension orale a cependant aussi été évaluée en mode mineur dans l'enquête IVQ. Dans Piaac, des compétences plus larges, mobilisées dans le cadre professionnel, sont approchées par un questionnaire déclaratif (*Job Requirement Approach*).

Pour définir plus précisément les dimensions à mesurer, il faut aussi tenir compte des objectifs de l'enquête. Cela fixe le cadre de définition et de mesure des compétences. La question se pose de façon cruciale pour les évaluations d'élèves : dans une optique nationale d'évaluation du système éducatif, il apparaît logique de se référer aux programmes en cours pour construire les épreuves, comme c'est le cas pour le

programme d'évaluations Cedre [Trosseille et Rocher, dans ce numéro, p. 15]. Dans un cadre international, on a souvent une vision plus large, en se référant plutôt à l'usage des compétences dans la vie personnelle et professionnelle, comme c'est le cas par exemple dans l'évaluation internationale PISA.

Dans le cas des évaluations d'adultes, nationales ou internationales, la référence aux programmes scolaires paraît peu pertinente et c'est donc la deuxième approche qui a été retenue. Ces enquêtes ont ainsi recours aux concepts de « littératie » et de « numératie » pour désigner les compétences de lecture et de calcul mobilisées sur des supports de la vie quotidienne. Ces dimensions sont très larges, elles renvoient à de nombreuses compétences, envisagées dans une perspective utilitariste. Le cadre de construction des épreuves reste donc assez empirique bien qu'il intègre les apports de la psychologie cognitive, s'agissant de notions telles que la compréhension de l'écrit [Besse, Luis *et alii,* 2009 ; Megherbi, Rocher *et alii*, 2009] ou la dyscalculie [Fisher et Charron, 2009] dans le cas d'IVQ. L'effort porte en fait surtout sur la variété des supports et sur la bonne acceptation des exercices lors des tests.

Cependant, la référence à l'usage des compétences dans la vie quotidienne peut poser problème. En effet, la tentation est grande alors pour certains enquêtés de répondre aux questions à partir de leurs connaissances et expériences personnelles. Selon que ces connaissances et expériences correspondront ou non aux informations données par les supports d'évaluation, la personne interrogée donnera ou non la bonne réponse, sans que ses compétences de compréhension et de raisonnement soient mobilisées. Une des questions de l'enquête Piaac portait sur la langue parlée à la Guadeloupe : on peut se demander combien de personnes en France se sont référées au dépliant touristique faisant office de support pour y répondre.

Prendre en compte la situation d'évaluation

Pour essayer de diminuer ce problème, la compréhension et la restitution des objectifs de l'enquête par les enquêteurs sont fondamentales. D'autant plus que les enquêteurs en charge de l'enquête ne sont pas des spécialistes de l'évaluation. Il est donc important que le protocole (présentation des supports d'évaluation, recueil des réponses, réactions aux demandes de précisions éventuelles, etc.) soit bien standardisé, mais tienne aussi compte des incertitudes d'une enquête au domicile des personnes interrogées, cadre bien moins approprié que la classe dans le cas des évaluations d'élèves. La présence des autres membres de la famille peut ainsi être source de nombreuses perturbations. Ils peuvent ne pas bien prendre l'« intrusion » assez longue de l'enquêteur dans le logement. Ils peuvent aussi vouloir participer à l'évaluation, en aidant la personne interrogée ou en cherchant à prendre sa place. Les enquêteurs doivent savoir gérer ces difficultés, en rappelant la nécessité d'une mesure individuelle des compétences ou par exemple en donnant une copie des épreuves au « perturbateur », en lui demandant d'y répondre de son côté. Proposer de baisser le son de la télévision est aussi un réflexe que les enquêteurs acquièrent assez vite.

Les formations aux enquêteurs permettent de bien définir les enjeux de l'enquête et de donner quelques consignes pour gérer ce type de situation. La qualité des résultats dépend du professionnalisme des enquêteurs dans le respect des consignes et

de leur capacité à s'adapter aux situations imprévues. En particulier, les critiques adressées à l'enquête IALS concernant les conditions de passation [Blum et Guérin-Pace, 2000 ; Carey, 2000] ont bien montré qu'une attention particulière doit porter sur les relations entre enquêteurs et enquêtés, afin de favoriser et de maintenir la motivation des enquêtés au cours de l'évaluation. La création de l'enquête IVQ a été une réponse à ce problème, en proposant un protocole plus naturel [Vallet, Bonnet *et alii,* 2002]. En particulier, il a été décidé de donner les exercices les uns après les autres et non pas globalement dans un livret. Le test paraît alors moins lourd et les échanges, bien cadrés, entre enquêteurs et enquêtés entre deux exercices, permettent d'atténuer l'impression d'examen.

Si des efforts importants peuvent être faits pour obtenir des conditions de passation les plus standardisées et confortables possibles, la situation sur le terrain n'est pas toujours idéale. C'est pourquoi les enquêteurs doivent recueillir des informations sur le déroulement de l'enquête. Cela peut prendre la forme d'une grille d'observation, où l'enquêteur va rendre compte, parfois exercice par exercice, de la façon dont l'enquêté s'est investi dans l'évaluation. L'informatisation des épreuves permet aussi facilement d'enregistrer les temps passés sur chaque exercice, qui, trop courts, peuvent être le signe d'un manque de motivation [Murat et Zamora, 2002].

Adapter l'épreuve

Les enquêtes auprès des adultes portent sur une population beaucoup plus hétérogène, en termes de niveaux de compétences, que celles portant sur des élèves, généralement menées à un niveau scolaire donné ou à un âge donné. Les personnes interrogées ont connu des parcours scolaires très variés, parfois hors de France, dans une autre langue que le français, à des époques plus ou moins éloignées. Il faut donc construire un protocole d'évaluation qui puisse s'adapter à toutes les situations. C'est pourquoi les évaluations d'adultes comportent généralement un processus d'« orientation », qui permet d'adapter la difficulté des épreuves au niveau de compétences de la personne. Il ne faut pas décourager les personnes en difficulté par des questions trop difficiles ; il ne faut pas non plus perturber les personnes très compétentes avec des questions trop faciles, car elles risquent de se démotiver ou de chercher des difficultés où il n'y en a pas. Pour les plus en difficulté, c'est la nature même des épreuves qui peut être modifiée, la mesure de la compréhension de texte pouvant être utilement complétée par celle du décodage des mots et des procédures élémentaires d'écriture, comme c'est le cas dans IVQ.

Cependant, cette nécessité d'adapter les épreuves conduit à des protocoles d'évaluation assez complexes, composés de filtres permettant l'orientation des individus vers des épreuves différentes, en fonction de leurs résultats obtenus, au fur et à mesure qu'ils avancent dans l'évaluation. Avec cette procédure, appelée *multistage adaptive testing*, l'estimation des niveaux de compétences des individus est complexe [Murat et Rocher, 2009]. En effet, par construction, le niveau de difficulté des épreuves proposées dépend du niveau de compétence des individus. Les individus les plus performants passent des épreuves difficiles ; les individus les moins performants des épreuves faciles. Dès lors, il n'est pas possible de comparer directement les individus en fonction de leurs résultats obtenus aux différentes épreuves. Les modèles psychométriques permettent précisément d'envisager de

façon séparée les deux concepts que sont le niveau de compétence des individus et le niveau de difficulté des épreuves [Rocher, dans ce numéro, p. 37]. Pour ce faire, il est nécessaire de disposer d'épreuves communes à des individus de niveaux différents. Les enquêtes IVQ et Piaac adoptent ce principe, cependant avec des modalités différentes que nous détaillons plus loin.

DEUX FAMILLES D'ÉVALUATION DES ADULTES

Les enquêtes internationales

La première enquête à grande échelle sur les compétences en littératie des adultes a pris place en 1985 aux États-Unis. Le *Young Adult Literacy Survey* (YALS) a été organisé par ETS avec le soutien du *National Center for Educational Statistics* (NCES). Restreinte au champ des jeunes adultes, l'enquête YALS a permis d'expérimenter pour la première fois un dispositif d'évaluation dont les épreuves sont fondées sur des supports diversifiés, avec des niveaux de difficulté divers. S'appuyant sur les résultats de cette première enquête, le *National Adult Literacy Survey* (NALS) a été conduit en 1989-1990 sur un échantillon couvrant l'ensemble de la population adulte des États-Unis. Trois grandes familles d'épreuves ont été proposées. La dimension *Prose Literacy* mesure la capacité à comprendre et à utiliser de l'information organisée à travers des phrases, elles-mêmes structurées en paragraphes. Des textes narratifs, mais aussi de la poésie, ont ainsi été repris à partir de journaux, magazines ou brochures, en préservant la typographie et la mise en page originale. La dimension *Document Literacy* s'intéresse à l'aisance pour manipuler de l'information structurée en matrice, c'est-à-dire à travers des lignes et des colonnes. Les supports reprennent des tableaux, des tickets, des graphiques, des grilles horaires, etc. La dimension *Quantitative Literacy* correspond à la notion d'arithmétique quotidienne : additions, soustractions, multiplications et divisions prennent place à travers des mises en situation (calculer une remise sur un achat, trouver le coût d'un emprunt, etc.).

Ces méthodes ont également inspiré, au niveau international, l'enquête *International Adult Literacy Survey* (IALS), organisée par Statistique Canada et ETS, puis coordonnée par l'OCDE dans une vingtaine de pays entre 1994 et 1999. En se fondant sur des travaux de chercheurs américains spécialisés en psychométrie, l'enquête IALS a donné lieu à des études comparatives [OCDE, 2000] sur les niveaux de compétences observés dans les pays ayant participé au dispositif. Une approche fondée sur les modèles de réponse à l'item (MRI) a ainsi été mobilisée pour construire une échelle internationale commune, à partir d'items traduits dans chaque langue [Rocher, dans ce numéro, p. 37].

L'OCDE, assistée de Statistique Canada et de ETS, a décidé de lancer en 1999 une autre enquête sur les compétences des adultes, *Adult Literacy and Life Skills Survey* (ALLS), étendant son champ d'investigation à d'autres compétences comme la « résolution de problèmes » (*Problem Solving*). Les deux dimensions *Prose Literacy* et *Document Literacy* ont été fusionnées en une seule échelle. Une évaluation en numératie, portant plus spécifiquement sur les compétences de calcul, s'est substituée à la *Quantitative Literacy*. Les principes de conception du questionnaire IALS ayant été repris à l'identique, malgré

les critiques adressées, la France a décidé de ne pas participer à ce projet et de développer en premier lieu un cadre national d'évaluation des compétences.
En 2007, l'OCDE a lancé l'opération Piaac (*Program for the International Assessment of Adult*) auquel ont participé 24 pays. Les compétences mesurées restent les mêmes que pour ALLS (avec une informatisation du protocole sur laquelle nous allons revenir). Cependant, le protocole tenait compte en partie des enseignements de l'expérience française. La France a décidé de participer à cette enquête en 2012, qui a concerné 7 000 personnes âgées de 16 ans à 65 ans. Les premiers résultats ont été diffusés à partir de la fin de l'année 2013 [OCDE, 2013 ; Jonas, 2013].

L'enquête IVQ

Suite aux problèmes rencontrés par l'enquête IALS et au refus de la France de participer à l'enquête ALLS, un comité de pilotage a été institué pour mettre au point un protocole rigoureux et adapté à la réalisation d'une enquête ménage en France[1]. Les objectifs définis dans le cadre de ce partenariat ont façonné la forme prise par l'enquête Information et vie quotidienne. La rencontre entre la recherche en psychologie[2] et l'ingénierie statistique a permis de donner corps à ces objectifs, en adaptant des exercices d'évaluation des compétences puis en les orchestrant sous la forme d'une enquête ménage, tout en prenant soin de répondre aux biais potentiels posés par le protocole et en particulier les relations entre enquêteurs et enquêtés. Un premier test de l'opération a été mené en décembre 2000 sur quelques centaines d'individus pour s'assurer que le principe même d'une évaluation à domicile était possible. On a constaté que les personnes interrogées se prenaient au jeu, grâce à des supports variés et ancrés dans la vie quotidienne (un programme TV, un CD, des cartes routières, etc.), mais que les problèmes de motivation (attention accordée au questionnaire, ou lassitude en fin d'épreuve) se posaient toujours de façon cruciale. C'est pourquoi un effort particulier a été fait pour capter de l'information sur le degré de motivation de chaque enquêté. Le deuxième test, sur un échantillon du même ordre, en avril 2002, a servi à choisir et améliorer des épreuves pour une première édition de l'enquête sur un échantillon important [Vallet, Bonnet *et alii,* 2002]. L'expertise fine des données a permis de valider le protocole et d'envisager la mise en place d'une enquête IVQ 2004 sur un échantillon étendu à l'ensemble de la France métropolitaine. Le projet a été présenté devant les instances nationales en charge de coordonner les travaux de la statistique publique et d'en vérifier la qualité – le Conseil national de l'information statistique et le Comité du label. Confortée par les premiers résultats méthodologiques, l'enquête IVQ 2004 a bénéficié du label d'intérêt général et de qualité statistique.

1. Dans le groupe de pilotage, se trouvent représentés l'ANLCI (Agence de lutte contre l'illettrisme), le CREST (Centre de recherche en économie et en statistique), la DARES (direction de l'animation de la recherche et des études sociales du ministère du Travail), la DEPP (direction de l'évaluation, de la prospective et de la performance, du ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche), l'INED (Institut national des études démographiques), l'Inetop (Institut national d'étude du travail et de l'orientation professionnelle), l'Insee (Institut national de la statistique et des études économiques).
2. Plusieurs équipes universitaires ont travaillé sur les exercices proposés : PsyEf (Lyon 2) sur le module d'orientation et module « illettrés », EVA (Rennes 2) sur le module d'orientation et module « numératie », Paris V et Paris XII sur le module « haut ».

Conduite dans l'ensemble des régions de France métropolitaine, l'enquête IVQ 2004 s'est appuyée sur un échantillon de 17 300 logements, avec des critères de pondération qui augmentaient les probabilités de trouver une personne en difficulté (notamment, chef de ménage peu diplômé ou né à l'étranger). Un sur-échantillonnage a été réalisé pour obtenir des résultats sur les zones urbaines sensibles (ZUS). Le protocole a également été repris dans les DOM par la suite avec des adaptations pour préserver la proximité sémantique des exercices à la vie quotidienne des populations enquêtées : par exemple, les noms des villes dans les textes ont été modifiés, tout en préservant les difficultés formelles qu'ils doivent présenter.

En 2011, une nouvelle édition de l'enquête a eu lieu en France métropolitaine. Elle a concerné 14 000 personnes et a permis d'établir des comparaisons avec l'enquête de 2004 [Jonas, 2012].

Comparaison entre les deux enquêtes

La participation de la France à l'enquête Piaac montre que, depuis IALS, les évaluations internationales d'adultes ont sensiblement amélioré leur méthodologie et pris en compte les travaux en France sur IVQ. Cependant, il reste un certain nombre de différences, qui peuvent expliquer des divergences dans les résultats [Jonas, Lebrère *et alii,* 2013].

Dans l'ensemble, pour ce qui concerne la maîtrise de l'écrit, les deux enquêtes s'appuient sur une définition assez proche de ce que l'on souhaite évaluer : la capacité à utiliser l'information dans la vie quotidienne. Les deux enquêtes se distinguent plutôt dans le traitement des personnes les plus en difficulté. Les responsables de Piaac ont développé un module spécifique pour les personnes qui montraient très vite des difficultés particulières aux premiers exercices, mais la France n'a pas souhaité utiliser ce module. En effet, l'étude des processus élémentaires de la lecture, incluant le déchiffrage de mots, est apparue peu pertinente dans une perspective internationale, avec des langues aux orthographes et aux alphabets très variés. L'enquête IVQ a développé des exercices spécifiques à ce sujet. Par ailleurs, la mesure des compétences en calcul, évaluée en mode mineur dans IVQ, s'appuie sur des principes théoriques assez différents de la numératie utilisée dans Piaac. Dans IVQ, il s'agit, à partir de problèmes courts et d'une forme un peu scolaire, de mesurer la maîtrise des outils de bases des mathématiques (additions, soustractions, proportionnalité, logique), alors que Piaac contextualise davantage les exercices, en essayant de les rapprocher de l'utilisation des compétences en mathématiques dans la vie quotidienne.

Si les deux enquêtes adoptent le principe d'une procédure adaptative, d'un point de vue pratique, le protocole adaptatif de Piaac est plus complexe que celui d'IVQ. L'enquête Piaac distingue dans un premier temps, à partir des réponses au questionnaire biographique et à travers un court test informatique, ceux qui maîtrisent l'outil informatique et les autres. Ces derniers passeront une version papier des épreuves, commençant par huit questions. À l'issue de ces huit questions, l'évaluation est interrompue pour les individus les plus faibles (dans certains pays, ils passent alors une évaluation spécifique). Les autres passent soit une épreuve en littératie, soit une épreuve en numératie. Pour les personnes qui ont des compétences suffisantes en informatique, l'évaluation se fait en cinq étapes : une première épreuve permet

d'interrompre l'évaluation pour les individus les plus faibles. Une deuxième épreuve est proposée en littératie ou en numératie (la passation d'un des deux domaines en premier est aléatoire) en fonction des résultats à l'épreuve précédente et du niveau d'études connu par le questionnaire biographique : les individus ayant eu de bons résultats et/ou un bon niveau d'études ont plus de chances de se voir proposer des exercices difficiles[3]. Une troisième épreuve est proposée dans le même domaine, là encore adaptée en fonction des réponses aux exercices précédents. Enfin, les deux dernières épreuves reprennent le même mécanisme pour le domaine qui n'a pas été évalué en premier. Cette procédure nécessite à la fois une estimation très précise en amont de l'enquête de la difficulté des exercices et une correction automatique des premières réponses des enquêtés.

Le protocole d'IVQ est plus simple dans la mesure où il ne comporte qu'un seul processus d'orientation (*two-stage adaptive test*). Après une première épreuve dite d'orientation, trois choix sont possibles en fonction des résultats obtenus : un module spécifique est destiné aux personnes en difficulté face à l'écrit ; un module « haut » est proposé aux personnes qui maîtrisent les compétences élémentaires de la lecture ; un module intermédiaire est proposé aux personnes ayant eu des résultats passables à l'épreuve d'orientation et permet de préciser laquelle des deux orientations ci-dessus est préférable. En 2011, deux innovations ont été introduites. D'une part, le module « haut » a été enrichi de manière à mieux décrire les compétences des individus qui n'ont pas de difficulté de lecture : ces individus sont alors orientés aléatoirement vers un module « haut A » ou bien vers un module « haut B », deux modules qui comportent un exercice en commun afin de relier les résultats. D'autre part, un « module commun » est proposé à tous les individus en fin d'évaluation. Ce module permet d'affiner la mesure du niveau de compétences des individus situés à la frontière du seuil d'orientation et de consolider la construction d'un score commun à tous les individus.

Les principales différences entre les deux enquêtes portent très certainement sur les conditions de collecte : durée de la passation, relations enquêteurs-enquêtés, support de l'évaluation (papier/informatique).
Notons tout d'abord que l'enquête Piaac est particulièrement longue : elle peut facilement atteindre 2 heures, alors que l'enquête IVQ a été calibrée pour une passation d'un peu plus d'une heure. De plus, dans Piaac, le choix a été fait de faire passer d'abord un long questionnaire biographique (une heure environ), au risque de proposer les exercices à des enquêtés déjà lassés. Le fait de faire passer le questionnaire biographique avant les épreuves cognitives apparaît en cohérence avec les principes d'estimation des compétences qui utilisent les informations biographiques et les résultats aux épreuves pour en déduire des « valeurs plausibles » des niveaux de compétences [Rocher, dans ce numéro, p. 37].
Un point crucial concerne les relations entre enquêteurs et enquêtés. En effet, l'expertise française sur IALS avait montré que la passation sous forme de cahier avait un caractère peu naturel et stressant pour l'enquêté : pendant qu'il travaillait

3. L'orientation n'est pas déterministe et comporte un facteur aléatoire pour contourner le problème d'estimation évoqué dans la partie « adapter l'épreuve », p. 89. En outre, il existe de nombreux recouvrements entre les différentes épreuves qui comportent des questions communes.

sur les exercices, l'enquêteur était inactif et l'enquêté pouvait avoir l'impression de lui faire perdre son temps. Dans l'enquête IVQ, ce problème avait été contourné en faisant donner les exercices un par un par l'enquêteur et en l'impliquant dans la saisie des réponses (ainsi cela permet de tester plus précisément la compréhension sans interférence avec les capacités d'expression, qui sont sollicitées si l'enquêté doit écrire ou saisir sa réponse). De plus, cette procédure permet de concentrer l'attention de l'enquêté sur chaque exercice, sans qu'il soit tenté de faire un choix en feuilletant le cahier et en s'arrêtant sur les supports qui l'intéressent le plus. L'enquête Piaac est restée sur le principe d'une évaluation auto-administrée, en passant sur un support électronique. Le rôle de l'enquêteur est volontairement limité au minimum et il reste un témoin silencieux, inutile, voire un peu gênant pendant la passation des épreuves. Cette approche interroge sur la possibilité de maintenir un degré suffisant de motivation des enquêtés au cours de la passation d'épreuves n'ayant aucun enjeu pour eux. Ainsi, dans IVQ, les enquêteurs suivent une formation préalable approfondie, pour que les interactions avec les enquêtés soient les plus naturelles possible. L'enquêteur doit maintenir la motivation de l'enquêté, sans pour autant lui donner des indications sur la qualité de ses réponses. En particulier, les demandes de correction et d'explication sont renvoyées à la fin de l'enquête, pour que les personnes passent les exercices dans les mêmes conditions : au sein d'un même exercice, corriger la première question peut ainsi aider ceux qui le demandent, à répondre aux questions suivantes.

Les procédures de passation constituent un élément clé de la qualité de ces enquêtes. Murat [2009] a pu montrer qu'à partir des mêmes items utilisés par IALS, mais avec les procédures de passation d'IVQ, le taux de mauvais lecteurs s'établissait à 15 %, contre 40 % dans IALS. Ce résultat issu d'IVQ était d'ailleurs proche de celui obtenu à partir de PISA 2000 pour les élèves de 15 ans, ce qui renforce sa vraisemblance et confirme que la collecte de IALS n'a pas dû se faire dans les meilleures conditions. Le professionnalisme des enquêteurs et le soin mis à les former ont donc une importance fondamentale dans les résultats. Cependant, sur ce point, l'enquête Piaac ne se distingue pas autant d'IVQ que IALS, car les enquêteurs sont issus du réseau des enquêteurs professionnels de l'INSEE, ce qui renforce la standardisation et limite les biais de passation. Les enquêteurs ont par ailleurs bénéficié d'une formation poussée pour les aider à gérer la situation particulière d'évaluation dans le cadre d'une enquête à domicile.

Le passage au support électronique est une innovation propre à l'enquête Piaac. Les responsables de Piaac ont ainsi fait l'hypothèse qu'il y a une équivalence parfaite entre passation électronique et passation sur support papier. En particulier, ils pensent être en mesure de comparer les résultats de Piaac avec ceux de IALS et de ALLS, dont les épreuves ont été passées sur support papier. Or, des travaux réalisés sur des évaluations d'élèves montrent la fragilité de l'hypothèse d'équivalence entre les deux supports [Bessonneau, Arzoumanian, Pastor, dans ce numéro, p. 159]. Ce problème a d'ailleurs été mis en exergue lors de l'expertise de l'enquête pilote française de Piaac qui avait montré des différences de difficulté très sensibles pour certains exercices entre les personnes qui le passaient sous forme électronique et celles qui le passaient sous format papier (ces décalages apparaissaient beaucoup plus nets pour les exercices de littératie que de numératie). De plus, la procédure d'orientation de Piaac peut sembler en contradiction avec ce principe d'équivalence : les individus qui montrent une maîtrise trop médiocre des procédures électroniques (cliquer, surligner, etc.),

passent une version papier aussi difficile que les exercices électroniques, en faisant l'hypothèse, par ailleurs vérifiée, qu'ils peuvent obtenir des résultats tout aussi bons que les personnes qui maîtrisent les procédures informatiques. Cela semble signifier implicitement qu'une dimension « maîtrise de l'informatique » peut perturber la mesure des compétences en compréhension. Il est regrettable que le protocole final n'ait pas prévu la passation d'exercices papier pour une partie au moins des personnes n'ayant pas de difficulté avec l'informatique.

Enfin, nous ne développons pas ici les aspects liés aux comparaisons internationales, qui sont de fait une différence entre les deux enquêtes. Cependant, il est à noter que ces problèmes d'équivalence entre les deux supports posent de sérieuses questions quant à la validité des comparaisons internationales. En effet, il existe des différences importantes entre pays du niveau de maîtrise, objectif ou déclaré, des compétences informatiques. Par conséquent, le nombre de personnes concernées par la passation informatique ou la passation papier-crayon peut être très variable d'un pays à l'autre : ainsi, il s'avère que 37 % des Japonais ont passé la version papier-crayon alors que cela n'a concerné que 10 % des Néerlandais. La validité des comparaisons internationales est donc fortement soumise à l'hypothèse d'équivalence des deux versions, dans chacun des pays.

QUEL AVENIR POUR LES ÉVALUATIONS D'ADULTES ?

Malgré une certaine proximité dans les principes et dans les méthodes, les enquêtes internationales et françaises se distinguent sur de nombreux aspects méthodologiques, en particulier pour ce qui concerne les conditions de collecte des données. Les résultats globaux des dernières enquêtes IVQ et Piaac apparaissent cependant relativement cohérents : par exemple le lien entre le niveau de compétence et le niveau de diplôme est assez comparable d'une enquête à l'autre [Jonas, Lebrère *et alii,* 2013]. Au final, c'est sans doute la capacité de ces enquêtes à répondre ou non à certaines problématiques qui constitue la véritable ligne de démarcation.

Ainsi, la possibilité de situer la France par rapport aux autres pays est évidemment un atout des enquêtes internationales. Il faut cependant en rappeler la contrainte principale : une enquête internationale est issue d'un consensus entre pays qui empêche d'adapter au mieux le protocole à la population d'un pays particulier. Le chiffrage de l'illettrisme, par exemple, s'appuie sur la mesure de compétences élémentaires, étroitement associée aux caractéristiques de la langue française. Comme nous l'avons indiqué dans la partie « comparaison entre les deux enquêtes » p. 92, la recherche d'une comparabilité internationale dans ce domaine apparaît peu pertinente, en particulier en ce qui concerne le décodage de mots, dont la traduction ne rend pas compte des différences de difficulté de lecture d'un pays à l'autre, en raison notamment de la différence de prononciation, inhérente à la langue considérée.
Dans une enquête nationale, il est aussi plus facile d'intégrer dans le questionnaire biographique des questions précises répondant aux préoccupations des

institutions et des chercheurs français (lien avec la pauvreté, avec la politique de la ville, etc.). Un questionnaire biographique international doit à la fois garantir la comparabilité des concepts et répondre aux demandes des différents partenaires qui s'intéressent à des domaines divers (lien des compétences avec le marché du travail, avec la formation continue, avec la santé, etc.). Cela permet de découvrir des problématiques variées et parfois novatrices dans une perspective française (la mesure indirecte des compétences utilisées au travail, la *Job Requirement Approach* par exemple, est apparue très séduisante), mais cela conduit à proposer aux enquêtés un questionnaire très long, dépassant une heure, ce qui complique la collecte et peut provoquer la lassitude des enquêtés en particulier lors du passage des épreuves.

Enfin, la territorialisation des résultats est un point particulièrement important. L'enquête Piaac permet naturellement des comparaisons internationales, mais l'enquête IVQ a l'intérêt de proposer des déclinaisons régionales, répondant aux besoins des institutions locales. On ne pourrait pas régler le problème posé par la disparition d'IVQ en proposant des extensions régionales de Piaac. En effet, l'intérêt des institutions locales pour IVQ vient surtout de la mesure de l'illettrisme que cette enquête propose. L'enquête Piaac est, on l'a dit, beaucoup moins bien placée pour répondre à ce type de demande.

Les enquêtes IVQ et Piaac apparaissent donc éminemment complémentaires. Bien que le coût important des évaluations d'adultes amène régulièrement à s'interroger sur leur opportunité, la coexistance de ces deux dispositifs est justifiée, tout comme il existe à la fois des évaluations internationales et nationales d'élèves.

BIBLIOGRAPHIE

Besse J.-M., Luis M.-H., Bouchut A.-L., Martinez F., 2009, « La mesure des compétences en traitement de l'écrit chez les adultes en grande difficulté », *Économie et statistique*, n° 424-425, Insee.

Blum A., Guérin-Pace F., 2000, *Des lettres et des chiffres – Des tests d'intelligence à l'évaluation du « savoir lire », un siècle de polémiques*, Paris, Fayard.

Carey S., 2000, *Measuring Adult Literacy – The International Adult Literacy Survey in the European Context*, London, Office for National Statistics.

Degorre A., Murat F., 2009, « La mesure des compétences des adultes, un nouvel enjeu pour la statistique publique », *Économie et statistique*, n° 424-425, Insee.

Djider Z., Murat F., 2006, « Des chiffres pour les hommes… des lettres pour les femmes », *Insee première*, n° 1071, Insee.

Fischer J.-P., Charron C., 2009, « Une étude de la dyscalculie à l'âge adulte », *Économie et statistique*, n° 424-425, Insee.

Green D. A., Riddle W. C., 2001, « Les capacités de lecture et de calcul et la situation sur le marché du travail », *Statistique Canada*, n° 8, coll. « Enquête internationale sur l'alphabétisation des adultes ».

Guérin-Pace F., 2009, « Illettrisme et parcours individuels », *Économie et statistique*, n° 424-425, Insee.

Jonas N., 2013, « Les capacités des adultes à maîtriser des informations écrites et chiffrées – Résultats de l'enquête Piaac 2012 », *Insee Première,* n° 1469, Insee.

Jonas N., 2012, « Pour les générations les plus récentes, les difficultés des adultes diminuent à l'écrit, mais augmentent en calcul », *Insee Première,* n° 1426, Insee.

Jonas N., Lebrère A., Pommier P., Trosseille B., 2013, « Mesurer les compétences des adultes – Comparaison de deux enquêtes », *Insee Analyses,* n° 13, Insee.

Megherbi H., Rocher T., Gyselinck V., Trosseille B., Tardieu H., 2009, « Évaluation de la compréhension de l'écrit chez l'adulte », *Économie et statistique*, n° 424-425, Insee.

Micheaux S., Murat F., 2006, « Les compétences à l'écrit, en calcul et en compréhension orale selon l'âge », *Données Sociales – La société françai*se, Insee.

Murat F., 2009, « Le retard scolaire en fonction du milieu parental : l'influence des compétences des parents », *Économie et statistique*, n° 424-425, Insee.

Murat F., 2008, « L'évaluation des adultes : des méthodes en plein développement », *Éducation & Formations*, n° 78, MEN-DEPP.

Murat F., 2007, « Maîtrise du français et du calcul chez les adultes dans les ZUS », *rapport 2006 de l'ONZUS*.

Murat F., 2006, « Les compétences des adultes et l'exclusion sociale », *Travaux de l'Observatoire – 2005-2006,* ONPES.

Murat F., 2005, « Les compétences des adultes à l'écrit, en calcul et en compréhension orale », *Insee première*, n° 1044, Insee.

Murat F., Rocher T., 2009, « Création d'un score global dans le cadre d'une épreuve adaptative », *Économie et statistique*, n° 424-425, Insee.

Murat F., Zamora P., 2002, « Les performances d'adultes à des tests en lecture : comment séparer motivation et compétences ? », *contribution aux journées de méthodologie statistique*.

NCES, 1998, *Adult Literacy in* OECD *countries: Technical Report on the first International Adult Literacy Survey*, NCES.

Place D., Vincent B., 2009, « L'influence des caractéristiques sociodémographiques sur les diplômes et les compétences », *Économie et statistique*, n°424-425, Insee.

OCDE, 2013, *Perspectives de l'OCDE sur les compétences 2013 – Premiers résultats de l'évaluation des compétences des adultes*, Paris, OCDE.

OCDE, Statistique Canada, 2000, *La littératie à l'ère de l'information : rapport final de l'enquête internationale sur la littératie des adultes*, Paris, OCDE, Statistique Canada.

Vallet L.-A., Bonnet G., Emin J.-C., Levasseur J., Rocher T., Blum A., Guérin-Pace F., Vrignaud P., Haultfoeuille X. D'., Murat F., Verger D., Zamora P., 2002, « Enquête méthodologique Information et vie quotidienne – Tome 1 : bilan du test 1, novembre 2002 », *Document de travail, série Méthodologie statistique*, n° C0202, Insee.

Méthodologie des évaluations

MÉTHODES DE SONDAGES UTILISÉES DANS LES PROGRAMMES D'ÉVALUATIONS DES ÉLÈVES

Émilie Garcia, Marion Le Cam et Thierry Rocher
MENESR-DEPP, bureau de l'évaluation des élèves

Cet article porte sur les méthodes de sondages utilisées à la DEPP dans le cadre des dispositifs d'évaluations standardisées des acquis des élèves. Chaque année, plusieurs échantillons d'élèves sont tirés au sort pour passer ces évaluations. Des problématiques classiques du domaine des sondages se posent, concernant par exemple la définition du champ, les bases de sondage, les modalités de tirage, etc. qui doivent répondre à certaines contraintes pratiques. En outre, dans la mesure où plusieurs échantillons sont tirés à partir des mêmes bases, la question de la coordination de leur tirage doit être traitée. Dans un premier temps, nous présentons les choix faits en matière de méthode de sondage, à toutes les étapes, du tirage des échantillons au redressement de la non-réponse. Dans un second temps, nous conduisons plusieurs simulations qui visent à montrer l'intérêt d'utiliser des informations auxiliaires, c'est-à-dire disponibles pour l'ensemble des élèves. Ces informations peuvent être prises en compte lors du tirage, avec les méthodes d'équilibrage, ou lors du redressement de la non-réponse, avec les méthodes de calage sur marges. Nous montrons que les stratégies prenant en compte l'information auxiliaire, employées dans les évaluations nationales menées par la DEPP, améliorent la qualité des estimateurs, en comparaison d'autres stratégies telles que celles employées dans le cadre des évaluations internationales comme PIRLS ou PISA.

En France, chaque année, la direction de l'évaluation, de la prospective et de la performance (DEPP) conduit des programmes d'évaluation des acquis des élèves [Trosseille et Rocher, dans ce numéro, p. 15]. Il peut s'agir d'évaluations nationales comme les évaluations des compétences du socle commun ou les évaluations Cedre (Cycle des évaluations disciplinaires réalisées sur échantillons), mais également d'évaluations internationales telles que PISA (*Programme for International Student Assessment*) ou PIRLS (*Progress in International Reading Literacy Study*).

Ces évaluations sont réalisées sur des échantillons composés de plusieurs milliers d'élèves, le plus souvent scolarisés soit en fin de CM2, soit en fin de troisième. Le **tableau 1** liste les échantillons concernés pour l'année 2013-2014 et montre qu'au total, près de 80 000 élèves ont été échantillonnés cette année-là. Notre article se concentre sur les programmes d'évaluations des élèves, mais notons que la DEPP est amenée à tirer des échantillons pour répondre à d'autres problématiques, par exemple pour l'enquête nationale de climat scolaire et de victimation [Hubert, 2014].

Tableau 1 Échantillons des programmes d'évaluations des élèves 2014

	Établissements	Élèves
Primaire		
Cedre mathématiques	290	7 952
Cedre maîtrise de la langue	175	4 532
LSE (lecture sur support électronique)	386	9 932
Socle CE1 compétences 1 et 3	628	20 160
TIMSS CM1 - expérimentation	40	1 280
Total	**1 519**	**43 856**
Secondaire		
Cedre mathématiques	323	8 026
Cedre compétences générales et langagières – expérimentation	179	4 551
LSE (lecture sur support électronique)	315	8 070
Socle sixième – expérimentation	391	10 071
PISA – expérimentation	55	2 360
TIMSS *Advanced* – expérimentation	32	1 821
Total	**1 295**	**34 899**

Lecture : en 2014, l'échantillon Cedre mathématiques au primaire comptait 7 952 élèves répartis dans 290 écoles.

L'objet de cet article est double : descriptif et analytique. Tout d'abord, il s'agit de présenter les méthodes de sondage utilisées dans le cadre de ces programmes d'évaluation. Les problématiques soulevées concernent des aspects très divers du domaine des sondages : définition des plans de sondage, modalités de tirage, gestion du recouvrement des échantillons, redressement de la non-réponse, calcul de précision, etc. Des choix ont été opérés pour chacune de ces étapes, en fonction d'un ensemble de contraintes, et font l'objet d'une description précise dans la première partie de cet article. Dans un second temps, nous analysons dans quelle mesure l'échantillonnage peut profiter des informations disponibles dans les bases de données de la DEPP, pour tous les élèves scolarisés en France. La prise en compte de cette information, appelée information auxiliaire, permet en théorie d'améliorer la qualité du tirage ainsi que du redressement de la non-réponse. L'information auxiliaire peut être utilisée en amont au moment du tirage des échantillons (échantillon équilibré) mais aussi en aval pour le redressement de la non-réponse (calage sur marges). Les méthodes de sondage appliquées dans les évaluations internationales ne font pas appel à de l'information auxiliaire du fait de la trop grande diversité des informations disponibles dans chaque pays. Nous vérifions empiriquement l'intérêt de prendre en compte l'information auxiliaire, au moyen de simulations qui portent sur chacune des étapes du sondage : tirage de l'échantillon, redressement de la non-réponse, calcul de précision.

PLANS DE SONDAGE

Pour un échantillon donné, la définition du plan de sondage découle d'un ensemble de contraintes liées aux objectifs de l'évaluation, à la précision recherchée et aux coûts induits. En outre, les informations disponibles dans les bases de sondage impliquent également certaines contraintes.

Champ et exclusions

De manière générale, il est possible de distinguer au moins trois possibilités pour définir la population-cible d'une évaluation :
– les élèves d'un niveau scolaire donné (par exemple, Cedre avec les niveaux CM2 et troisième) ;
– les élèves entrant dans un niveau (par exemple, le panel d'élèves entrant en sixième) ;
– les élèves d'un âge donné (par exemple, l'évaluation PISA avec les élèves de 15 ans quel que soit leur niveau scolaire).
Dans le premier degré, le champ des évaluations comprend les écoles publiques et privées sous contrat en France métropolitaine et DOM, à l'exception du cycle Cedre qui concerne uniquement la France métropolitaine. Sont toujours exclues du champ les écoles des COM, les écoles privées hors contrat, les écoles à l'étranger, les écoles spécialisées. Enfin, pour des raisons de coût, les écoles de moins de 6 élèves du niveau scolaire visé sont exclues de la base de sondage. En guise d'illustration, en 2014, pour l'évaluation des compétences du socle en CE1, il y avait environ 2 500 écoles qui accueillaient moins de 6 élèves de CE1. Ces écoles représentent près de 7,5 % de l'ensemble des écoles, mais elles accueillent moins de 1 % des élèves.
Dans le second degré, les établissements publics et privés sous contrat en France métropolitaine et DOM constituent le champ. Sont toujours exclus les établissements des COM, les établissements privés hors contrat, les établissements à l'étranger, les EREA (Établissements régionaux d'enseignement adapté). Comme pour le premier degré, les évaluations Cedre ne visent que la France métropolitaine. En outre, PISA n'interroge ni La Réunion ni Mayotte, en raison d'une différence de calendrier scolaire.

Les bases de sondage

La base de sondage est la base de données dans laquelle sont tirés les échantillons. Pour les échantillons relevant du second degré, la base de sondage utilisée est la base dite Scolarité construite par la DEPP. C'est une base de données individuelles et anonymes contenant de nombreuses informations sur les élèves scolarisés une année scolaire donnée (date de naissance, profession et catégorie socioprofessionnelle des parents, etc.). Nous disposons également d'informations sur les établissements scolaires (le secteur d'enseignement, par exemple). Ces informations, qualifiées de variables auxiliaires, peuvent être utilisées au moment du tirage des échantillons, pour définir les variables de stratification. Dans le premier degré, le système d'information est nettement moins développé et les plans de sondage s'appuient sur des données qui concernent les écoles, pas les élèves.

Modalités de tirage

Très majoritairement, les plans de sondage sont à deux degrés : un premier degré qui concerne le tirage d'établissements scolaires ou de classes ; un second degré qui concerne le tirage des élèves eux-mêmes. Ce type de sondage est soumis à ce qu'on appelle des effets de grappe : concrètement, les élèves d'un même établissement ou d'une même classe ont tendance à avoir des caractéristiques communes. Ainsi, la variabilité totale va dépendre de la variabilité entre établissements ou entre classes et moins de la variabilité entre élèves. Ce phénomène a une conséquence directe : à effectif égal, un sondage par grappe est moins précis qu'un sondage aléatoire simple qui vise à tirer directement les élèves dans une liste.
Pour le primaire, les écoles sont sélectionnées puis tous les élèves du niveau scolaire visé sont évalués. Il s'agit donc d'un sondage par grappe. Le nombre d'élèves d'un niveau s'élève en moyenne à 25 et ne dépasse que rarement 40 élèves, ce qui limite les effets de grappe.
Pour le secondaire, deux options sont considérées : soit un sondage par grappe en sélectionnant un échantillon de classes et où tous les élèves des classes tirées au sort participent à l'évaluation ; soit un premier degré qui concerne les établissements puis un second degré où un nombre d'élèves fixe dans chaque établissement est sélectionné[1]. Les programmes nationaux suivent la première option tandis que l'évaluation PISA suit la seconde [OCDE, 2012].

Le choix de sondages par grappe est motivé par la facilité de gestion. En effet, le fait de sélectionner tous les élèves d'un niveau scolaire donné au primaire ou tous les élèves d'une classe au collège permet d'éviter de mettre en place des procédures de tirage au sort d'élèves une fois les établissements tirés. Comme la DEPP ne dispose pas de la liste nominative des élèves, et pour être certain que le tirage au sort soit respecté dans les établissements, dans le cadre de PISA, les établissements sélectionnés envoient à la DEPP la liste des élèves de 15 ans, à charge pour la DEPP de sélectionner un nombre fixe d'élèves parmi eux.

Stratification et choix de l'allocation

De manière générale, les échantillons sont stratifiés avec une allocation proportionnelle à la répartition selon la zone de scolarisation, à savoir : public hors éducation prioritaire ; éducation prioritaire ; privé. Cela signifie que le tirage de l'échantillon est réalisé de manière à respecter la répartition des élèves selon la zone de scolarisation de l'établissement scolaire, telle qu'observée sur l'ensemble de la population-cible. Le nombre d'établissements (ou de classes) alloués à chaque zone est ainsi déterminé selon ce principe.

Pour les évaluations des compétences du socle commun, qui alimentent les indicateurs de résultats de la LOLF, un indicateur est demandé pour chacune des quatre strates suivantes : public hors éducation prioritaire ; réseau de réussite scolaire (RRS) ; écoles, collèges et lycées pour l'ambition, l'innovation et la réussite (Éclair) ;

1. Dans ce second cas, les établissements sont tirés au sort proportionnellement à leur taille (nombre d'élèves). En effet, une fois que les établissements sont échantillonnés, un nombre fixe d'élèves est alors sélectionné quel que soit l'établissement. Par conséquent, les élèves des grands établissements ont moins de chance d'être tirés au sort que les élèves des petits établissements. Le tirage proportionnel à la taille permet ainsi de rétablir l'égalité des probabilités de tirage.

privé. Pour ces évaluations, les strates de l'éducation prioritaire vont donc être « sur-représentées » afin de garantir une précision suffisante de nos estimateurs. Par exemple, les élèves des écoles Éclair peuvent représenter environ un quart des élèves de l'échantillon, alors qu'ils représentent en réalité environ 5 % de l'ensemble des élèves.
Le nombre d'élèves sélectionnés est établi en fonction du coût de l'évaluation et du degré de précision attendu. Par exemple, dans le cadre des indicateurs de la LOLF, une précision de 2 points de pourcentage est demandée concernant la proportion d'élèves maîtrisant les compétences du socle commun. L'élaboration du plan de sondage est préparée en fonction de cette attente.

TIRAGE DES ÉCHANTILLONS

De manière classique, on note U la population visée par une évaluation donnée, Y la variable d'intérêt (typiquement le score à l'évaluation, ou bien une indicatrice de difficulté), X une variable auxiliaire, c'est-à-dire connue pour l'ensemble des élèves de la population U. Un échantillon S d'élèves est sélectionné dans la population U. Chaque élève i a la probabilité π_i d'être sélectionné dans l'échantillon S (probabilité d'inclusion). Enfin, les poids de sondages, définis comme les inverses des probabilités d'inclusion π_i, sont notés d_i.

Tirage équilibré

Un échantillon équilibré est un échantillon qui est représentatif de la population au regard de certaines variables auxiliaires. Cela signifie que dans un échantillon équilibré, l'estimateur du total d'une variable auxiliaire X sera exactement égal au vrai total de la variable X dans la population.

Cette propriété s'écrit[2] :

$$\sum_{i \in S} \frac{X_i}{\pi_i} = \sum_{i \in U} X_i \qquad (1)$$

Les échantillons équilibrés ont donc comme propriété de fournir une photographie parfaite de la population, au regard des variables auxiliaires connues, ce que ne garantit pas une procédure aléatoire simple d'échantillonnage. En théorie, ils permettent également d'améliorer la précision des estimateurs s'il existe un lien entre la variable d'intérêt et les variables auxiliaires.

Le tirage équilibré est réalisé grâce au programme CUBE développé par l'Insee et mis à disposition sous forme de macros SAS. La documentation complète est disponible sur le site Internet de l'Insee [Rousseau et Tardieu, 2004]. L'algorithme CUBE permet de choisir de manière aléatoire un échantillon parmi tous les échantillons possibles

2. Le terme de gauche de l'équation (1) représente l'estimateur du total, dit estimateur de Horitz-Thompson. En outre, l'indice i peut représenter ici aussi bien les élèves que les établissements ou les classes, en considérant les sous-totaux par unités comme variables auxiliaires.

respectant les contraintes reposant sur les variables auxiliaires. Il se déroule en deux phases : une « phase de vol » et une « phase d'atterrissage ». Durant la phase de vol, toutes les contraintes sont respectées. Elle se termine si un échantillon équilibré de manière parfaite est trouvé ou s'il n'est pas possible de trouver un échantillon en respectant toutes les contraintes. Si la phase de vol n'a pas abouti à un échantillon, la phase d'atterrissage débute. Elle consiste au relâchement des contraintes et au choix optimal de l'échantillon selon le critère choisi par l'utilisateur (ordre de priorité sur les contraintes, relâchement de la contrainte avec un coût minimal sur l'équilibrage ou garantie d'un échantillon de taille fixe).

Afin de conduire un tirage équilibré, il est nécessaire de disposer d'informations auxiliaires pour l'ensemble des élèves. Concernant les évaluations réalisées à l'école primaire, comme nous l'avons signalé précédemment, les bases de sondage contiennent très peu d'informations. C'est pourquoi les échantillons sont simplement stratifiés selon la zone de scolarisation de l'école, mais ils ne sont pas équilibrés sur d'autres variables.
En revanche, dans le secondaire, de nombreuses informations sont disponibles. Nous réalisons un tirage équilibré des établissements (ou des classes) selon les variables suivantes :
– l'effectif : le nombre total d'élèves du niveau visé dans la population (typiquement, les élèves de troisième) ;
– l'indice de position sociale, qui consiste en une transformation de la PCS des parents des élèves, tel que défini par Le Donné et Rocher [2010] : la somme de cet indice par établissements (ou par classes) est retenue comme variable d'équilibrage ;
– le retard scolaire : le nombre d'élèves en retard du niveau visé ;
– le sexe : le nombre de filles du niveau visé.

Non-recouvrement entre les échantillons

Chaque année, plusieurs programmes d'évaluation étant conduits au même moment, certains établissements peuvent être échantillonnés pour participer à plusieurs évaluations. Afin d'alléger la charge de travail des établissements, et pour assurer un taux de retour maximal, une procédure de non-recouvrement des échantillons est employée. Au moment du tirage d'un échantillon, les écoles ou les collèges dont une classe a déjà été sélectionnée pour une autre évaluation la même année sont exclus de la base de sondage. Les probabilités d'inclusion doivent donc être recalculées pour tenir compte de ces exclusions tout en gardant une représentativité nationale. En outre, il convient d'adapter la procédure de tirage équilibré à cette contrainte de non-recouvrement. L'**encadré** précise la méthode retenue pour corriger les probabilités de sélection.

REDRESSEMENT DE LA NON-RÉPONSE : CALAGE SUR MARGES

Comme toute enquête réalisée par sondage, les évaluations des élèves sont exposées à la non-réponse. Cependant, les taux de participation sont généralement très satisfaisants. Par exemple, pour l'évaluation Cedre histoire-géographie au collège, 96,5 % des établissements ont participé, et au final le taux de réponse des élèves s'est élevé à 90 %.

TIRAGE ÉQUILIBRÉ APRÈS ÉLIMINATION DE LA BASE DES ÉCHANTILLONS PRÉCÉDEMMENT TIRÉS

La situation est la suivante : un échantillon d'établissements a été sélectionné pour participer à une évaluation ; un deuxième échantillon doit être tiré pour une autre évaluation. Nous souhaitons éviter que des établissements soient interrogés deux fois. Il s'agit donc de gérer le non-recouvrement entre les échantillons et d'assurer également un tirage équilibré du deuxième échantillon.

Nous nous concentrons ici sur le non-recouvrement des échantillons, mais notons qu'une approche plus générale incluant un taux de recouvrement non nul (pour permettre des analyses croisées entre enquêtes) est en cours de développement avec une application à des données issues d'évaluations standardisées [Christine et Rocher, 2012].

Formulation du problème et notations

Un échantillon S_1 a été tiré. Il est connu et les probabilités d'inclusion des établissements π_j^1 sont également connues.

On souhaite alors tirer un échantillon S_2 dans la population U avec les probabilités π_j^2, mais sans aucun recouvrement avec l'échantillon S_1.

On va donc tirer l'échantillon S_2 dans la population $U(S_1)$, c'est-à-dire la population U privée des établissements de l'échantillon S_1 qui appartiennent à U. Notons d'emblée que S_1 n'a pas nécessairement été tiré dans U, mais potentiellement dans une autre population, plus large ou plus réduite ; cela n'affecte en rien la formulation envisagée ici.

Notons également que l'indice j est utilisé ici : il concerne les établissements et non les élèves, représentés par l'indice i.

Il s'agit donc de procéder à un tirage conditionnel. On note π_j^{2/S_1} les probabilités d'inclusion conditionnelles des établissements dans le second échantillon S_2, sachant que le premier échantillon est connu.

Ces probabilités conditionnelles peuvent s'écrire :

$$\pi_j^{2/S_1} = \begin{cases} \lambda_j & \text{si } j \notin S_1 \\ 0 & \text{si } j \in S_1 \end{cases}, \text{ avec } \lambda_j \in [0,1]$$

On a $\pi_j^2 = E(\pi_j^{2/S_1}) = \lambda_j(1-\pi_j^1)$, d'où : $\lambda_j = \dfrac{\pi_j^2}{1-\pi_j^1}$

Équilibrage

On souhaite maintenant que l'échantillon S_2 soit équilibré selon certaines variables (nombre d'élèves en retard, etc.). Soit X une variable d'équilibrage, la condition s'écrit :

$$\sum_{j \in S_2} \frac{X_j}{\pi_j^2} = \sum_{j \in U} X_j$$

Pour arriver à ce résultat, le principe est de tirer S_2 dans $U(S_1)$ avec les probabilités d'inclusion λ_j et avec une condition d'équilibrage sur la variable $X_j / (1-\pi_j^1)$.

Ainsi, on aura :

$$\sum_{j \in S_2} \frac{X_j}{\pi_j^2} = \sum_{j \in S_2} \frac{X_j}{\lambda_j(1-\pi_j^1)} = \sum_{j \in U(S_1)} \frac{X_j}{(1-\pi_j^1)}$$

Or, en espérance, on a :

$$E\left(\sum_{j \in U(S_1)} \frac{X_j}{(1-\pi_j^1)} \right) = E\left(\sum_{j \in U} \frac{X_j}{(1-\pi_j^1)} I_{j \notin S_1} \right) = \sum_{j \in U} X_j$$

La condition d'équilibrage initiale est donc remplie.

Condition fondamentale

Comme il s'agit d'une probabilité, la condition fondamentale est que $\lambda_j \in [0,1]$.

Comme $\lambda_j = \dfrac{\pi_j^2}{1-\pi_j^1}$, la condition est en fait que :

$$\pi_j^1 + \pi_j^2 \leq 1$$

Dans certains cas, par exemple, des strates souvent sur-représentées comme les établissements situés dans des zones spécifiques concernant peu d'élèves (ex. : Éclair), cette condition pourrait ne pas être satisfaite. Cependant, de façon concrète, la condition a toujours été respectée dans les plans de sondage réalisés.

Bien que ces taux de retour soient élevés, il est nécessaire de tenir compte de la non-réponse dans les estimations, car celle-ci n'est pas purement aléatoire (par exemple, la non-réponse est plus élevée chez les élèves « en retard »). Afin de la prendre en compte, un calage sur marges est effectué à l'aide de la macro CALMAR, également disponible sur le site Internet de l'Insee. La méthode de calage sur marges consiste à modifier les poids de sondage d_i des répondants de manière à ce que l'échantillon ainsi repondéré soit représentatif de certaines variables auxiliaires dont on connaît les totaux sur la population [Sautory, 1993]. C'est une méthode qui permet de corriger la non-réponse, mais également d'améliorer la précision des estimateurs. En outre, elle a pour avantage de rendre cohérents les résultats observés sur l'échantillon pour ce qui concerne des informations connues sur l'ensemble de la population.
Les nouveaux poids w_i, calculés sur l'échantillon des répondants S', vérifient l'équation suivante pour les K variables auxiliaires sur lesquelles porte le calage :

$$\forall k = 1...K, \sum_{i \in S'} w_i X_i^k = \sum_{i \in U} X_i^k \qquad (2)$$

Ils sont obtenus par minimisation de l'expression $\sum_{i \in S'} d_i G(\frac{w_i}{d_i})$ où G désigne une fonction de distance, sous les contraintes définies dans l'équation (2).
Pour le redressement de la non-réponse, seules les marges des variables sont nécessaires, c'est-à-dire leur somme sur l'ensemble de la population. À l'école comme au collège, les variables auxiliaires de calage utilisées sont :
– le nombre total d'élèves du niveau visé dans la population ;
– le nombre d'élèves du niveau visé en retard dans la population ;
– le nombre de garçons du niveau visé dans la population.
Au collège, le nombre d'élèves de classe sociale défavorisée est également introduit.

CALCUL DE PRÉCISION

Les résultats des évaluations sont soumis à une variabilité qui dépend notamment des erreurs d'échantillonnage. Il est possible d'estimer statistiquement ces erreurs d'échantillonnage et de produire des intervalles de confiance sur les différents estimateurs calculés.

On note Y la variable d'intérêt (typiquement le score obtenu à une évaluation) et $\hat{\bar{Y}}$ l'estimateur de la moyenne de Y, qui constitue un estimateur essentiel sur lequel nous insistons dans la suite, bien que d'autres soient également au centre des analyses, comme ceux concernant la dispersion. La méthode retenue est cependant applicable à différents types d'estimateurs.
Nous souhaitons estimer la variance de cet estimateur, c'est-à-dire $V(\hat{\bar{Y}})$. En absence de formule théorique pour calculer $V(\hat{\bar{Y}})$, il existe plusieurs procédures permettant de l'estimer, c'est-à-dire de calculer $\hat{V}(\hat{\bar{Y}})$, l'estimateur de la variance d'échantillonnage. Il peut s'agir de méthodes de linéarisation des formules (Taylor) ou bien de méthodes empiriques (méthodes de réplication, jackknife, etc.). Ces méthodes sont bien décrites dans la littérature. Le lecteur est invité à consulter Tillé [2001] ou Ardilly [2006].

Cependant, lorsqu'un calage sur marges a été effectué, il faut en tenir compte pour le calcul de la précision. Dans ce cas, la variance de $\hat{\bar{Y}}$ est asymptotiquement équivalente à la variance des résidus de la régression de la variable d'intérêt sur les variables de calage [Deville et Särndal, 1992]. En pratique, pour estimer la variance d'échantillonnage de $\hat{\bar{Y}}$, tenant compte du calage effectué, il convient alors d'appliquer la procédure suivante :
1 – On effectue la régression linéaire de la variable d'intérêt sur les variables de calage, en pondérant par les poids initiaux. Les résidus e_i de cette régression sont calculés.
2 – Les valeurs $g_i e_i$ sont calculées, où g_i représente le rapport entre les poids CALMAR (w_i) et les poids initiaux (d_i) : $g_i = \frac{w_i}{d_i}$.
3 – La variance d'échantillonnage de $\hat{\bar{Y}}$ est alors obtenue en calculant la variance d'échantillonnage de $g_i e_i$.

LES AVANTAGES DE L'UTILISATION DE L'INFORMATION AUXILIAIRE : TROIS SIMULATIONS

Comme nous l'avons vu en première partie, dans le second degré, la DEPP dispose de bases de sondage avec de l'information disponible sur l'ensemble de la population. L'information auxiliaire est utilisée à chaque étape, du tirage des échantillons au calcul de précision. Cette approche est assez spécifique aux programmes nationaux d'évaluation. En effet, les évaluations internationales, telles que PISA, empruntent une autre voie, qui n'utilise que très peu l'information auxiliaire. Les responsables de ces enquêtes avancent que la disponibilité d'informations auxiliaires est très variable selon les pays et que les variables elles-mêmes sont différentes selon les pays, ce qui nécessiterait une investigation coûteuse pour définir les méthodes, car personnalisée pour chaque pays.

Dans cette partie, nous souhaitons cependant montrer l'intérêt de recourir à l'information auxiliaire, ainsi que cela est fait dans les programmes nationaux d'évaluation. Trois simulations sont présentées : la première concerne le tirage équilibré des échantillons, la seconde le redressement de la non-réponse *via* un calage sur marges et la troisième le calcul de précision. Ces simulations sont menées de manière à évaluer les deux types de procédures : celle utilisée dans les évaluations internationales (tirage non équilibré, redressement par ajustement, calcul simple de précision) et celle utilisée dans les programmes nationaux tels que Cedre (tirage équilibré, redressement par calage sur marges, calcul de précision tenant compte du calage). Ainsi, les simulations ne sont pas envisagées de manière indépendante, mais dans l'optique d'une comparaison globale des deux approches.
Les simulations ont été réalisées sur la base exhaustive des notes obtenues à l'examen terminal du brevet en 2009 en mathématiques, français et histoire-géographie. Dans cette base, nous disposons aussi des caractéristiques des élèves

(sexe, année de naissance, PCS des parents) et des établissements. Ces informations serviront de variables auxiliaires. Les notes fournissent un *proxy* intéressant par rapport aux scores obtenus aux évaluations, car portant sur des objets proches en principe. Nous estimons ainsi qu'en matière de sondage, les deux formes d'évaluation – scores aux évaluations et notes à l'examen – devraient conduire aux mêmes types de conclusion. Pour nos simulations, les variables d'intérêt sont donc les notes à l'examen, à défaut des scores obtenus aux évaluations standardisées.

Simulation 1 : impact du tirage équilibré

Pour montrer les avantages de l'utilisation de l'information auxiliaire lors du tirage des échantillons, nous avons comparé quatre stratégies d'échantillonnage :
1 – sondage à probabilités proportionnelles à la taille sur les établissements, puis tirage aléatoire simple de 30 élèves dans chacun des établissements ;
2 – sondage à probabilités proportionnelles à la taille et équilibré des établissements, puis tirage aléatoire simple de 30 élèves dans chacun des établissements ;
3 – tirage aléatoire simple de classes, puis tous les élèves des classes tirées au sort participent à l'évaluation ;
4 – tirage équilibré de classes, puis tous les élèves des classes tirées au sort participent à l'évaluation.

Le premier plan de sondage est celui utilisé pour PISA tandis que le dernier est celui utilisé à la DEPP. Les plans 2 et 4 sont les versions équilibrées des plans 1 et 3. Tous les échantillons sont stratifiés selon le secteur d'enseignement : public hors éducation prioritaire, RAR (réseaux ambition réussite), RRS (réseaux de réussite scolaire) et privé. Dans chacune des quatre strates, 2 000 élèves sont sélectionnés. En effet, il est souvent demandé un indicateur pour chacune de ces strates. Les variables d'intérêt sont les notes au brevet obtenues en français, en mathématiques et en histoire-géographie ; les paramètres d'intérêt sont les moyennes et les pourcentages d'élèves obtenant une note inférieure à un certain seuil.
Pour les plans de sondage 2 et 4, les variables auxiliaires utilisées pour l'équilibrage sont :
– le sexe ;
– le fait d'être en retard ;
– l'indice social moyen de l'établissement (en quatre groupes selon les quartiles).

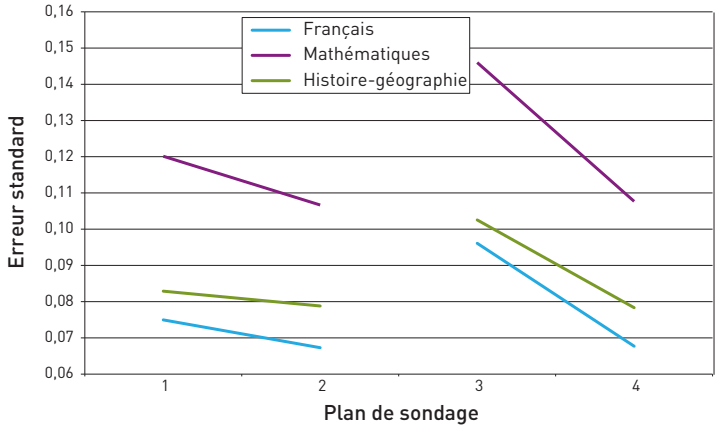
Pour chaque plan de sondage, 1 000 échantillons ont été tirés. Pour chaque échantillon s on calcule la moyenne des notes $\widehat{\bar{y}_s}$, ensuite on calcule la moyenne des $\widehat{\bar{y}_s}$ ainsi que l'écart-type des $\widehat{\bar{y}_s}$ qui correspond à l'erreur standard, c'est-à-dire à la racine carrée de la variance d'échantillonnage.
Le **tableau 2** présente les notes moyennes estimées pour chacun des plans de sondage et les compare avec la note moyenne réelle de l'ensemble de la population. Pour chaque simulation, on retrouve la note moyenne observée dans la population. Dans le cadre des plans de sondage employés, la moyenne simple sur l'échantillon est en effet un estimateur sans biais de la moyenne sur la population. On observe de légers écarts (à la deuxième décimale) pour les proportions d'élèves ayant obtenu une note inférieure à un certain seuil.
Si les biais sont très faibles, voire nuls, la précision est variable selon les plans de sondage considérés. Ainsi, comme le montre la **figure 1**, l'erreur standard est plus

Tableau 2 Simulation 1 – notes moyennes selon les plans de sondage

	Population	Plans de sondage			
		1	2	3	4
Français	11,41	11,41	11,41	11,42	11,41
Mathématiques	9,68	9,68	9,68	9,68	9,68
Histoire-géographie	10,74	10,74	10,74	10,74	10,74
Français < 8 (en %)	13,82	13,82	13,85	13,80	13,82
Mathématiques < 6 (en %)	21,49	21,52	21,52	21,51	21,48
Histoire-géographie < 8 (en %)	20,46	20,45	20,50	20,42	20,44

Lecture : la note moyenne des élèves en français est de 11,41 sur l'ensemble de la population.
La moyenne des 1 000 estimations de la note moyenne en français pour le plan de sondage 1 est aussi de 11,41.
Le pourcentage d'élèves ayant une note de français inférieure à 8 est au total de 13,82 % ; les quatre plans de sondage donnent des valeurs proches en moyenne.

Figure 1 Simulation 1 – erreurs standard selon les plans de sondage



Lecture : en mathématiques, l'erreur standard de la note est d'environ 0,12 pour le premier plan de sondage et elle est inférieure à 0,11 pour le deuxième.

faible dans le cas d'un sondage équilibré (plans 2 et 4). Dès lors que l'on utilise de l'information auxiliaire pour réaliser un tirage équilibré, la précision est nettement améliorée. Ainsi, il apparaît que l'enquête PISA gagnerait à adopter une démarche de tirage équilibré (plan 2 par rapport au plan 1 de PISA). Ce constat est encore plus marqué dans le cas où le tirage au premier degré concerne des classes et pas des établissements : le fait d'échantillonner des classes au lieu d'établissements dégrade la précision, en l'absence de tirage équilibré (plan 3 en comparaison du plan 1). Ce phénomène est à relier au fait que l'effet de grappe est plus important avec un tirage de classes qu'avec un tirage d'établissements puis d'élèves. En revanche, dès lors que l'on utilise l'information auxiliaire disponible, le mode de tirage (établissements ou classes) conduit à des résultats comparables du point de vue de la précision (plans 2 et 4). Ce point est important pour les aspects logistiques gérés par la DEPP qui sont moins coûteux dans le cadre d'un tirage de classes entières plutôt que d'un tirage d'établissements puis d'élèves.

Simulation 2 : impact du calage sur marge

Dans cette partie, au-delà du choix de tirage des échantillons, nous comparons également deux stratégies de repondération en présence de non-réponse. La première, utilisée dans le cadre des évaluations internationales, est basée sur des coefficients d'ajustement concernant les établissements et les élèves, de manière à ce que les répondants représentent les non-répondants. La seconde approche, utilisée à la DEPP, emploie une procédure de calage sur marges qui consiste à modifier les poids de sondage des élèves de manière à ce que l'échantillon soit représentatif de la population au regard de certaines variables auxiliaires choisies.

Plus précisément, concernant la première approche, nous avons repris la démarche suivie dans les évaluations internationales PIRLS [Martin, Mullis, Kennedy, 2007] ou PISA [OCDE, 2012], et également appliquée sur les premières évaluations du cycle Cedre. Il s'agit d'appliquer aux poids de sondage des coefficients d'ajustement. Pour un élève i d'un l'établissement j, nous définissons les nouveaux poids de la manière suivante[3] :

$$w_{ij}^{1} = f_{1j} f_{2ij} d_{2ij} d_{1j} \qquad (3)$$

avec :

– d_{1j} le poids de sondage initial de l'établissement j ;

– d_{2ij} le poids de sondage initial de l'élève i au sein de l'établissement j ;

– f_{1j} le coefficient d'ajustement pour la non-réponse des établissements (rapport entre le nombre d'établissements de la strate échantillonnés au départ et le nombre d'établissements répondants de la strate) ;

– f_{2ij} le coefficient d'ajustement de la non-réponse des élèves au sein des établissements répondants (rapport entre le nombre d'élèves échantillonnés de l'établissement et le nombre d'élèves répondants), distingué selon le sexe.

La deuxième approche repose sur un calage sur marges selon les totaux des distributions suivantes :

– le nombre d'élèves dans chaque strate ;

– la répartition par sexe dans la population ;

– le nombre d'élèves en retard dans la population ;

– le nombre d'élèves dans chaque quartile de la population découpée grâce à l'indice social.

Afin de comparer ces deux stratégies dans le cadre d'une nouvelle simulation, nous utilisons la même base de sondage que pour la simulation 1. Notre simulation procède selon les étapes suivantes :

1 – tirage d'un échantillon ;

2 – génération de non-réponse ;

3 – repondération des élèves répondants.

Concernant le **tirage des échantillons**, nous avons repris deux plans de sondage utilisés précédemment. Le premier, celui de PISA, est un sondage proportionnel à la taille des établissements puis une sélection aléatoire de 30 élèves dans chacun des établissements (plan de sondage 1 de la simulation précédente). Le second plan de

3. C'est une version simplifiée de ce qui est fait dans PIRLS et PISA où les coefficients d'ajustement sont plus nombreux, mais la démarche est similaire.

sondage est un échantillon équilibré sur les établissements puis sélection de 30 élèves dans chacun des établissements (plan de sondage 2 de la simulation précédente). Nous n'avons pas retenu le plan de sondage 4 utilisé à la DEPP car il est fondé sur un échantillon de classes et non d'établissements. Ainsi, les modalités de modélisation de la non-réponse n'auraient pas été comparables. En outre, nous avons pu observer que les plans de sondage 2 et 4 étaient assez proches en termes de précision. Au final, les résultats de la simulation 2 nous renseigneront sur l'impact de la procédure de repondération mais dans deux cadres de tirage différents, l'un équilibré et l'autre pas. Cette simulation nous permet ainsi de nous prononcer sur une démarche globale intégrant le tirage et la repondération.

S'agissant de la **génération de non-réponse**, nous avons au préalable modélisé la non-réponse, sachant qu'elle comporte deux types : celle des établissements scolaires échantillonnés et celle des élèves au sein des établissements répondants. Ces non-réponses ont été modélisées à l'aide de régressions logistiques appliquées à des données réelles de programmes d'évaluation.
Pour caractériser la non-réponse des établissements, nous avons utilisé l'évaluation Cedre compétences générales de 2009 en fin de troisième. Pour cette évaluation, les établissements qui n'ont pas répondu sont connus, en particulier nous disposons pour chacun d'entre eux de la note moyenne au brevet et de l'indice moyen de position socio-scolaire. Le modèle retenu prédit la non-réponse des établissements en fonction de ces deux variables.
Pour caractériser la non-réponse des élèves au sein des établissements répondants, nous avons dû utiliser une autre opération : l'évaluation Cedre histoire-géographie de 2012 en fin de troisième. Pour cette évaluation, nous disposons d'informations sur les élèves non-répondants, en particulier le sexe, le retard scolaire et l'origine sociale. La seule variable significative du modèle explicatif de la non-réponse est le retard scolaire, variable retenue pour la génération de non-réponse des élèves, qui a consisté à tirer au sort des élèves non-répondants, en distinguant les élèves, « en retard » des élèves « à l'heure ».

Pour chaque **stratégie de repondération**, 1 000 échantillons ont été tirés. La première stratégie est envisagée à partir du premier plan de sondage (PISA) et la seconde stratégie est appliquée au deuxième plan de sondage (PISA équilibré). Pour chaque échantillon de chaque stratégie, on calcule la moyenne des notes aux épreuves finales du brevet. Au final, le **tableau 3 p. 114** présente la moyenne de ces moyennes. Il ressort que la première stratégie, fondée sur les coefficients d'ajustement, apparaît légèrement biaisée, surtout pour les fractiles, c'est-à-dire les indicateurs concernant le pourcentage d'élèves en-deçà d'un certain seuil de note.

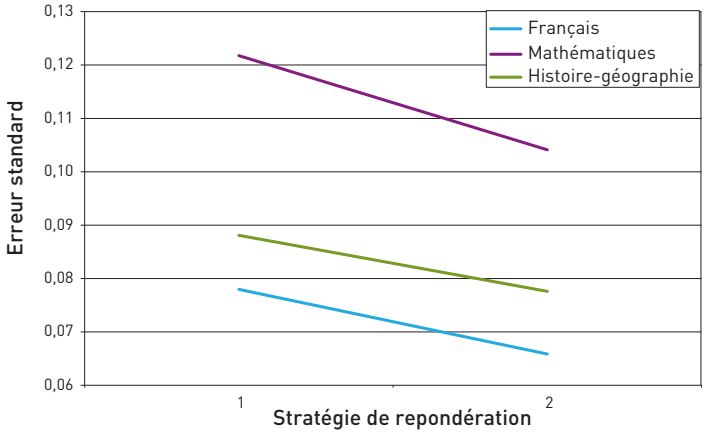
Au-delà du biais, nous avons également étudié la précision respective de ces deux stratégies. Nous avons ainsi calculé l'écart-type des 1 000 moyennes des échantillons simulés, c'est-à-dire l'erreur standard de la note moyenne. Il ressort que l'erreur standard est plus faible dans le cas d'un sondage équilibré et d'un redressement de la non-réponse avec un calage sur marges ▸ **Figure 2 p. 114**. Notons que la stratégie 2 conduit à des niveaux de précision comparables à ceux observés dans le cas de sondages équilibrés sans non-réponse (plans 2 et 4 de la figure 1). La stratégie 2 corrige donc la non-réponse de manière très satisfaisante tandis que la stratégie 1 est moins efficace.

Tableau 3 Simulation 2 – moyennes après repondération

Notes		Population	Stratégies	
			1	2
Français		11,41	11,44	11,41
Mathématiques		9,68	**9,78**	9,70
Histoire-géographie		10,74	**10,79**	10,75
Français < 8	(en %)	13,82	**13,54**	13,84
Mathématiques < 6	(en %)	21,49	**20,77**	**21,42**
Histoire-géographie < 8	(en %)	20,46	**20,02**	20,45

Lecture : la note moyenne des élèves en français est de 11,4. La moyenne des 1 000 estimations de la note moyenne en français pour la stratégie 1 est de 11,44.

Figure 2 Simulation 2 – erreurs standard selon la stratégie de repondération



Lecture : en mathématiques, l'erreur standard de la note est d'environ 0,12 pour la première stratégie de repondération et elle est supérieure à 0,10 pour la seconde.

Ces écarts peuvent s'expliquer par le fait que dans le cadre de la première stratégie, le calcul des coefficients d'ajustement repose sur l'hypothèse que les répondants et les non-répondants sont « similaires », s'agissant des élèves au sein d'un établissement, ou des établissements au sein d'une strate. Cette hypothèse a pu induire un biais dans l'estimation, dans la mesure où certaines variables individuelles, comme le retard scolaire, expliquent la non-réponse, quels que soient les établissements.

Quoi qu'il en soit, la comparaison des résultats de la simulation 1 avec ceux de la simulation 2 conduit à souligner l'importance du redressement de la non-réponse. En effet, la stratégie 2 donne des niveaux de précision du même ordre que ceux du plan 2 de la simulation 1, ce qui montre que le calage sur marges corrige parfaitement la non-réponse. En revanche, la méthode de redressement de la stratégie 1 est moins efficace et conduit à des niveaux de précision moins élevés.

Simulation 3 : impact de l'utilisation de l'information auxiliaire sur le calcul de la précision

Dans cette dernière section, nous nous intéressons aux procédures de calcul de précision. En effet, la publication des résultats des évaluations standardisées est accompagnée d'une estimation des erreurs liées à l'échantillonnage. Plus précisément, nous calculons l'erreur standard – soit la racine carrée de la variance d'échantillonnage – des différents paramètres d'intérêt, tels que le score moyen. De nombreuses méthodes de calcul existent. Dans le cadre des évaluations standardisées, les méthodes employées sont de nature empirique, car il peut être délicat de déterminer la formule théorique de l'erreur standard, au vu des plans de sondage et des redressements effectués. Nous avons conduit une simulation de manière à déterminer quelle était la meilleure méthode pour effectuer les calculs de précision pour chacune des deux stratégies de repondération présentées précédemment.

Dans le cadre de la simulation 2, pour chacun des 1 000 échantillons de chaque stratégie de repondération, nous avons utilisé la procédure *Surveymeans* de SAS pour calculer la précision des estimateurs[4]. Comme nous l'avons vu dans la section précédente, lorsqu'un calage sur marges a été effectué, il faut en tenir compte dans le calcul de la précision. Nous avons alors, pour chacune des deux méthodes, calculé la moyenne des 1 000 valeurs obtenues pour estimer l'espérance de l'erreur-standard en fonction de la méthode retenue. Par ailleurs, nous considérons que, pour une variable donnée, l'erreur standard « vraie » peut être approchée par l'écart-type de la distribution des 1 000 estimateurs obtenus par simulations.

Le **tableau 4** montre que, dans le cas de la première stratégie, le calcul de précision conduit à surestimer l'erreur standard, en particulier pour les fractiles. En revanche, dans le cadre de la deuxième stratégie procédant par calage sur marges, l'emploi de la procédure reposant sur les $g_i e_i$ présentés dans la partie précédente, permet d'aboutir à des estimations correctes de la précision, avec cependant là encore de légers biais subsistant en ce qui concerne les fractiles ▸ **Tableau 5**. Des investigations supplémentaires mériteraient d'être conduites afin d'identifier les raisons des biais d'estimation observés selon la stratégie 1 (tableau 4). Une explication pourrait

▸ **Tableau 4** **Stratégie 1 : précision pour un sondage à allocation proportionnelle à la taille et un redressement de la non-réponse avec des coefficients d'ajustement**

	Erreur standard	Estimation
Français	0,08	0,09
Mathématiques	0,12	0,14
Histoire-géographie	0,09	0,10
Français < 8	0,73	0,83
Mathématiques < 6	0,94	1,07
Histoire-géographie < 8	0,87	0,96

Lecture : l'erreur standard de la note moyenne en français est de 0,08 (*i.e.* l'écart-type de la distribution des 1 000 notes moyennes obtenues). La moyenne des 1 000 estimations de la précision de la note moyenne en français est de 0,09.

4. Nous avons utilisé deux méthodes, l'une empirique (Jackknife), l'autre par linéarisation des formules de variance (Taylor). Elles donnent des résultats quasi-identiques, nous reproduisons ici la méthode empirique (Jackknife).

Tableau 5 Stratégie 2 : précision pour un tirage équilibré et un redressement de la non-réponse avec un calage sur marges

	Erreur standard	Estimation
Français	0,07	0,07
Mathématiques	0,10	0,11
Histoire-géographie	0,08	0,08
Français < 8	0,70	0,72
Mathématiques < 6	0,89	0,93
Histoire-géographie < 8	0,83	0,85

Lecture : l'erreur standard de la note moyenne en français est de 0,07. La moyenne des 1 000 estimations de la précision de la note moyenne en français est de 0,07.

être avancée en observant que le redressement effectué avec les coefficients d'ajustement consiste en une forme de correction comparable à un « calage » (sur le sexe et la variable de stratification), correction qui n'est pas prise en compte explicitement dans le calcul de précision.

CONCLUSION

Plusieurs types de questions se posent au moment du tirage d'un échantillon et les choix sont toujours contraints. Concernant les évaluations, une des premières contraintes porte sur les bases de sondage disponibles. Pour le second degré, en France, elles sont relativement riches et elles permettent d'utiliser l'information auxiliaire pour le tirage de l'échantillon et pour le redressement de la non-réponse. Ce n'est malheureusement pas le cas pour les évaluations réalisées dans le premier degré, car les bases de données sont très limitées.

Or, les simulations montrent que l'utilisation de l'information auxiliaire, quand elle est pertinente, permet d'améliorer la précision des estimateurs. Lorsque des contraintes pratiques dégradent la précision, comme le fait de sélectionner des classes plutôt que des établissements pour les évaluations nationales, les simulations montrent que l'utilisation de l'information auxiliaire dans ce cas permet de ne pas perdre en précision.

Ces résultats doivent interroger les pratiques des évaluations internationales qui mobilisent très peu l'information auxiliaire, avec l'argument de la standardisation, c'est-à-dire de la même procédure appliquée dans tous les pays, même si certains pays disposent de nombreuses informations susceptibles d'améliorer la qualité des échantillons. Or, des procédures différentes selon les pays pourraient être étudiées sans que cela nuise à la comparabilité des résultats, mais au contraire pour que chaque pays optimise son plan d'échantillonnage.

BIBLIOGRAPHIE

Ardilly P., 2006, *Les techniques de sondage*, Paris, Technip.

Christine M., Rocher T., 2012, « Construction d'échantillons astreints à des conditions de recouvrement par rapport à un échantillon antérieur et à des conditions d'équilibrage par rapport à des variables courantes : aspects théoriques et mise en œuvre dans le cadre du renouvellement des échantillons des enquêtes d'évaluation des élèves », *Journées de Méthodologie Statistique*, Paris, janvier 2012.

Deville J. C., Särndal C. E., 1992, "Calibration Estimators in Survey Sampling", *Journal of the American Statistical Association*, vol. 87, No. 418, p. 376-382.

Hubert T., 2014, « Un collégien sur cinq concerné par la "cyber-violence" », *Note d'information*, n ° 39, MENESR-DEPP, Paris.

Le Donné N., Rocher T., 2010, « Une meilleure mesure du contexte socio-éducatif des élèves et des écoles – Construction d'un indice de position sociale à partir des professions des parents », *Éducation et formations*, n° 79, MENJVA-DEPP, p. 103-115.

Martin M., Mullis I., Kennedy A., 2007, *PIRLS 2006 Technical Report*, Chestnut Hill, TIMSS & PIRLS International Study Center, Boston College.

OCDE, 2012, *PISA 2009 – Technical Report*, Paris, OCDE.

Rousseau S., Tardieu F., 2004, *La macro SAS CUBE d'échantillonnage équilibré – Documentation de l'utilisateur*, Paris, Insee.

Sautory O., 1993, « La macro CALMAR – Redressement d'un échantillon par calage sur marges », *Série des documents de travail*, Document n° F9310, Paris, Insee.

Tillé Y., 2001, *Théorie des sondages – Échantillonnage et estimation en populations finies – Cours et exercices avec solutions*, Paris, Dunod.

LA MOTIVATION DES ÉLÈVES FRANÇAIS FACE À DES ÉVALUATIONS À FAIBLES ENJEUX

Comment la mesurer ? Son impact sur les réponses[1]

Saskia Keskpaik et **Thierry Rocher**
MENESR-DEPP, bureau de l'évaluation des élèves

Les évaluations standardisées des élèves, telles que Cedre ou PISA, renvoient à des enjeux politiques croissants, alors qu'elles restent à faibles enjeux pour les élèves participants. Dans le système éducatif français, où la notation tient une place prépondérante, la question de la motivation des élèves face à ces évaluations mérite d'être posée. En 2011, afin d'explorer cette question, une expérience a été menée en France à partir du test PISA. Suite à cette expérience, un instrument pour mesurer la motivation a été adapté à partir du « thermomètre d'effort » proposé dans PISA.
Cet instrument a été introduit dans plusieurs évaluations conduites au niveau national par la DEPP, sur des échantillons de plusieurs milliers d'élèves, en fin de primaire (CM2) et en fin de collège (troisième).
Ces données permettent de distinguer la motivation de l'élève de la difficulté perçue du test, et ainsi de mieux appréhender le lien entre la motivation des élèves français et leur performance. L'analyse de ces données renseigne en outre sur le rôle de certaines caractéristiques, des élèves ou des évaluations elles-mêmes, dans le degré de motivation à répondre aux questions de l'évaluation.

Dans le système éducatif français où la notation tient une place prépondérante, la question de la motivation des élèves face à une évaluation sans enjeux pour eux mérite d'être posée. En effet, si les élèves ne sont pas motivés à faire de leur mieux lors de telles évaluations, la validité des résultats et leur interprétation peuvent être interrogées. De même, lorsque certains élèves ou des sous-populations d'élèves s'avèrent systématiquement moins motivés que d'autres, la comparabilité des résultats risque d'être biaisée.

1. Cet article fait suite à une communication lors du congrès international « Actualité de la Recherche en Éducation et en Formation » (AREF) à Montpellier en août 2013.

Plusieurs travaux ont été mis en œuvre, ces dernières années, pour étudier la motivation des élèves à répondre à des tests à faibles enjeux pour eux ainsi que la relation entre cette motivation et la performance [Butler et Adams, 2007 ; Eklöf, 2008 ; O'Neil *et alii*, 2004 ; Penk, Poehlmann, Roppelt, 2013]. Ces études, divergentes dans leurs conclusions, varient considérablement selon les méthodes utilisées ainsi que selon les instruments employés pour mesurer la motivation. Malgré un intérêt croissant pour le sujet, peu de tentatives ont été menées afin de construire une mesure valide de la motivation à répondre à des tests [Ndinga et Frenette, 2010, Eklöf, 2008].

En 2011, une expérience a été menée en France à partir du pré-test de PISA 2012[2] [Keskpaik et Rocher, 2012]. Suite à cette expérience, un instrument pour mesurer la motivation a été adapté du « thermomètre d'effort » de PISA. Cet instrument a ensuite été introduit dans de nombreuses évaluations conduites, en 2012, au niveau national par la DEPP[3] sur des échantillons de plusieurs milliers d'élèves, en fin de primaire (CM2) et en fin de collège (troisième). La validité de cet instrument a été interrogée lors d'une étude qualitative menée par la DEPP en mai 2013 ▸ **Encadré p. 123**.

La présente étude s'intéresse aux résultats de ces évaluations, notamment aux réponses des élèves aux questions relatives à la motivation, et à la variation de ces réponses selon diverses caractéristiques des évaluations, des élèves et des établissements. Nous déterminons des profils d'élèves et étudions le lien entre la motivation des élèves face à ces évaluations et les scores qu'ils y ont obtenus.

MESURE DE MOTIVATION : « EFFORT » VERSUS « APPLICATION »

Effort, application et difficulté

PISA mesure l'investissement des élèves face au test à l'aide d'un « thermomètre d'effort » ▸ **Figure 1**. Une étude exploratoire des données de ce thermomètre [Keskpaik et Rocher, 2012] nous a suggéré que le terme « effort » est susceptible de poser un problème d'interprétation, en mélangeant la motivation de l'élève avec la difficulté du test. Ainsi, les élèves performants ont pu déclarer faire peu « d'effort », car les exercices proposés leur ont paru faciles. Suite à cette étude, un instrument de mesure de la motivation a été adapté à partir du « thermomètre » de PISA. Plus précisément, les énoncés des questions ont été simplifiés pour réduire la charge de lecture, le terme « effort » a été remplacé par le terme « application », les échelles ont été placées horizontalement, et une échelle supplémentaire a été ajoutée afin d'interroger les élèves sur la difficulté des exercices proposés. Cet instrument varie légèrement selon le niveau scolaire. Au collège, les échelles sont sur dix positions ▸ **Figure 2**. En revanche, à l'école, une échelle sur quatre positions a été préférée, jugée plus adaptée aux élèves de l'école primaire.
L'analyse des réponses d'élèves à ces items « d'application » indique tout d'abord que la motivation au test (question 2 de l'instrument) est liée à la difficulté perçue du test

2. Dans chaque cycle PISA un pré-test, appelé *Field Trial*, est organisé un an avant le test principal (*Main Study*) afin de tester les items ainsi que les procédures de passation.
3. Direction de l'évaluation, de la prospective et de la performance (DEPP), ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche.

Figure 1 Thermomètre d'effort de PISA

Quel effort avez-vous fourni pour répondre à ce test ?

Essayez de vous imaginer face à une situation de la vie réelle (à l'école ou dans un autre contexte) qui est très importante pour vous personnellement. Vous auriez envie de bien réussir, et pour cela, vous tentez de donner le meilleur de vous-même, en y consacrant le plus d'efforts possibles.

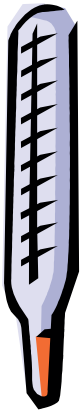
Dans cette situation, vous cocheriez la valeur la plus élevée sur le « thermomètre de l'effort », comme ci-dessous :	Par rapport à la situation que vous venez d'imaginer, quel effort pensez-vous avoir fourni en répondant à ce test ?	Si la note obtenue lors de ce test comptait pour votre bulletin scolaire, quel effort auriez-vous fourni ?
		
☒ 10	☐ 10	☐ 10
☐ 9	☐ 9	☐ 9
☐ 8	☐ 8	☐ 8
☐ 7	☐ 7	☐ 7
☐ 6	☐ 6	☐ 6
☐ 5	☐ 5	☐ 5
☐ 4	☐ 4	☐ 4
☐ 3	☐ 3	☐ 3
☐ 2	☐ 2	☐ 2
☐ 1	☐ 1	☐ 1

Figure 2 Instrument de mesure de la motivation au test (niveau collège)

Trois questions portant sur ce cahier

1. Sur une échelle de difficulté allant de 1 à 10, comment avez-vous trouvé les exercices de cette évaluation?

Très faciles — Très difficiles

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10

2. Comment vous êtes-vous appliqué(e) pour faire cette évaluation?

(indiquez votre degré d'application sur une échelle allant de 1 à 10)

Je ne me suis pas du tout appliqué(e) — Je me suis énormément appliqué(e)

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10

3. Si les résultats de cette évaluation comptaient pour votre bulletin scolaire, comment vous seriez-vous appliqué(e) ?

(indiquez votre degré d'application sur une échelle allant de 1 à 10)

Je ne me serais pas du tout appliqué(e) — Je me serais énormément appliqué(e)

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10

(question 1) et que ce lien est négatif. Plus les élèves ont jugé l'évaluation difficile, moins ils déclarent s'être appliqués pour la faire ▸ **Tableau 1**. Lorsque l'on étudie ce lien dans PISA, on note que le coefficient de corrélation est positif. Cela signifie que plus les exercices sont jugés difficiles par les élèves, plus ceux-ci déclarent avoir fourni d'effort pour y répondre. Les élèves qui disent avoir fourni beaucoup d'effort en répondant au test l'ont-ils fait car ils se sentaient très investis ou parce que la difficulté des exercices leur a demandé un effort considérable ? Le terme « effort » semble en effet ambigu pour informer sur la motivation des élèves face au test.

▸ **Tableau 1** **Relation entre la difficulté perçue du test et la motivation au test**

Évaluation	Coefficient de corrélation
Cedre histoire-géographie (3e)	**- 0,109**
Socle compétences 1 et 3 (3e)	**- 0,053**
Cedre histoire-géographie (CM2)	**- 0,178**
Cedre sciences (CM2)	**- 0,181**
Socle compétence 1 (CM2)	**- 0,324**
Socle compétence 2 (CM2)	**- 0,281**
Socle compétence 3 (CM2)	**- 0,213**
PISA (élèves de 15 ans)	**0,069**

Lecture : dans l'évaluation Cedre histoire-géographie niveau collège, la difficulté perçue du test est liée négativement à la motivation au test (coefficient de corrélation - 0,109) alors que dans PISA, où la motivation est mesurée en termes d'effort, ce lien est positif (0,069).
Note : les corrélations significatives au niveau 0,01 sont indiquées en gras.
Champs : France métropolitaine, public et privé sous contrat (Cedre) ; France métropolitaine + DOM, public et privé sous contrat (socle) ; France métropolitaine + DOM (sauf La Réunion), public et privé sous contrat (PISA).
Sources : MENESR-DEPP, évaluations Cedre et socle ; OCDE-PISA 2012.

La difficulté perçue des exercices entretient ainsi une relation avec la motivation à répondre au test, relation qui est par ailleurs plus forte au niveau primaire qu'au niveau secondaire. Ceci s'observe également au niveau agrégé : les évaluations jugées, en moyenne, les plus faciles, sont à la fois celles pour lesquelles le niveau de motivation des élèves s'avère le plus élevé. Dans la recherche des éléments d'explication, on peut noter le rôle du format d'exercices – les évaluations jugées les plus faciles et caractérisées par un degré plus grand de la motivation déclarée par les élèves sont celles qui se composent principalement des questions à choix multiples (QCM). Les informations qualitatives recueillies à l'aide des entretiens collectifs avec des élèves confirment cette observation : les élèves sont unanimes lorsqu'ils disent préférer les QCM aux questions à réponse construite ▸ **Encadré ci-contre**.

Profils de motivation

Butler et Adams [2007] proposent une analyse de l'investissement différentiel des élèves, et de l'effet de cet investissement sur la performance, en construisant un indicateur d'effort relatif à partir du « thermomètre d'effort » de PISA. Les auteurs

ÉTUDE QUALITATIVE DES CONDITIONS DE PASSATION

Une étude qualitative a été menée par la DEPP en mai 2013. Cette étude consistait à observer des conditions de passation et à conduire des entretiens collectifs (des *focus group*) avec des élèves dans une dizaine de collèges participant à une évaluation en sciences expérimentales [Bobineau, 2013]. L'objectif de l'étude était de connaître l'avis des élèves sur différents aspects de l'évaluation ainsi que d'obtenir des informations d'ordre qualitatif concernant leur investissement.

L'étude interrogeait aussi la validité de l'instrument de mesure de la motivation adapté à partir du « thermomètre d'effort ». Les renseignements provenant des entretiens collectifs avec des élèves montrent que ceux-ci ont bien compris les trois questions relatives au test qui leur étaient posées à la fin de l'évaluation (figure 2 p.121). La compréhension de l'instrument ne semble ainsi pas prêter à confusion. En outre, les élèves ont eu la possibilité de donner leur avis et de proposer des améliorations pour ces trois questions. De manière assez collégiale, ils proposent de réduire les échelles. Plus précisément, de les ramener à cinq possibilités de réponses, avec un milieu identifiable, deux extrêmes mais aussi la possibilité de mitiger leur réponse grâce aux entre-deux.

L'étude qualitative a mis en évidence le rôle que peut jouer le personnel de l'établissement dans la propension des élèves à participer à l'évaluation. Les entretiens conduits dans des établissements ayant pris le soin d'informer les élèves sur l'importance et l'utilité de l'évaluation ont révélé que les élèves de ces collèges comprenaient mieux pourquoi ils passaient ce test, sans conséquence directe pour eux.

classent les élèves en fonction de l'écart entre l'effort que ceux-ci déclarent avoir fourni en répondant au test PISA et l'effort qu'ils auraient fourni si les résultats du test avaient compté pour leur bulletin scolaire.

Cet indicateur très intéressant se prête aussi à quelques critiques. Notamment, il ne prend pas en compte le niveau général d'investissement des élèves. À titre d'exemple, un élève qui déclare avoir fourni un effort de 5 en répondant au test PISA et qui dit qu'il aurait fourni un effort de 7 si la note avait compté pour son bulletin a le même score sur l'échelle d'effort relatif qu'un élève qui coche respectivement 8 et 10. Ces deux élèves ont un écart de deux points entre les deux échelles et auront un score de l'effort relatif égal à 8 sur 10 (le score 10 représentant un même effort pour PISA que pour un test noté). Or, on peut supposer que pour un même effort relatif, deux élèves peuvent indiquer des niveaux de motivation générale différents, pouvant aboutir à des performances différentes au test.

Nous avons voulu vérifier cette hypothèse en positionnant les élèves simultanément sur les deux échelles – investissement au test et investissement si la note au test avait compté – et en étudiant leur score moyen au test en fonction de cette position. Observons les deux figures suivantes qui représentent les données des évaluations Cedre [4] histoire-géographie 2012 et PISA 2006 ▸ **Figures 3 et 4 p. 124-125**. Notons tout d'abord que la majorité des élèves se situent sur des niveaux élevés des échelles (7 points ou plus) et le plus souvent sous la ligne bissectrice, ce qui signifie qu'ils ont déclaré un investissement moins important pour un test standardisé que pour une épreuve notée. Le plus grand nombre a coché 10 sur l'échelle « si noté » et 8 sur l'échelle « test » (490

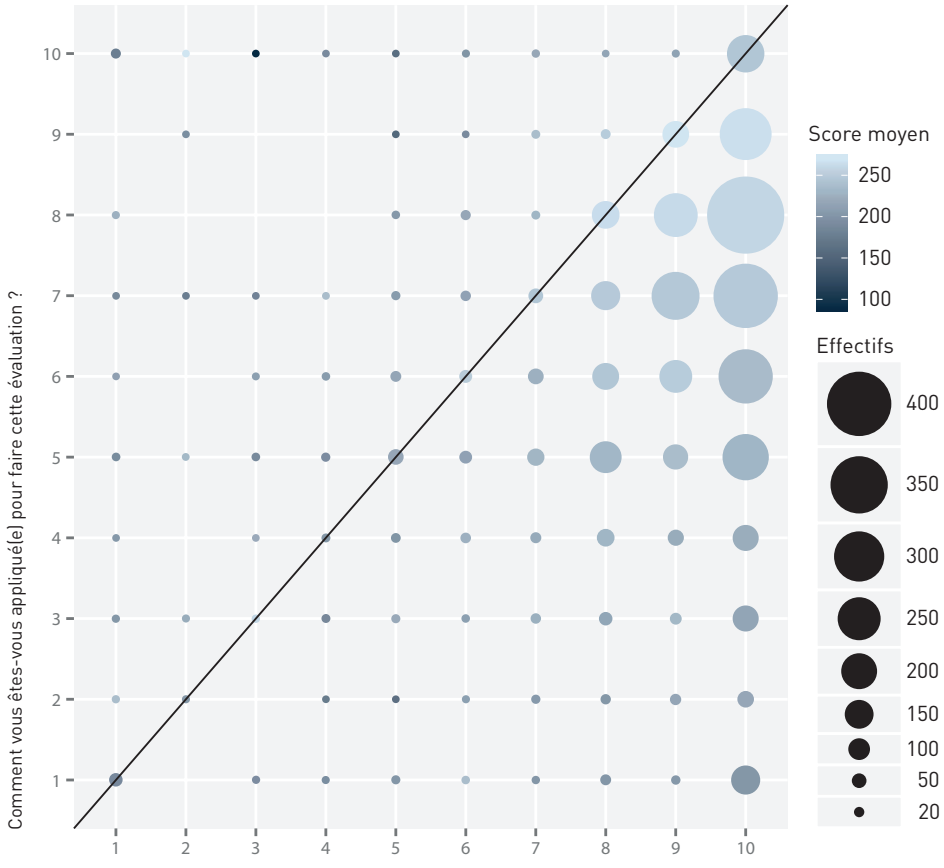
4. L'évaluation en histoire-géographie fait partie du cycle des évaluations disciplinaires réalisées sur échantillons (Cedre) que la DEPP a commencé à mettre en place en 2003 afin de rendre compte des résultats du système éducatif français au regard des objectifs fixés par les programmes [Trosseille et Rocher, ce numéro, p. 15].

élèves dans Cedre, soit 10 % des répondants, et 576 élèves dans PISA, soit 12 % des répondants).

On observe des variations de score considérables entre les élèves manifestant le même investissement relatif. Reprenons l'exemple évoqué plus haut et considérons deux groupes d'élèves qui ont coché respectivement sur les deux échelles 5 et 7 pour les uns, et 8 et 10 pour les autres. Si le score moyen en Cedre histoire-géographie est de 234 pour le premier groupe (n = 68), il s'élève à 260 pour le deuxième (n = 490), ce qui correspond à une différence d'environ un demi écart-type. Les données provenant de PISA montrent la même tendance : ces deux groupes d'élèves ont obtenu respectivement un score moyen de 490 (n = 46) et de 517 (n = 576) en 2006, soit une différence d'environ un tiers d'écart-type.

Nous nous concentrerons désormais sur l'enseignement secondaire où la motivation

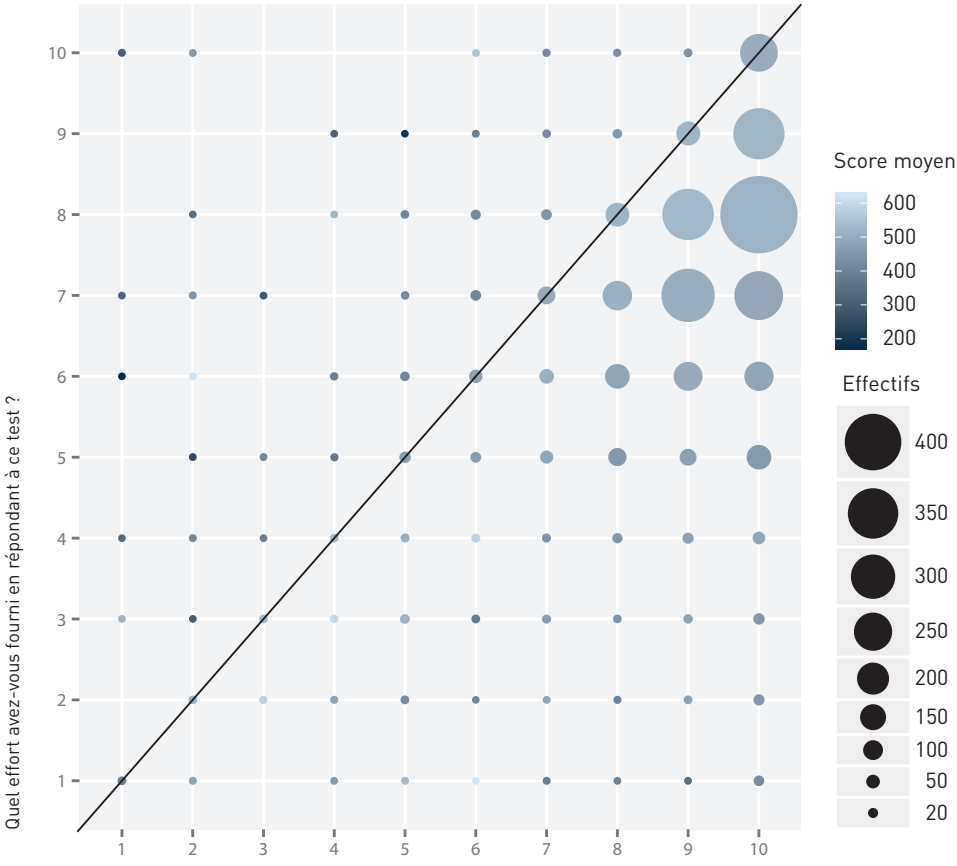
Figure 3 Score moyen à l'évaluation Cedre histoire-géographie selon le niveau d'investissement (en termes « d'application »)



Lecture : 490 élèves ont coché 8 sur l'échelle « d'application » au test et 10 sur l'échelle « d'application » au test si les résultats avaient compté pour leur bulletin scolaire. Le score moyen en histoire-géographie de ce groupe d'élèves est de 260 points.
Champ : France métropolitaine, public et privé sous contrat.
Source : MENESR-DEPP, évaluation Cedre histoire-géographie 2012.

des élèves semble davantage poser problème qu'à l'école. En nous inspirant du travail de Butler et Adams [*op. cit.*], nous avons regroupé les élèves en fonction de leur position sur ces deux échelles. Des regroupements ont été testés en vue de réunir des élèves ayant des niveaux de motivation et des performances similaires mais aussi pour avoir des effectifs suffisamment importants pour pouvoir effectuer des analyses secondaires en croisant ces groupes avec d'autres variables. Le regroupement retenu vise à prendre en compte le niveau général des élèves et à assembler ceux qui se ressemblent à l'égard de l'interaction entre la motivation et le score.
Nous avons réparti les élèves en sept « profils de motivation » ▸ **Figure 5**. Comme Butler et Adams, nous nommons « irréalistes » les élèves qui déclarent s'être davantage appliqués pour faire l'évaluation standardisée qu'ils ne l'auraient fait si la note au test avait compté pour leur bulletin scolaire (donc tous ceux qui se trouvent au-dessus de la

▸ Figure 4 Score moyen à l'évaluation PISA selon le niveau d'investissement (en termes « d'effort »)



Lecture : 576 élèves ont coché 8 sur l'échelle d'effort fourni au test et 10 sur l'échelle d'effort fourni si la note au test avait compté pour leur bulletin scolaire. Le score moyen en culture scientifique de ce groupe d'élèves est de 517 points.
Champ : France métropolitaine + DOM (sauf La Réunion), public et privé sous contrat.
Source : OCDE-PISA 2006.

diagonale exprimant un degré de motivation égal pour le test et pour une épreuve notée). Nous appelons « démotivés » ceux qui se considèrent aussi peu motivés pour le test standardisé que pour une épreuve notée (scores de 1 à 7 sur les deux échelles). Les « pragmatiques » sont les élèves qui se disent très investis dans une évaluation lorsque la note obtenue compte (scores de 8 à 10) mais très peu motivés pour un test à faibles enjeux (scores de 1 à 5). Les « peu motivés » sont ceux qui déclarent s'appliquer pour faire une évaluation qui compte (scores de 8 à 10) mais qui le font moins si les résultats n'ont pas de conséquence directe pour eux (scores 6 et 7). Enfin, les élèves motivés de manière générale sont appelés « réalistes » lorsque la différence entre les deux échelles est de 2 points, « assidus » si cet écart est égal à un point et « partisans » dans le cas où il n'y a pas de différence.

La présente étude se concentre sur la motivation des élèves face à des évaluations mais il serait très intéressant de comparer notre classification à la théorie de l'autodétermination et aux échelles de motivation développées antérieurement [Deci et Ryan, 1985]. Ainsi, on pourrait qualifier de « pragmatiques » les élèves qui sont motivés de manière extrinsèque, et notamment ceux en régulation externe. De même, les « démotivés » font fortement penser à l'amotivation. De telles comparaisons pourraient faire l'objet d'études ultérieures où les deux types d'échelles seraient proposés aux élèves.
La répartition de ces profils varie selon l'évaluation. Si les « pragmatiques » sont les plus représentés dans Cedre histoire-géographie (HG) ainsi que dans la session 5 d'EIST (en troisième), et les « peu motivés » en session 4 d'EIST (quatrième)[5], ce sont les « assidus » qui prévalent dans l'évaluation du socle

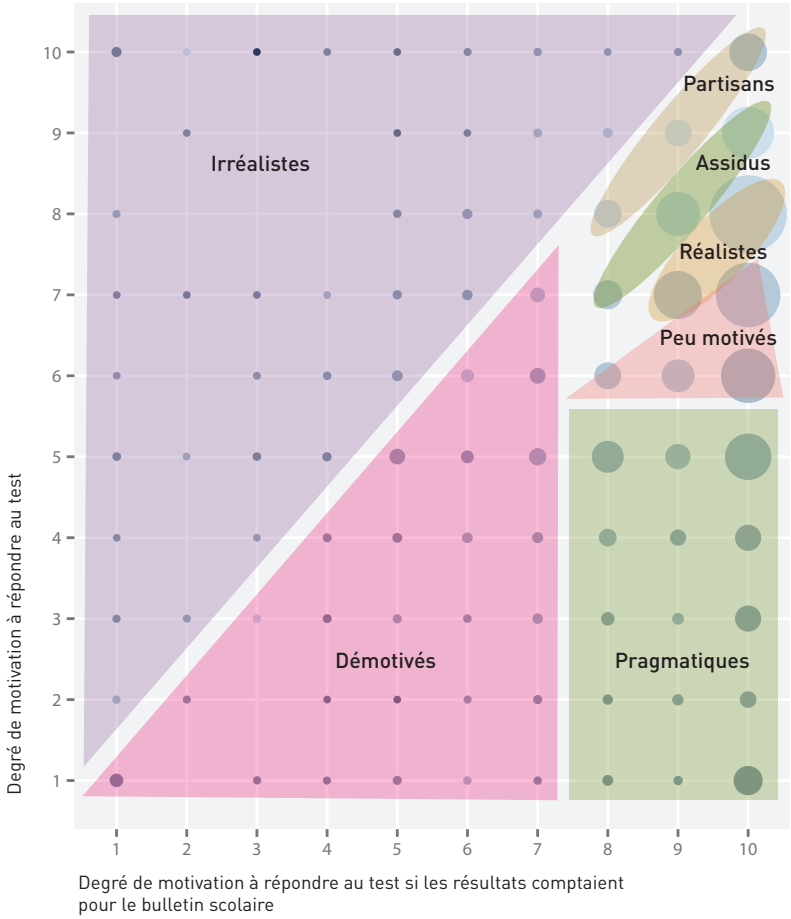
Tableau 2 Répartition des profils de motivation selon l'évaluation (en %)

	Cedre HG	Socle	PISA	EIST session 4	EIST session 5
Irréalistes	5	4	5	6	5
Démotivés	10	9	10	9	8
Pragmatiques	26	16	12	20	27
Peu motivés	20	14	19	23	25
Réalistes	15	18	22	19	16
Assidus	14	23	21	14	12
Partisans	10	15	12	9	7
Non-réponse	12	4	8	21	12

Lecture : dans l'évaluation Cedre histoire-géographie (HG) la part des élèves « démotivés » est de 10 %.
Note : la somme est égale à 100 % des répondants.
Champs : France métropolitaine, public et privé sous contrat (Cedre) ; France métropolitaine + DOM, public et privé sous contrat (Socle) ; France métropolitaine + DOM (sauf La Réunion), public et privé sous contrat (PISA) ; France métropolitaine, public (EIST).
Sources : MENESR-DEPP, évaluations Cedre, socle et EIST ; OCDE-PISA 2006.

5. Évaluation du dispositif d'enseignement intégré de sciences et technologie (EIST). Il s'agit d'une étude longitudinale : les élèves profitant du dispositif et les élèves témoins ont été évalués à cinq reprises entre novembre 2008 et mai 2012 [Le Cam et Cosnefroy, ce numéro, p. 283].

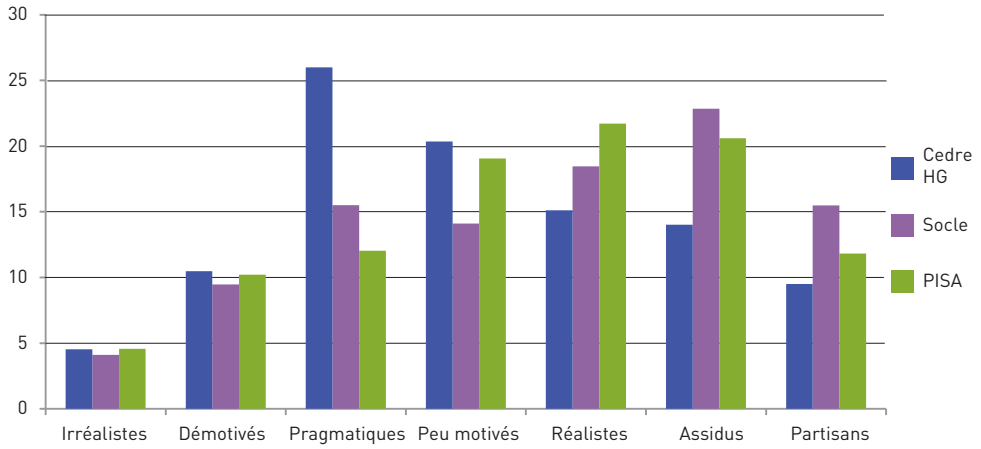
Figure 5 **Profils de motivation**



commun [6] et les « réalistes » dans PISA **Tableau 2**. Les différences entre les évaluations peuvent s'expliquer, en partie, par le terme utilisé dans l'instrument de mesure de motivation. Les évaluations Cedre histoire-géographie et Socle emploient le terme « application », les autres tests mesurant l'investissement des élèves à l'aide du « thermomètre d'effort ». Une représentation graphique de trois évaluations – Cedre histoire-géographie, socle et PISA – donne une vision plus marquée de ces différences **Figure 6.** Observons de plus près ces profils en fonction de quelques caractéristiques d'élèves et d'établissements. Pour ne pas surcharger les tableaux, nous ne considérerons que les évaluations Cedre histoire-géographie, socle et PISA qui constituent de bons exemples selon plusieurs aspects. Cedre est une évaluation disciplinaire destinée à mesurer l'atteinte des objectifs fixés par des programmes scolaires officiels. Les tests du socle visent à évaluer les proportions d'élèves ayant acquis les compétences 1 (la maîtrise de la langue française) et 3 (les principaux éléments de mathématiques et la culture scientifique et technologique)

6. Les évaluations du socle commun des connaissances et des compétences [Miconnet et Vourc'h, ce numéro, p. 141].

Figure 6 Répartition des profils de motivation dans Cedre HG, socle et PISA



Lecture : dans l'évaluation Cedre histoire-géographie la part des élèves « assidus » est de 14 %, contre 21 % dans PISA.
Champs : France métropolitaine, public et privé sous contrat (Cedre) ; France métropolitaine + DOM, public et privé sous contrat (Socle) ; France métropolitaine + DOM (sauf La Réunion), public et privé sous contrat (PISA).
Sources : MENESR-DEPP, évaluations Cedre et socle ; OCDE-PISA 2006.

qui font partie des sept grandes compétences constituant le bagage minimum censé être acquis par tous les élèves à la fin de la scolarité obligatoire. L'évaluation PISA s'intéresse à une génération d'élèves (élèves de 15 ans) et les évalue non sur des connaissances au sens strict mais sur leurs capacités à mobiliser et à appliquer celles-ci dans des situations variées, parfois éloignées de celles rencontrées dans le cadre scolaire. Ces trois évaluations se distinguent également par le format de questions, les tests du socle ne comportant que des QCM alors que Cedre et PISA demandent souvent aux élèves de construire leur réponse. Au-delà de ces différences, ce sont les variations dans l'instrument de mesure qui nous intéressent ici, PISA comportant le « thermomètre d'effort » et les deux autres évaluations le nouvel instrument modifié.
La répartition des profils de motivation varie selon le secteur de scolarisation. Ainsi, les « démotivés » et les « pragmatiques » sont proportionnellement plus nombreux dans des établissements publics – et surtout dans ceux de l'éducation prioritaire (EP) – que dans des collèges privés ▸ **Tableau 3**.

L'analyse des profils en fonction du sexe met en évidence une moindre motivation de la part des garçons : les « démotivés » et les « pragmatiques » sont proportionnellement plus nombreux parmi eux que parmi les filles ▸ **Tableau 4**. Les filles semblent ainsi davantage motivées non seulement lors des évaluations à faibles enjeux, mais aussi lorsque la note obtenue compte pour leur bulletin scolaire. Le retard scolaire a également un impact sur la motivation, les « démotivés » étant plus nombreux parmi les élèves ayant redoublé que parmi ceux « à l'heure » ▸ **Tableau 5**. On note par ailleurs que les « irréalistes » – les élèves qui déclarent s'être plus appliqués (ou avoir fait plus d'effort) en faisant le test que cela aurait été le cas si la note obtenue au test avait compté pour leur bulletin scolaire – sont également plus représentés parmi les élèves redoublants. En revanche,

Tableau 3 Répartition des profils de motivation selon le secteur de scolarisation (en %)

Profils de motivation	Cedre HG			Socle		
	Public hors EP	EP	Privé	Public hors EP	EP	Privé
Irréalistes	4	6	4	4	6	2
Démotivés	11	12	10	10	13	7
Pragmatiques	27	29	21	16	18	12
Peu motivés	19	22	22	14	15	15
Réalistes	15	12	16	18	17	21
Assidus	14	12	16	23	19	26
Partisans	10	8	10	16	13	17

Lecture : selon les résultats de l'évaluation Cedre histoire-géographie (HG), la part des élèves « démotivés » est de 12 % dans les établissements de l'éducation prioritaire (EP).
Champs : France métropolitaine, public et privé sous contrat (Cedre) ; France métropolitaine + DOM, public et privé sous contrat (socle).
Sources : MENESR-DEPP, évaluations Cedre et socle.

Tableau 4 Répartition des profils de motivation selon le genre (en %)

Profils de motivation	Cedre HG		Socle		PISA	
	Garçons	Filles	Garçons	Filles	Garçons	Filles
Irréalistes	5	4	5	3	5	4
Démotivés	12	9	12	7	12	9
Pragmatiques	28	24	20	11	16	9
Peu motivés	19	22	16	13	19	19
Réalistes	14	16	16	21	19	24
Assidus	12	16	17	29	17	24
Partisans	11	9	14	17	11	12

Sources : MENESR-DEPP, évaluations Cedre et socle ; OCDE-PISA 2006.

Lecture : selon les résultats de l'évaluation Cedre histoire-géographie (HG), la part des élèves « démotivés » est de 12 % parmi les garçons.
Champs : France métropolitaine, public et privé sous contrat (Cedre) ; France métropolitaine + DOM, public et privé sous contrat (socle) ; France métropolitaine + DOM (sauf La Réunion), public et privé sous contrat (PISA).

on observe relativement moins d'écarts selon le retard scolaire pour les profils « pragmatiques » et « peu motivés » que pour les « démotivés », ce qui amène à supposer que les élèves en retard se sentent moins motivés de manière générale (même dans le cas des épreuves scolaires notées), et pas seulement lors des tests à faibles enjeux. Lorsque l'on s'intéresse au statut socio-économique des élèves[7] selon leur profil de motivation, on observe que ce sont souvent les « irréalistes » qui proviennent des milieux socio-économiques les moins favorisés ▸ Tableau 6 p. 130. Si, dans le cas de l'évaluation Cedre histoire-géographie, l'indice moyen reflétant le statut socio-économique tend à être d'autant plus élevé que l'élève se manifeste motivé, la

7. Mesuré en termes de l'indice de position sociale (IPS) défini par Le Donné et Rocher [2010] dans Cedre et en termes de l'indice socio-économique international de statut professionnel (HISEI) dans PISA (OCDE, 2009].

Tableau 5 Répartition des profils de motivation selon le retard scolaire (en %)

Profils de motivation	Cedre HG		Socle		PISA	
	À l'heure	En retard	À l'heure	En retard	À l'heure	En retard
Irréalistes	4	7	3	7	2	8
Démotivés	9	16	7	16	9	12
Pragmatiques	25	31	14	18	11	13
Peu motivés	21	18	14	13	19	19
Réalistes	16	11	20	14	23	19
Assidus	15	9	25	17	23	16
Partisans	10	8	16	14	11	13

Lecture : selon les résultats de l'évaluation Cedre histoire-géographie (HG), la part des élèves « démotivés » est de 16 % parmi les élèves en retard dans leurs cursus scolaire.
Champs : France métropolitaine, public et privé sous contrat (Cedre) ; France métropolitaine + DOM, public et privé sous contrat (socle) ; France métropolitaine + DOM (sauf La Réunion), public et privé sous contrat (PISA).

Sources : MENESR-DEPP, évaluations Cedre et socle ; OCDE-PISA 2006.

relation est moins nette pour l'évaluation PISA. Dans PISA, contrairement à Cedre, les « démotivés » et les « pragmatiques » n'ont pas un statut socio-économique moins élevé. Rappelons encore une fois que l'instrument de mesure de motivation n'est pas le même pour les deux évaluations, le terme « application » étant employé dans Cedre et celui « d'effort » dans PISA. On sait que le score des élèves est lié à leur statut socio-économique, les élèves provenant des milieux sociaux moins favorisés obtenant des résultats moins bons [OCDE, 2007]. Il se peut ainsi que les élèves qui déclarent avoir fait peu d'effort pour faire le test – les « démotivés » et les « pragmatiques » – l'aient fait parce que leur niveau est bon et que les exercices leur ont paru faciles. Testons cette hypothèse en observant les scores moyens par profil de motivation.

De manière générale, plus les élèves se déclarent motivés (en termes d'application ou d'effort), plus leur score moyen est élevé ▸ Tableau 7. Néanmoins, ce sont les « assidus » – et non les « partisans » – qui ont le score moyen le plus élevé. Les données PISA montrent que les « démotivés » ont obtenu un score relativement

Tableau 6 Statut socio-économique des élèves selon leur profil de motivation

Profils de motivation	Cedre HG (IPS)		PISA (HISEI)	
	Moyenne	Écart-type	Moyenne	Écart-type
Irréalistes	- 0,21	0,62	50	23
Démotivés	- 0,23	0,70	52	21
Pragmatiques	- 0,14	0,68	53	21
Peu motivés	- 0,06	0,68	51	19
Réalistes	- 0,003	0,66	51	18
Assidus	- 0,02	0,68	52	18
Partisans	- 0,03	0,69	51	18

Lecture : selon les résultats de l'évaluation Cedre histoire-géographie (HG), l'indice IPS moyen des élèves « démotivés » est de - 0,23.
Note : IPS – indice de position sociale défini par Le Donné et Rocher [2010] ; HISEI – indice socio-économique international de statut professionnel [OCDE, 2009].
Champs : France métropolitaine, public et privé sous contrat (Cedre) ; France métropolitaine + DOM (sauf La Réunion), public et privé sous contrat (PISA).

Sources : MENESR-DEPP, évaluations Cedre ; OCDE-PISA 2006.

élevé. Il s'agit des élèves qui déclarent avoir fourni relativement peu d'effort lors du test et qui disent à la fois que l'effort fourni n'aurait pas été considérablement plus élevé si la note obtenue au test avait compté pour leur bulletin scolaire. Notre hypothèse semble se confirmer : on peut supposer qu'un effort plutôt faible, associé à des scores relativement élevés indique que ces élèves n'ont pas besoin de fournir beaucoup d'efforts car les exercices du test (et les épreuves scolaires notées) sont faciles pour eux.

On peut ainsi constater que la motivation face aux évaluations à faibles enjeux varie selon les caractéristiques d'établissements et d'élèves. Ceci veut dire que lorsque la motivation impacte de manière significative les résultats qu'obtiennent les élèves lors de tels tests, un moindre investissement de certaines sous-populations d'élèves peut introduire des biais dans l'estimation de leur performance et amener à de mauvaises interprétations des résultats. Nous étudions plus en détail la relation entre la motivation et les résultats d'élèves.

Tableau 7 Score moyen des élèves selon leur profil de motivation

Profils de motivation	Cedre HG		Socle compétence 1 (français)		Socle compétence 3 (math. et sciences)		PISA (sciences)	
	Moyenne	Écart-type	Moyenne	Écart-type	Moyenne	Écart-type	Moyenne	Écart-type
Irréalistes	213	42	- 0,74	0,95	- 0,78	1,05	425	100
Démotivés	223	52	- 0,54	1,16	- 0,44	1,37	493	116
Pragmatiques	225	42	- 0,33	1,04	- 0,32	1,16	466	99
Peu motivés	246	44	0,10	0,97	0,01	1,00	495	93
Réalistes	256	45	0,32	1,07	0,25	1,07	514	87
Assidus	261	49	0,49	1,05	0,46	1,11	527	90
Partisans	259	55	0,46	1,26	0,46	1,23	517	95

Lecture : selon les résultats de l'évaluation Cedre histoire-géographie (HG), le score moyen des élèves « démotivés » est de 223.
Champs : France métropolitaine, public et privé sous contrat (Cedre) ; France métropolitaine + DOM, public et privé sous contrat (socle) ; France métropolitaine + DOM (sauf La Réunion), public et privé sous contrat (PISA).
Sources : MENESR-DEPP, évaluations Cedre et socle ; OCDE-PISA 2006.

MOTIVATION ET PERFORMANCE

L'évolution du score en fonction de l'évolution de la motivation

Les données longitudinales de l'évaluation EIST nous permettent d'étudier l'évolution des résultats d'élèves d'une session à l'autre. Avec deux sessions d'évaluation (sessions 4 et 5) comportant l'instrument de mesure de l'investissement face au test – en termes d'effort –, nous sommes en mesure de mettre en lien l'évolution du score des élèves avec l'évolution de leur investissement afin d'observer si une augmentation du degré d'investissement s'accompagne d'une évolution positive du score.

En outre, un questionnaire de contexte était adressé aux élèves dans le cadre de cette évaluation. Une partie des questions portait sur leur motivation et leur intérêt vis-à-vis de la discipline évaluée, en l'occurrence les sciences. Sur une échelle à quatre positions, les élèves étaient incités à exprimer leur degré d'accord avec les affirmations suivantes : *« Ce que je fais en science est intéressant », « Je participe en science parce que j'aime bien chercher », « Je travaille en science parce que j'aime bien cette discipline », « Je participe en science parce que j'aime bien faire des expériences »*. Nous avons construit un indicateur synthétisant ces questions à l'aide d'une analyse en composantes principales[8]. Avec cet indicateur de l'intérêt/motivation à l'égard des sciences à notre disposition, nous avons la possibilité de faire la distinction entre les aspects de la motivation spécifiques à la passation du test (*situation-specific motivation*) et les aspects spécifiques au domaine évalué (*domain-specific motivation*).

On peut constater une tendance positive en ce qui concerne l'évolution du score entre la session 4 et la session 5. En revanche, la motivation spécifique à la situation (l'effort fourni au test) ainsi que la motivation spécifique au domaine tendent à diminuer d'une session d'évaluation à l'autre. Lorsque l'on étudie le lien entre ces évolutions, on note qu'elles sont toutes positivement (et significativement) corrélées, mais que cette corrélation n'est pas très élevée ▸ **Tableau 8**. Ainsi, une augmentation de motivation – spécifique à la situation ou au domaine – s'accompagne d'une évolution positive du score. On observe aussi que l'évolution de l'effort fourni au test est un peu plus corrélée à l'évolution du score que l'évolution de l'intérêt vis-à-vis des sciences. Nous avons ensuite construit un modèle de régression dans le but « d'expliquer » le score obtenu lors de la session 5 en fonction de l'évolution de la motivation entre les sessions 4 et 5 ▸ **Tableau 9**. Comme l'indiquent les coefficients de régression, le score obtenu lors de la session 5 est d'autant plus élevé que l'évolution de l'effort fourni au test ainsi que de l'intérêt vis-à-vis des sciences a été positive. En revanche, le pouvoir explicatif du modèle s'avère très faible : l'évolution de ces deux indicateurs de motivation n'explique que 1 % de la variabilité du score obtenu lors de la session 5. Lorsque l'on ajoute au modèle le niveau initial de l'élève – le score obtenu lors de la

▸ **Tableau 8** **Relations entre les évolutions du score, de l'effort fourni au test et de l'intérêt vis-à-vis des sciences entre la session 4 et la session 5 (coefficients de corrélation)**

	Évolution du score	Évolution de l'effort fourni au test	Évolution de l'indicateur d'intérêt vis-à-vis des sciences
Évolution du score	**1**	**0,148**	**0,087**
Évolution de l'effort fourni au test	**0,148**	**1**	**0,154**
Évolution de l'indicateur d'intérêt vis-à-vis des sciences	**0,087**	**0,154**	**1**

Lecture : l'évolution du score entre les sessions 4 et 5 est liée positivement à l'évolution de l'effort fourni au test (coefficient de corrélation 0,148).
Note : les corrélations significatives au niveau 0,01 sont indiquées en gras.
Champ : France métropolitaine, public.
Sources : MENESR-DEPP, évaluations du dispositif EIST.

8. Nous avons effectué une analyse en composantes principales sur les données de la session 4 et nous avons utilisé ces résultats pour calculer les scores factoriels sur les données de la session 5. Seuls les individus participant aux deux sessions ont été retenus pour les analyses présentées ici.

session 4 – les autres coefficients restent significatifs et le coefficient de détermination (R^2) passe à 50 %. Cela nous permet de conclure que c'est surtout le niveau initial de l'élève qui explique son gain en performance entre les deux sessions d'évaluation, l'évolution des aspects motivationnels jouant un rôle considérablement moins important.

Tableau 9 Score obtenu lors de la session 5 en fonction de l'évolution des aspects motivationnels (coefficients de régression)

	Modèle 1	Modèle 2
Constante	**0,06**	- 0,01
Évolution de l'effort fourni au test	**0,08**	**0,11**
Évolution de l'indicateur d'intérêt vis-à-vis des sciences	**0,08**	**0,05**
Score obtenu lors de la session 4		**0,69**
R^2	0,01	0,5

Lecture : l'augmentation d'un écart-type du score obtenu lors de la session 4 s'accompagne d'une augmentation du score lors de la session 5 de 0,69 écart-type, l'évolution de l'effort fourni au test et l'évolution de l'indicateur d'intérêt vis-à-vis des sciences étant maintenues constantes.
Note : les coefficients significatifs au niveau 0,01 sont indiqués en gras. R^2 : coefficient de détermination.
Champ : France métropolitaine, public.
Sources : MENESR-DEPP, évaluations du dispositif EIST.

Variabilité inter-classes et intra-classe de la performance en fonction de la motivation

Afin de mieux distinguer l'effet de divers aspects motivationnels et des caractéristiques individuelles et d'établissements sur la variation du score, nous avons employé la méthode de l'analyse multiniveaux (encadré p. 123). Cette méthode permet d'étudier la variabilité, entre établissements et au sein de chaque établissement, de la performance des élèves en fonction de la motivation à répondre au test. Elle permet également de déterminer dans quelle mesure la relation entre la motivation et la performance varie d'un établissement à l'autre. Les analyses présentées ici sont effectuées sur les données de l'évaluation Cedre histoire-géographie. Il s'agit d'un échantillon de classes mais une seule classe par établissement a été tirée au sort, ce qui signifie que ces deux unités sont ici confondues.

Tout d'abord, nous nous sommes intéressés à la décomposition de la variabilité du score en histoire-géographie selon les deux niveaux étudiés (niveau élève et niveau classe). Nous avons construit un modèle « vide » pour connaître la part du score attribuable aux différences entre les établissements et la part attribuable aux différences entre les élèves, au sein d'un même établissement **(modèle 1)** Tableau 10. En calculant le coefficient de corrélation intra-classe ρ pour ce modèle (encadré p. 137), nous obtenons une valeur de 0,22, ce qui indique que 22 % de la variabilité du score sont dus aux différences entre les classes. Cela signifie que la performance varie plus à l'intérieur des classes qu'entre les classes.
Nous avons ensuite intégré la motivation face au test dans le modèle vide pour observer si cet ajout amène à une réduction de la variance du score entre classes **(modèle 2)**.

Tableau 10 Modèles multiniveaux

PARAMÈTRES	MODÈLE 1	MODÈLE 2	MODÈLE 3
	« Vide »	Motivation au test	Pente par établissement
Effets fixes			
Constante	- 0,034 (0,037)	- 0,031 (0,034)	- 0,04 (0,035)
Motivation au test		0,119*** (0,007)	0,119*** (0,008)
Motivation au test si les résultats comptaient pour le bulletin scolaire			
Difficulté perçue du test			
Intérêt vis-à-vis de l'histoire-géographie			
Garçon			
Retard scolaire			
Collège en éducation prioritaire			
Collège privé			
Motivation au test, moyenne par classe			
Effets aléatoires			
Niveau 2 : variance des constantes (variance inter-classes)	0,224*** (0,028)	0,188*** (0,024)	0,192*** (0,024)
Covariance constantes-pentes			0,016*** (0,004)
Variance des pentes			0,003*** (0,001)
Niveau 1 : variance inter-élèves	0,792*** (0,019)	0,734*** (0,017)	0,72*** (0,017)
AIC	10209,3	9906,3	9879,8
BIC	10219,1	9919,3	9899,4

Lecture : toutes les autres variables comprises dans le modèle 8 étant tenues constantes, un garçon obtient un score de 0,11 écart-type plus élevé qu'une fille.
Note : les indicateurs concernant la motivation à répondre au test et la difficulté perçue du test sont centrés par rapport à la moyenne générale. Les unités de ces variables sont exprimées en points. Le score en histoire-géographie ainsi que l'indicateur de l'intérêt vis-à-vis de l'histoire-géographie sont standardisés et ont l'écart-type comme unité de mesure. Les autres variables explicatives n'ont pas d'unité de mesure mais une catégorie de référence (les garçons par rapport aux filles, les élèves en retard par rapport aux élèves « à l'heure », etc.). Les erreurs standards associées aux effets sont présentées entre parenthèses.
*** désigne un effet significatif au niveau 0,01 ; ** désigne un effet significatif au niveau 0,05.
AIC : critère d'information d'Akaike. **BIC :** critère d'information bayésien.
Champ : France métropolitaine, public et privé sous contrat.
Source : MENESR-DEPP, évaluation cedre HG.

MODÈLE 4	MODÈLE 5	MODÈLE 6	MODÈLE 7	MODÈLE 8
Motivation si impact sur bulletin scolaire	Difficulté perçue du test	Intérêt vis-à-vis de l'histoire-géographie	Caractéristiques sociodémographiques et scolaires	Caractéristiques des établissements
- 0,039 (0,034)	- 0,035 (0,032)	- 0,032 (0,031)	0,022 (0,032)	0,011 (0,034)
0,108*** (0,008)	0,099*** (0,008)	0,08*** (0,008)	0,077*** (0,007)	0,07*** (0,007)
0,055*** (0,008)	0,059*** (0,008)	0,047*** (0,008)	0,037*** (0,008)	0,037*** (0,008)
	- 0,096*** (0,007)	- 0,079*** (0,007)	- 0,074*** (0,007)	-0,074*** (0,007)
		0,205*** (0,015)	0,196*** (0,014)	0,195*** (0,014)
			0,115*** (0,026)	0,11*** (0,026)
			- 0,527*** (0,033)	- 0,526*** (0,033)
				- 0,144** (0,07)
				0,227*** (0,064)
				0,138*** (0,031)
0,186*** (0,023)	0,162*** (0,021)	0,143*** (0,019)	0,125*** (0,017)	0,093*** (0,013)
0,016*** (0,004)	0,014*** (0,003)	0,012*** (0,003)	0,011*** (0,003)	0,01*** (0,003)
0,003*** (0,001)	0,003*** (0,001)	0,003*** (0,001)	0,002*** (0,001)	0,002*** (0,001)
0,712*** (0,017)	0,679*** (0,016)	0,646*** (0,016)	0,607*** (0,015)	0,607*** (0,015)
9839,9	9650,5	9460,6	9208,6	9168,4
9862,7	9676,5	9489,9	9244,4	9214

Autrement dit, les différences dans le degré de motivation expliquent-elles une partie de la variation du score d'un établissement à l'autre ? L'ajout de la motivation au test dans le modèle conduit à une réduction de la variabilité des scores, aux deux niveaux (classe et élève) : la variance inter-classes passe de 0,224 à 0,188 (soit une diminution de 16 %) et la variance inter-élèves passe de 0,792 à 0,734 (soit une diminution de 7 %). Les indices de sélection de modèle (AIC , BIC) diminuent aussi, ce qui indique que le pouvoir explicatif du modèle augmente. Le modèle 2 rend ainsi mieux compte de la structure des données que le modèle 1. On observe également que l'augmentation d'une unité sur l'échelle de la motivation s'accompagne d'un gain de score d'un dixième d'écart-type et que cet effet est significatif.

Dans l'étape suivante, on permet au modèle d'ajuster une pente par établissement afin de tester si la relation entre le score et la motivation à répondre au test varie en fonction de l'établissement **(modèle 3)**. En effet, on observe une relation positive entre le score moyen de la classe et la pente entre le score et la motivation : plus le niveau de la classe est élevé, plus la corrélation entre la motivation et le score est forte. Comme les effets aléatoires se sont avérés significatifs, nous les avons gardés pour les modèles suivants.
Nous avons progressivement ajouté des effets fixes : la motivation de l'élève si la note au test avait compté pour son bulletin scolaire **(modèle 4)**, la difficulté perçue du test **(modèle 5)** et l'indicateur de l'intérêt vis-à-vis de l'histoire-géographie **(modèle 6)**. Le questionnaire de contexte adressé aux élèves à la fin de leur cahier d'évaluation comportait un certain nombre d'items relatifs à l'intérêt vis-à-vis de l'histoire-géographie. Sur une échelle à cinq positions, les élèves étaient incités à indiquer à quelle fréquence ils s'intéressent, en dehors du collège, à l'histoire, à la géographie, à l'éducation civique *« en lisant des livres ou des revues sur ces sujets », « en regardant des émissions de télévision ou des films », « en cherchant des documents sur Internet », « en visitant des musées, des lieux historiques, des sites géographiques »* (jamais, rarement, de temps en temps, souvent, très souvent). Le questionnaire les interrogeait également sur leurs sentiments vis-à-vis de la discipline : *« À propos de l'histoire, la géographie, l'éducation civique, vous diriez… »* (j'adore, j'aime bien, j'aime moyennement, je n'aime pas du tout, je déteste). Nous avons employé l'analyse en composantes principales afin de construire un indicateur synthétisant tous ces items.

Ensuite, nous avons intégré dans le modèle les caractéristiques sociodémographiques et scolaires des élèves telles que le sexe et le retard scolaire **(modèle 7)** et, enfin, le secteur et le degré moyen de motivation caractérisant les établissements **(modèle 8)**. On observe que les indices AIC et le BIC diminuent à chaque étape, ce qui met en évidence une augmentation progressive du pouvoir explicatif du modèle. Par rapport au modèle vide, le modèle complet **(modèle 8)** amène à une réduction de la variance inter-classes de 59 % (de 0,224 à 0,093) et de la variance inter-élèves de 23 % (de 0,792 à 0,607).

On note que l'augmentation d'une unité sur l'échelle de la motivation au test s'accompagne d'un gain de 0,07 écart-type de score, toutes choses égales par ailleurs (modèle 8). Une unité supplémentaire sur l'échelle de la motivation, si l'épreuve était notée, amène quant à elle à une augmentation de score de 0,04 écart-type. La difficulté perçue du test est liée négativement au score : un point supplémentaire sur cette échelle entraîne une réduction du score de 0,07 écart-type. L'augmentation, d'un écart-type, de l'indicateur de l'intérêt/motivation vis-à-vis-de l'histoire-géographie est associé à un gain de score de 0,2 écart-type. Toutes choses égales par ailleurs, un garçon obtient un score d'un dixième d'écart-type plus élevé qu'une fille et un redoublant un score d'un demi écart-type moins élevé qu'un élève « à l'heure ». Enfin, le score varie également en fonction des caractéristiques d'établissements : toutes choses égales par ailleurs, un élève scolarisé dans le secteur de l'éducation prioritaire obtient un score en histoire-géographie moins élevé (de 0,14 écart-type) et celui inscrit dans un collège privé un score plus élevé (de 0,23 écart-type) par rapport à celui suivant sa scolarité dans un établissement public hors éducation prioritaire. Outre le niveau individuel de motivation à répondre au test, le niveau moyen

MODÈLES MULTINIVEAUX

L'analyse multiniveaux, comme le nom l'indique, prend en compte la structure hiérarchique des données, en l'occurrence organisée sur deux niveaux : élève et établissement (ou bien la classe). L'emploi des modèles multiniveaux permet ainsi d'évaluer la variabilité des résultats à l'intérieur des établissements (entre les élèves) ainsi qu'entre les établissements.

Dans l'analyse multiniveaux, on peut distinguer les effets fixes du modèle, qui sont les coefficients de régression, et les effets aléatoires du modèle, c'est-à-dire les estimations de variance.

Le modèle « vide » consiste en une décomposition de variance :

$$y_{ij} = \beta_{0j} + \varepsilon_{ij}$$
$$\beta_{0j} = \gamma_{00} + u_{0j}$$

où y_{ij} est le score de l'élève *i* dans l'établissement *j*, β_{0j} est la constante de l'établissement *j*, ε_{ij} le résidu pour l'élève *i* dans l'établissement *j*, γ_{00} la constante générale et u_{0j} l'écart entre la constante générale et la constante de l'établissement *j* (parfois appelé « effet établissement »).

Le coefficient de corrélation intra-classe ρ montre la part de la variance totale du score qui peut être expliquée par les différences entre les établissements :

$$\rho = \frac{\sigma^2_{u0}}{\sigma^2_{u0} + \sigma^2_{\varepsilon 0}}$$

où σ^2_{u0} est la variance inter-établissement – soit Var(u_{0j}) – et $\sigma^2_{\varepsilon 0}$ est la variance inter-élèves, soit Var(ε_{ij}).

Avec l'introduction d'une variable explicative – par exemple, la motivation à répondre au test *X* – comme un effet fixe, l'équation peut être écrite :

$$y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \varepsilon_{ij}$$
$$\beta_{0j} = \gamma_{00} + u_{0j}$$
$$\beta_{1j} = \gamma_{10}$$

Ce modèle a deux composantes aléatoires, la variance de ε_{ij} et la variance de u_{0j}, et deux paramètres fixes, γ_{00} et γ_{10} (*i.e.* le coefficient de régression de la variable *X*).

Nous avons considéré, plus haut, que l'effet de la variable indépendante sur la variable dépendante est fixe, c'est-à-dire que la relation entre la motivation à répondre au test et la performance est la même dans chaque établissement. Or, les établissements peuvent se distinguer au regard de ce lien. L'analyse multiniveaux permet de prendre en compte ces différences en définissant l'effet de la variable explicative comme un effet aléatoire, autrement dit, en ajustant une pente de régression par établissement. L'équation avec pentes aléatoires peut s'écrire :

$$y_{ij} = \alpha_j + \beta_{1j} X_{ij} + \varepsilon_{ij}$$
$$\alpha_j = \gamma_{00} + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + u_{1j}$$

Dans ce modèle, le coefficient de régression de la variable *X* a deux composantes : un coefficient de régression général identique pour tous les établissements (la part fixe, noté γ_{10}) et un coefficient relatif à l'établissement correspondant (la partie aléatoire, noté u_{1j}).

Pour en savoir plus sur les modèles multiniveaux, le lecteur est invité à consulter l'ouvrage de Goldstein [2003].

de motivation par classe joue un rôle sur le score, les élèves provenant des classes plus motivées en moyenne ayant des scores plus élevés, toutes choses égales par ailleurs (l'augmentation d'une unité de cet indicateur par classe est associée à un gain de score de 0,14 écart-type)[11]. Ce dernier résultat est sans doute à relier aux effets des conditions de passation, dont on sait qu'elles peuvent varier selon les établissements ▸ **Encadré ci-dessus**. L'influence sur la motivation des élèves du personnel de l'établissement en charge de l'évaluation doit être davantage explorée lors de futures études.

11. Afin de vérifier si de tels résultats sont spécifiques à cette évaluation, nous avons testé le même modèle sur les données de l'évaluation du socle commun. Nous observons quelques variations mais le profil de régression est identique, nous retrouvons les mêmes tendances et les mêmes ordres de grandeur des coefficients.

CONCLUSIONS, PRÉCONISATIONS

L'analyse et la comparaison des données recueillies par des instruments de mesure de motivation – « thermomètre d'effort » ou « échelles d'application » – mettent en évidence des différences dans les résultats ainsi que dans les conclusions qui peuvent en être tirées. La motivation des élèves étant liée à la difficulté (perçue) du test, la nécessité d'ajouter une échelle pour mesurer cette difficulté se manifeste. Les échelles « d'application » semblent ainsi mieux adaptées pour rendre compte du degré de motivation des élèves français que le thermomètre « d'effort » utilisé dans PISA. En outre, comportant une charge de lecture considérablement moindre que ce dernier, l'instrument modifié est censé moins désavantager les lecteurs faibles.

La motivation face au test, comme nous l'avons vu, varie selon les caractéristiques d'élèves, les garçons et les redoublants se manifestant relativement moins impliqués. De même, la motivation semble moindre dans le secteur de l'éducation prioritaire, ce qui confirme, dans le contexte français, les tendances repérées dans une étude précédente [Jakwerth, Stancavage, Reed, 1999] selon laquelle la motivation posait problème dans des établissements caractérisés par le plus faible niveau des élèves.

La relation entre la motivation à répondre au test et la performance semble, selon notre étude, plutôt modeste. En outre, les différences en termes de degré de motivation ne sont pas considérables entre les établissements. Cependant, pour mieux explorer la question de motivation des élèves face à un test sans enjeux pour eux, nous avons mené une expérience à partir de l'évaluation Cedre mathématiques en mai 2014. Cette expérience – en cours d'analyse – vise à comparer les performances de deux groupes d'élèves : un groupe expérimental ayant au préalable reçu l'information que l'épreuve sera notée, et un groupe témoin qui passe le test dans des conditions habituelles.

BIBLIOGRAPHIE

Bobineau M., 2013, *Évaluations dites « à faibles enjeux » : quelle perception et implication de la part des élèves ? Étude qualitative à partir de CEDRE Sciences 2013*, rapport de stage, MENESR-DEPP.

Butler J., Adams R. J., 2007, "The Impact of Differential Investment of Student Effort on the Outcomes of International Studies", *Journal of Applied Measurement*, vol. 8, n° 3, p. 279-304.

Deci E. L., Ryan R. M., 1985, *Intrinsic motivation and self-determination in human behavior*, New York, Plenum.

Eklöf H., 2008, "Test-taking motivation on low-stakes tests: A Swedish TIMSS 2003 example", *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, vol. 1, p. 9-21.

Goldstein H., 2003, *Multilevel Statistical Models*, 3e éd., Londres, Edward Arnold.

Jakwerth P. M., Stancavage F. B., Reed E., 1999, *An Investigation of Why Students Do Not Respond to Questions,* NAEP Validity Studies (NVS) Panel, Palo Alto, American Institute for Research.

Keskpaik S., Rocher T., 2012, « Les évaluations à faibles enjeux : quel rôle joue la motivation ? Une expérience à partir de PISA », communication dans le cadre du 24e colloque de l'ADMEE-Europe, Luxembourg.

Le Donné N., Rocher T., 2010, « Une meilleure mesure du contexte socio-éducatif des élèves et des écoles. Construction d'un indice de position sociale à partir des professions des parents », *Éducation & formations*, n° 79, MENJVA-DEPP, p. 103-115.

Ndinga P., Frenette E., 2010, « Élaboration et validation de l'Échelle de motivation à bien réussir un test (EMRT) », *Mesure et évaluation en éducation*, vol. 33, n° 3, p. 99-123.

OCDE, 2009, *PISA 2006*, *Technical Report*, Paris, OCDE.

OCDE, 2009, *PISA Data Analysis Manual*, 2e éd., Paris, OCDE.

OCDE, 2007, *PISA 2006, Les compétences en sciences, un atout pour réussir, vol. 2 : Données*, Paris, OCDE.

O'Neil H. F., Abedi J., Lee C., Miyoshi J., Mastergeorge A., 2004, *Monetary Incentives for Low-Stakes Tests*, CSE Report 625, Los Angeles, CRESST/UCLA.

Penk C., Poehlmann C., Roppelt A., 2013, "Do Test-Takers Give Their Best? Motivational Determinants of Test Performance in Low-Stakes Assessments", article présenté au congrès annuel de l'American Educational Research Association, San Francisco.

DÉTERMINATION DES STANDARDS MINIMAUX POUR ÉVALUER LES COMPÉTENCES DU SOCLE COMMUN

Nicolas Miconnet
MENESR-DEPP, bureau des études statistiques sur les élèves

Ronan Vourc'h
MENESR-DEPP, bureau de l'évaluation des élèves

Depuis 2012, la direction de l'évaluation, de la prospective et de la performance (DEPP) du ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche est en charge de la production d'indicateurs relatifs à la maîtrise des compétences du socle commun. Pour ce faire, elle a progressivement mis en place des évaluations standardisées auprès d'échantillons représentatifs d'élèves en fin de CM2 et en fin de troisième. Celles-ci permettent de recueillir des informations fiables et comparables dans le temps, alors que celles obtenues à partir de l'attribution des compétences du socle commun par les enseignants peuvent varier en fonction des caractéristiques individuelles des élèves, mais aussi de facteurs liés à leur établissement.
La mise au point de tels indicateurs impose d'établir des scores seuils permettant de distinguer ceux qui atteignent la compétence évaluée et ceux qui ne l'atteignent pas. Pour cela, on a recours à des méthodes qui confrontent les résultats issus des évaluations standardisées avec le jugement d'enseignants et d'experts sur le niveau des élèves et le contenu des évaluations.
Parmi les méthodes utilisées, celle dite « des marque-pages » se révèle la mieux adaptée à ce contexte d'évaluation. Elle permet, à l'exception des langues vivantes étrangères au collège, d'aboutir à des pourcentages de validation qui varient d'environ 70 % à 80 % selon les niveaux scolaires et les disciplines. Ces pourcentages ainsi déterminés diffèrent selon le secteur de scolarisation, le sexe et l'âge. Enfin, l'analyse du devenir d'un échantillon d'élèves de troisième vient conforter la démarche mise en œuvre pour déterminer les seuils de maîtrise.

Le socle commun a été inscrit dans la loi en 2005. Celle-ci arrête que *« la scolarité obligatoire doit au moins garantir à chaque élève les moyens nécessaires à l'acquisition d'un socle commun constitué d'un ensemble de connaissances et de compétences qu'il est indispensable de maîtriser pour accomplir avec succès sa scolarité, poursuivre sa formation, construire son avenir personnel et professionnel et réussir sa vie en société*[1]*»*.
Le socle commun est aujourd'hui détaillé en sept compétences – elles-mêmes définies au travers de plusieurs domaines –, qui sont évaluées par les enseignants à trois étapes de la scolarité : le palier 1 en fin de CE1 (uniquement les compétences 1, 3 et 6) ; le palier 2 en fin de CM2 et le palier 3 en fin de troisième ▸ **Encadré p. 143**.
Pour ce faire, ils disposent d'un livret définissant les différents domaines et compétences à valider, ainsi que de grilles d'aide à la décision.
Le décret relatif au socle commun[2] paru en 2006 a instauré la règle de non-compensation des compétences, considérant notamment que la maîtrise du socle commun ne pouvait être *« que globale, car les compétences qui le constituent, avec leur liste principale de connaissances, de capacités et d'attitudes, sont complémentaires et également nécessaires »*. À ce principe de non-compensation s'ajoute une règle de collégialité : la décision d'attribution des compétences pouvant, par exemple, intervenir à l'occasion d'un conseil de classe.

Ces modalités d'attribution des compétences du socle commun par les enseignants ne sont pas sans introduire une certaine variabilité en fonction des caractéristiques individuelles des élèves, mais aussi de facteurs liés à leur établissement. En effet, on sait que le jugement porté par les enseignants ne s'explique pas uniquement par les performances effectives des élèves, mais qu'il peut aussi être influencé par des éléments d'ordre contextuel, tel que le niveau de la classe par exemple [Bressoux et Pansu, 2003]. D'une façon générale, les facteurs susceptibles d'introduire des biais dans la notation sont bien connus et documentés par les études entreprises dès le début du XXe siècle par l'école d'Henri Piéron. Concernant le socle commun, de récents travaux menés par la DEPP ont montré combien, pour des résultats identiques aux tests, les attestations des enseignants pouvaient varier selon des variables sociodémographiques et scolaires [Daussin, Rocher, Troseille, 2010]. C'est, par exemple, le cas des élèves « en retard », qui, à score et caractéristiques fixés, ont moins de chances de recevoir une attestation de compétences en français et en mathématiques.

Ces analyses rappellent toute la légitimité des évaluations standardisées[3] réalisées auprès d'échantillons représentatifs d'élèves pour répondre à la demande d'indicateurs comparables dans le temps tels que ceux exigés par la loi organique relative aux lois de finances (LOLF).

De 2007 à 2012, la DEPP a ainsi calculé une série d'indicateurs mesurant les proportions d'élèves maîtrisant les compétences « de base » en français et en mathématiques en fin d'école et en fin de collège, déclinées selon le secteur de l'établissement (public hors éducation prioritaire, public relevant de l'éducation prioritaire[4], privé). Ces épreuves

1. Loi d'orientation et de programme pour l'avenir de l'École (article 9) – Loi n° 2005-380 du 23 avril 2005.
2. Décret relatif au socle commun de connaissances et de compétences et annexe – Décret n° 2006-830 du 11 juillet 2006.
3. On définit ici les évaluations standardisées comme des dispositifs qui *« visent à mesurer les acquis cognitifs des élèves sur la base d'épreuves dont les conception, administration et correction sont uniformisées »* [Mons, 2009].
4. L'éducation prioritaire comprend les établissements qui relèvent du dispositif Éclair (Écoles, collèges et lycées pour l'ambition, l'innovation et la réussite) et du réseau de réussite scolaire (RRS).

LES SEPT COMPÉTENCES DU SOCLE COMMUN

Le socle commun s'organise actuellement en sept compétences : la maîtrise de la langue française (compétence 1), la pratique d'une langue vivante étrangère (compétence 2), les principaux éléments de mathématiques et la culture scientifique et technologique (compétence 3), la maîtrise des techniques usuelles de l'information et de la communication (compétence 4), la culture humaniste (compétence 5), les compétences sociales et civiques (compétence 6), l'autonomie et l'initiative (compétence 7).

Il convient de préciser que la présente étude se fonde sur la précédente version du socle. Le nouveau socle a été publié en avril 2015, après une consultation nationale lancée en 2014.

avaient été élaborées en 2005 et testées en 2006 [Rocher, Chesné, Fumel, 2008]. Au moment de leur conception, la notion de socle commun existait bien, mais aucun texte n'en définissait précisément le contenu. Les tests avaient donc été établis à partir d'éléments issus des programmes scolaires en relation avec le socle commun de connaissances et de compétences. Les résultats obtenus indiquaient une stabilité dans le temps en CM2 alors qu'en troisième ils mettaient en évidence une baisse des taux de maîtrise dans les établissements relevant de l'éducation prioritaire [*L'état de l'École*, 2012].

Les indicateurs mesurant les proportions d'élèves maîtrisant les compétences « de base » ont aujourd'hui laissé la place aux indicateurs de maîtrise des compétences du socle commun pour lesquels des expérimentations ont été entreprises dès 2009. La mise au point de tels indicateurs impose d'établir des scores seuils permettant de distinguer ceux qui atteignent le niveau souhaité de ceux qui ne l'atteignent pas. Mais comment déterminer ces seuils dont la définition n'est pas univoque ? C'est ce que cet article se propose d'étudier. Pour cela, il décrit trois méthodes de détermination des seuils proposées dans la littérature psychométrique. Il compare ensuite leur mise en œuvre respective dans le cadre de la détermination des seuils de maîtrise des compétences du socle commun évaluées par des tests standardisés en fin de CM2 et en fin de troisième. Il analyse enfin la pertinence des résultats obtenus au regard des caractéristiques des élèves, de leur environnement et de leur parcours scolaire.

ÉVALUATION DE LA MAÎTRISE DU SOCLE AUX PALIERS 2 ET 3 : QUELS OUTILS ?

Les données recueillies

C'est en 2011 que des tests standardisés ont été utilisés pour la première fois pour renseigner les indicateurs de maîtrise des compétences du socle commun de connaissances et de compétences en fin de CM2 et en fin de troisième. Cette année-là, les tests ont permis de fournir des indicateurs pour les compétences 1 et 3 en fin d'école et pour la compétence 1 en fin de collège. En 2012, cette démarche a été élargie aux compétences 2 et 5 à l'école et aux compétences 2, 3 et 5 au collège. En moyenne, les échantillons se composent d'environ 7 000 élèves par évaluation et les taux de réponse dépassent les 90 %. À l'école comme au collège, il s'agit

d'un sondage stratifié selon le secteur (public hors éducation prioritaire, public relevant de l'éducation prioritaire, privé). À l'école, les évaluations concernent tous les élèves de CM2 des établissements sélectionnés. Au collège, elles concernent tous les élèves d'une même classe de troisième. Pour les compétences 1 et 3, la dimension des échantillons a été augmentée en 2013 pour améliorer la précision des estimations. En particulier, le secteur de l'éducation prioritaire a été surreprésenté. En effet, pour ces compétences, les indicateurs doivent être déclinés pour le secteur public et le secteur privé, mais aussi, au sein du secteur public, pour les établissements relevant de l'éducation prioritaire (Éclair et RRS). En outre, les indicateurs de la LOLF demandent de renseigner les écarts observés entre les résultats des élèves de l'éducation prioritaire et ceux du secteur public hors éducation prioritaire.

Construction des épreuves

L'élaboration de ce dispositif d'évaluations standardisées tient compte de deux contraintes principales : le temps (les indicateurs doivent être mis à jour chaque année au mois de janvier) et le coût. Pour répondre à ces contraintes, la construction d'un test sous forme de QCM (questions à choix multiples) a été retenue. Les épreuves ont été élaborées spécifiquement pour chaque niveau évalué par des groupes de concepteurs composés d'enseignants et de conseillers pédagogiques en collaboration avec l'inspection générale. Ce format de questions assure une correction rapide, fiable et économique. En contrepartie, une interrogation sous cette forme exclut d'emblée l'évaluation de certains domaines de compétences. Par exemple, le domaine « dire » pour la compétence 1 aux paliers 2 et 3 et le domaine « écrire » pour la compétence 1 au palier 3.

Pour la compétence 5, à l'école et au collège, l'indicateur a été construit à partir d'items issus des évaluations Cedre (Cycle des évaluations disciplinaires réalisées sur échantillon)[5] pour l'histoire-géographie et l'éducation civique. Il convient donc de noter que cette évaluation ne concerne qu'une partie de la culture humaniste couverte par la compétence 5.

Les modèles de réponse à l'item

Une fois les données recueillies, une échelle de performances a été élaborée pour chaque évaluation en utilisant les modèles de réponse à l'item [voir notamment Grégoire et Laveault, 2002]. Ces derniers, développés dans la seconde moitié du XX^{e} siècle, modélisent la probabilité de réussite à un item en fonction de certaines de leurs caractéristiques, dont la difficulté, et en fonction du niveau de compétence des élèves.

Le modèle de réponse à l'item le plus simple a été développé en 1960 par Georg Rasch, le seul paramètre d'item à estimer étant la difficulté de l'item. Dans le cadre des évaluations des élèves conduites par la DEPP, un modèle de réponse à l'item à deux paramètres est utilisé. Ces deux paramètres d'items sont d'une part la difficulté et d'autre part la discrimination. Formellement, un modèle de réponse à l'item à deux paramètres s'écrit :

5. Les évaluations Cedre ont pour finalité de mesurer les atteintes des objectifs fixés par les programmes dans une discipline donnée et de comparer les performances des élèves dans le temps. Elles portent donc sur des contenus plus élargis que ceux évalués par les épreuves du socle.

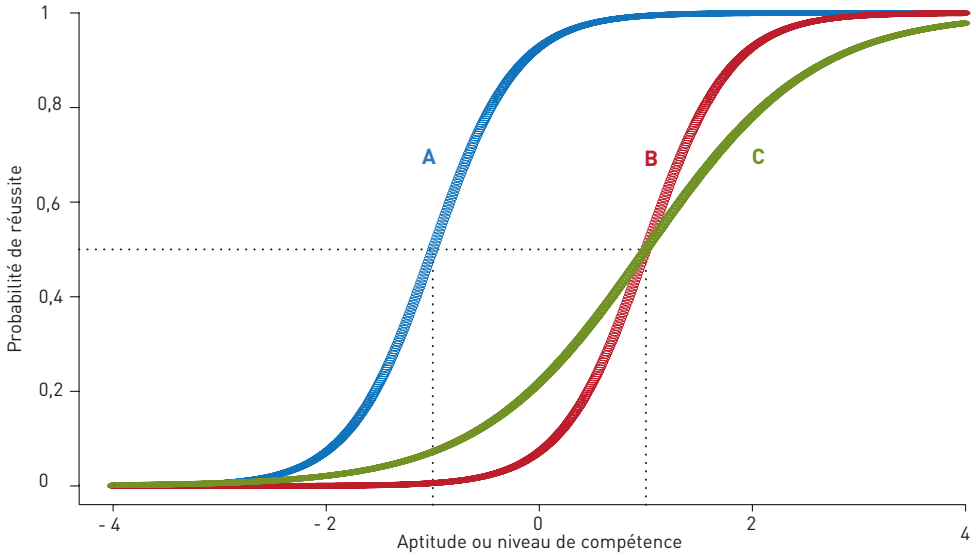
$$P_j(\theta_i) = \frac{e^{Da_j(\theta_i - b_j)}}{1 + e^{Da_j(\theta_i - b_j)}}$$

où $P_j(\theta_i)$ est la probabilité qu'un élève *i*, possédant une aptitude θ_i, réponde correctement à l'item *j*, b_j et a_j sont respectivement le paramètre de difficulté et le paramètre de discrimination de l'item *j*, et D un facteur d'échelonnement, constante fixée à 1,7 permettant ainsi de se rapprocher de la loi normale.

Cette dernière fonction peut être représentée par une courbe appelée courbe caractéristique de l'item qui met en relation l'aptitude ou la compétence de l'élève avec la probabilité de réussir un item donné. Généralement, l'aptitude θ des élèves est comprise entre – 4 et + 4. La **figure 1** illustre trois courbes caractéristiques pour des items dont le niveau de difficulté ou de discrimination varie. Par convention, la valeur qui représente la difficulté d'un item est la valeur de θ pour laquelle la probabilité de donner une réponse correcte est de 0,5, c'est-à-dire une chance sur deux. Ainsi, l'aptitude des élèves et les difficultés des items sont représentées sur une même échelle. Il s'agit là d'un avantage indéniable des modèles de réponses à l'item par rapport à l'approche usuelle (nombre de bonnes réponses), notamment lors de la détermination de seuils de maîtrise.

En effet, cette échelle permet de décrire les compétences d'un niveau donné en se servant des items correspondants comme descripteurs de ces compétences quand l'approche classique ne permet que de situer les élèves au sein de la distribution des scores sans renseigner sur les compétences attribuables aux niveaux où ils se situent. Sur la figure 1, la difficulté de l'item A est de – 1 et celle des items B et C de 1. Quant à la discrimination de l'item, elle est représentée par la pente de la courbe

Figure 1 **Différentes courbes caractéristiques des items (A, B, C) selon la valeur du paramètre de difficulté et de discrimination**



Source : MENESR-DEPP.

au point d'inflexion (plus elle est forte, plus l'item est discriminant). Ici, les items A et B présentent la même discrimination alors que l'item C est moins discriminant (la pente de la courbe étant plus faible).

Une fois le modèle théorique défini, il s'agit d'estimer les paramètres du modèle : la difficulté et la discrimination de chaque item et l'aptitude de chaque élève. L'estimation simultanée de tous ces paramètres s'avère complexe et dépasse le cadre de cet article (le lecteur intéressé par cette problématique pouvant se reporter à Rocher, dans ce numéro, p. 37).
Notons que l'aptitude des élèves, telle que définie par le modèle de réponse à l'item est très fortement corrélée avec le score (défini par le nombre de bonnes réponses à l'évaluation). Ainsi, le modèle de réponse à l'item donne des résultats proches de la théorie classique consistant à sommer le nombre de bonnes réponses. L'aptitude estimée par le modèle de réponse à l'item le plus simple (un seul paramètre, la difficulté) est d'ailleurs en bijection avec le nombre de bonnes réponses. L'ajout du paramètre de discrimination permet de casser ce lien univoque entre nombre de bonnes réponses et aptitude estimée par le modèle. En quelque sorte, ce modèle à deux paramètres revient à pondérer les items selon leurs discriminations (à titre d'illustration, si deux élèves ont le même nombre de bonnes réponses, mais obtenues sur des items différents, celui ayant réussi les items les plus discriminants aura une aptitude plus élevée).

Les modèles de réponse à l'item présentent aussi l'avantage de pouvoir situer sur une même échelle de compétences des élèves qui n'ont pas nécessairement été soumis aux mêmes items. C'est notamment le cas lorsque l'on utilise la méthode des « cahiers tournants » pour évaluer un nombre important d'items sans allonger le temps de passation. Cette méthode consiste à répartir les items dans des cahiers différents qui comportent des items communs.

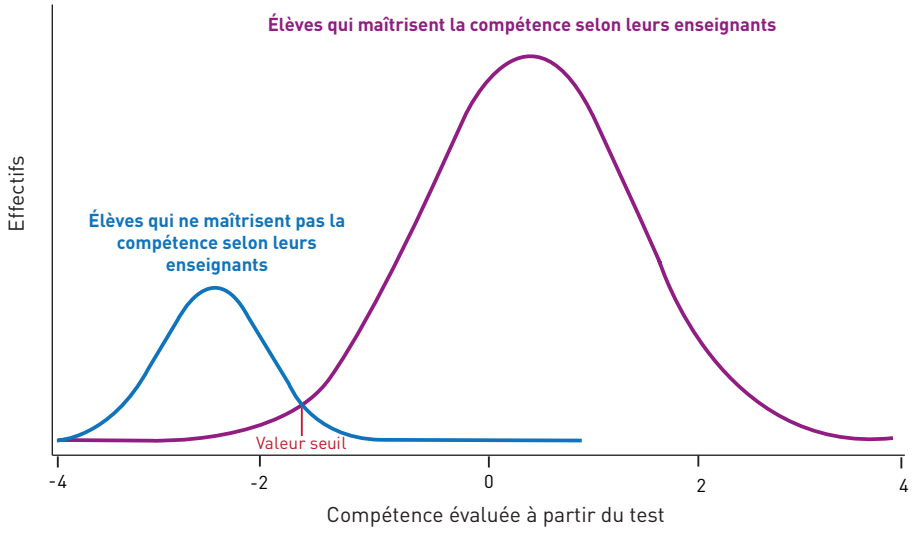
TROIS MÉTHODES POUR DÉTERMINER LE SEUIL DE MAÎTRISE DES COMPÉTENCES

Le calcul des taux de réponses correctes aux items ne permet donc pas d'estimer directement la proportion des élèves qui maîtrisent ces compétences, car le « degré de maîtrise » n'a pas encore été défini à ce stade. En effet, les items peuvent être de difficulté très variable, quand bien même ils portent sur une compétence dite du « socle ». De ce fait, pour être considéré comme un élève qui maîtrise les compétences du socle, l'élève doit-il réussir toutes les questions qui lui sont proposées ? Les trois quarts ? La moitié ? C'est ce seuil qui doit être fixé, seuil à partir duquel on considérera qu'il maîtrise les compétences du socle. La détermination de ce seuil ne s'impose pas d'elle-même.
Différentes méthodes ont été proposées dans la littérature pour déterminer ce point de bascule entre maîtrise et non-maîtrise. Quelle que soit la méthode utilisée, ce point de césure est défini par la confrontation entre les attentes des enseignants ou d'un groupe d'experts et les résultats statistiques. Trois de ces méthodes ont été retenues dans le cadre de la détermination des seuils de maîtrise des compétences du socle commun. La première repose sur le jugement des enseignants sur leurs élèves et les deux autres sur le jugement des items par des enseignants et des experts.

La première méthode possible pour déterminer le point de césure est celle dite des contrastes décrite par Grégoire et Laveault [2002]. Cette méthode confronte le jugement des enseignants sur les élèves avec les résultats des élèves aux tests standardisés. Il est demandé aux enseignants d'indiquer pour chacun de leurs élèves si celui-ci maîtrise ou non la compétence évaluée. On répartit alors les élèves en deux groupes, selon leur maîtrise supposée de la compétence évaluée, et reporte alors sur un même graphique la distribution de l'aptitude des élèves, estimée à partir du modèle de réponse à l'item pour chacun des deux groupes.
Pour que cette méthode puisse être utilisée, les deux distributions doivent être suffisamment disjointes. Un cas d'école, où les deux distributions se croisent, est reporté sur la **figure 2**. Le point d'intersection correspond alors au point de césure entre maîtrise et non-maîtrise (valeur seuil sur la figure 2). Cette méthode présente notamment l'avantage de pouvoir comparer un dispositif de classement des élèves par les enseignants – dont nous avons déjà présenté les limites – avec les résultats d'évaluations standardisées.

Une deuxième méthode peut aussi être utilisée pour déterminer des seuils de maîtrise. Il s'agit d'une adaptation de la méthode dite des « zones de jugement », ou Hofstee, décrite par Bunch et Cizek [2007]. En théorie, cette méthode utilise le score, c'est-à-dire le nombre de bonnes réponses au test, mais elle peut également être adaptée en utilisant l'aptitude estimée par le modèle de réponse à l'item. C'est notamment le cas lorsque l'évaluation utilise un nombre important d'items sans que les élèves les passent nécessairement tous (méthode des « cahiers tournants » par exemple).
La présentation ci-dessous fait référence à l'utilisation de la méthode Hofstee dans le cas où le score est défini en termes de réponses correctes.
Comme avec la méthode des contrastes, l'information apportée par les enseignants des élèves est utilisée, mais également celle tirée des experts ayant participé à la

Figure 2 **Exemple théorique de la méthode des contrastes**



Source : MENESR-DEPP.

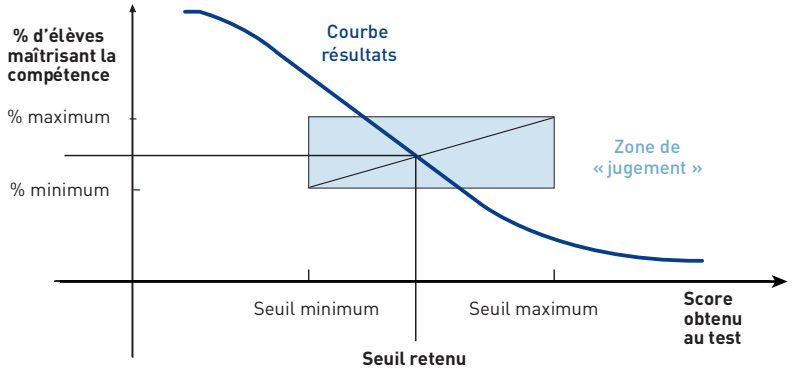
conception du test. Deux questions leur sont posées. La première leur demande de se prononcer, en se basant sur leur expérience professionnelle, sur le pourcentage minimum et le pourcentage maximum d'élèves maîtrisant, selon eux, la compétence en question. Il leur est ensuite demandé de déterminer à partir de quel score les élèves maîtrisent cette compétence. Puisqu'il est difficile de se déterminer sur le score, les enseignants et les experts travaillent sur les items. Plus précisément, ils doivent classer chaque item du test en trois catégories (A, B, C) selon le degré d'importance accordé à la réussite de l'item pour la validation de la compétence ▸ **Tableau 1**. Le nombre minimum de bonnes réponses pour maîtriser la compétence, correspondant à une notation souple, est obtenu en sommant les items classés « A » et le score maximum, correspondant à une notation plus sévère, est obtenu en sommant les items classés « A » ou « B ».
À chaque expert correspond alors une zone de jugement ▸ **Figure 3**. Cette zone témoigne à la fois des attentes et du niveau d'exigence de chaque expert. À partir des données issues de l'expérimentation, il est possible de calculer pour chaque score le pourcentage d'élèves situés au-delà de ce score et de tracer la courbe correspondante. Le point d'intersection entre cette courbe et la diagonale de la zone de jugement fournit le score-seuil retenu.

▸ Tableau 1 Classification des items en trois catégories (méthode Hofstee)

Selon vous, les élèves maîtrisant la compétence...	... doivent absolument réussir cet item (A)	... devraient pouvoir réussir cet item (B)	... ne doivent pas forcément réussir cet item (C)
Item 1			
Item 2			
...			
Item N			

Source : MENESR-DEPP.

▸ Figure 3 Méthode des « zones de jugement »

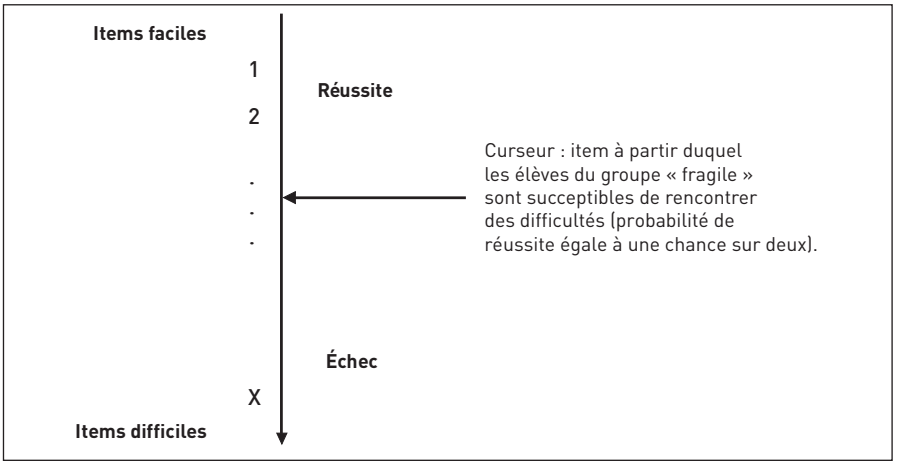


Source : MENESR-DEPP.

Enfin, une troisième méthode est utilisée pour déterminer le point de césure entre maîtrise et non-maîtrise des compétences du socle. Il s'agit de la méthode des marque-pages (bookmarks). Simple à mettre en œuvre et plus couramment utilisée que les précédentes, elle est aussi décrite par Bunch et Cizek [2007]. À partir des résultats de l'estimation du modèle de réponse à l'item, les items sont classés par ordre croissant selon la valeur de leur paramètre de difficulté. Les items du début de la liste correspondent à des items faciles, c'est-à-dire très réussis, et ceux de la fin sont plus difficiles ▸ **Figure 4**.

Comme nous l'avons vu, grâce aux modèles de réponse à l'item, les paramètres de difficulté des items et les niveaux de compétences des élèves sont positionnés sur une même échelle. Plus précisément, chaque item est positionné à un niveau tel que les jeunes situés à ce niveau ont une chance sur deux de réussir cet item et ceux qui se situent en dessous ont une probabilité de réussite plus faible. Il est alors demandé à chaque expert d'imaginer un groupe d'élèves fragile (élèves situés à la frontière entre le groupe « maîtrise » et le groupe « non-maîtrise ») et d'indiquer l'item (ou une zone réduite d'items) à partir duquel (de laquelle) ces élèves sont susceptibles de rencontrer des difficultés et d'avoir une chance sur deux de réussir. Ainsi, pour chaque expert, un pourcentage (s'il a fourni un seul item comme « zone de bascule ») ou un intervalle (si plusieurs items fournis) de maîtrise de la compétence est déterminé.
La seconde phase consiste à faire converger les attentes des différents experts pour aboutir à un consensus autour de la définition d'un seuil de maîtrise au regard des résultats obtenus à partir de la mise en œuvre de ces trois méthodes.

▸ **Figure 4** **Méthode des marque-pages**



Source : MENESR-DEPP.

DÉTERMINATION DES POINTS DE CÉSURE POUR L'ÉVALUATION DES COMPÉTENCES DU SOCLE COMMUN

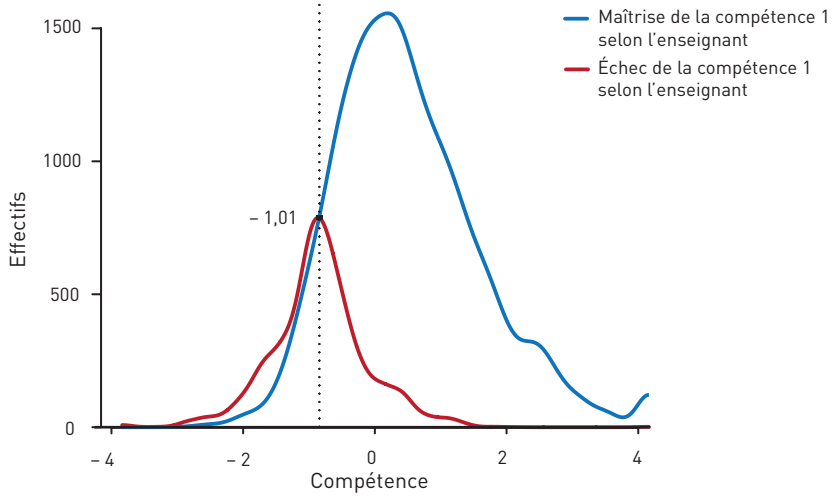
Selon les évaluations, le point de césure entre maîtrise et non-maîtrise a été défini en 2011 ou en 2012 en mettant en application les trois méthodes présentées précédemment. Selon les évaluations, le point de bascule a aussi été défini en 2011 ou en 2012. Il a été repris à l'identique les années suivantes, les évaluations étant composées exclusivement d'items communs dont les paramètres sont supposés invariants entre les différentes années.

Exemple pour la compétence 1 à l'école

Le point de césure pour la compétence 1 à l'école, c'est-à-dire le niveau d'aptitude à partir duquel un élève de CM2 maîtrise la compétence, a été déterminé en 2011 sur la base d'un échantillon d'environ 7 500 élèves.

La méthode des contrastes a été utilisée dans un premier temps. On a donc demandé aux enseignants de l'échantillon d'indiquer pour chacun de ses élèves si celui-ci maîtrise ou non la compétence 1, sur la base du travail effectué tout au long de l'année scolaire. La distribution des scores des élèves pour chacun des deux groupes (maîtrise / non maîtrise) a ensuite été représentée en fonction de cette information fournie par chaque enseignant ▸ **Figure 5**. Le point d'intersection entre ces deux distributions correspond à un niveau de compétence de – 1,01. Il conduit à un niveau de validation de 83 %. Cependant, le recouvrement des deux distributions ne permet pas de classer avec confiance un élève dans l'une ou l'autre des deux catégories à partir de son score au test. On ne saurait donc déterminer un seuil pertinent à partir de ces seuls résultats.
Quant à la méthode Hofstee (zone de jugement), utilisée sur la base du score estimé par le modèle de réponses à l'item, elle conduit à des pourcentages de validation

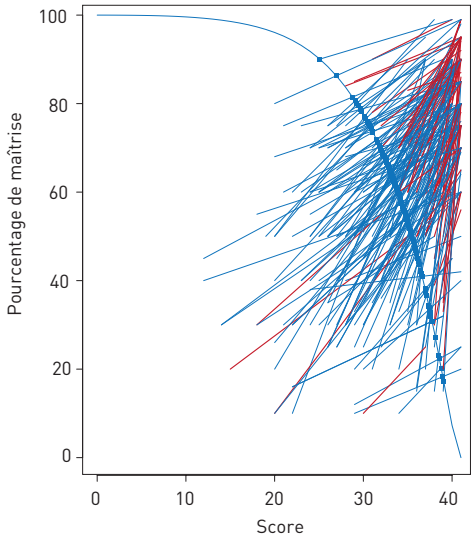
▸ **Figure 5** **Méthode des contrastes pour la compétence 1 à l'école en 2011**



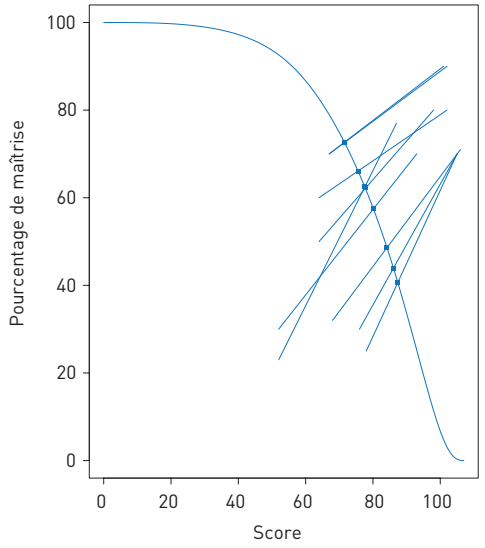
Source : MENESR-DEPP.

de la compétence 1 très variables à partir des données recueillies auprès des 263 professeurs des classes de l'échantillon ▸ Figure 6. La médiane est de 57 %, mais pour certains enseignants la proportion d'élèves maîtrisant la compétence ne dépasse pas 20 %, alors que pour d'autres, elle se situe au-delà de 80 %. Une des explications à cette dispersion, validée par des analyses secondaires entreprises autour de la méthode Hofstee, réside dans les différences de jugement des enseignants selon le secteur de leur établissement. Les résultats obtenus à partir des travaux effectués par neuf experts indiquent eux aussi des disparités de jugement comparables à celles observées parmi les enseignants de l'échantillon ▸ Figure 7. D'une manière générale, la mise en application de cette méthode ne permet donc pas d'obtenir un consensus compte tenu de la dispersion des taux de maîtrise qui en résultent.

▸ Figure 6 Méthode des « zones de jugement » pour la compétence 1 à l'école à partir des données recueillies auprès des professeurs de l'échantillon en 2011



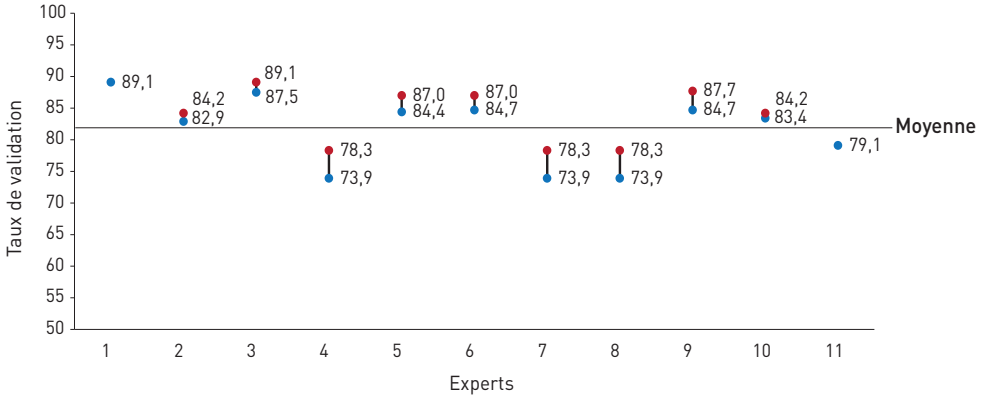
▸ Figure 7 Méthode des « zones de jugement » pour la compétence 1 à l'école à partir des données recueillies auprès des experts en 2011



Note : l'axe des abscisses est le score à l'évaluation, l'axe des ordonnées est le pourcentage d'élèves maîtrisant la compétence 1. La courbe représente, pour chaque score, le pourcentage d'élèves situés au-delà de ce score. Le point d'intersection entre cette courbe et la diagonale de la zone de jugement fournit le score-seuil retenu. Seules les diagonales représentées en bleu coupent la courbe.
Source : MENESR-DEPP.

Le seuil de coupure entre maîtrise et non-maîtrise a été obtenu par la méthode des marque-pages appliquée avec onze experts, dont les neuf ayant appliqué la méthode des « zones de jugement ». Les items de l'évaluation ont été classés par ordre de difficulté croissante (tri sur les paramètres de difficulté estimés par le modèle de réponse à l'item) et chaque expert a déterminé un item (ou un intervalle d'items) comme point de césure. La variabilité inter-experts est marquée, mais est tout de même sensiblement plus faible que celle observée lors de l'utilisation de la méthode des « zones de jugement ». Le pourcentage de maîtrise de la compétence sur les données 2011 oscille entre 73,9 % et 89,1 %. La moyenne est à 82,6 % ▸ Figure 8.

Figure 8 Méthode des marque-pages pour la compétence 1 à l'école en 2011



Note de lecture : l'intervalle d'items choisi par l'expert numéro 4 conduit à un taux de validation de la compétence 1 variant de 73,9 % à 78,3 %.
Source : MENESR-DEPP.

Conformément à la méthode, une discussion s'est alors engagée pour dégager un éventuel consensus. Finalement, l'item retenu comme point de césure présente une difficulté de - 0,9 conduisant à une maîtrise de la compétence 1 par 79 % des élèves. Les experts ont considéré que les élèves fragiles commençaient à échouer sur des items ayant trait au traitement de l'implicite. Ces élèves parviennent à mettre en relation les informations explicites d'un texte, mais peinent à s'engager dans un processus d'interprétation.

Les résultats de 2012 et de 2013 s'appuient sur les enseignements tirés lors de la session précédente. En effet, les items de ces évaluations ont été extraits de l'évaluation 2011. Dans le cadre de la théorie du modèle de réponse à l'item, les paramètres des items sont supposés invariants entre ces deux sessions [Rocher, dans ce numéro, p. 37]. Plus précisément, le paramètre de difficulté de chaque item est supposé être constant. Le point de césure déterminé en 2011 est de nouveau appliqué les années suivantes et conduit, en 2013, à un taux de validation de 79,8 % pour la compétence 1 à l'école ▸ Tableau 2 p. 155.

Pour les autres compétences, une démarche similaire à celle qui vient d'être exposée a permis de déterminer un item-seuil défini comme point de bascule entre maîtrise et non-maîtrise. Les trois méthodes présentées (contrastes, zone de jugement, marque-pages) ont aussi été mises en œuvre, mais ce sont les résultats fournis par la méthode des marque-pages qui se sont révélés les plus probants. En effet, la méthode des contrastes n'a pas toujours permis d'aboutir à des résultats utilisables. Pour la majorité des compétences, l'aptitude des élèves ne différait pas suffisamment selon que les enseignants des élèves validaient ou non la compétence. Cependant, même lorsque cette méthode ne peut pas être utilisée, elle n'en est pas moins utile puisque son échec tend une nouvelle fois à prouver que la validation des compétences par les enseignants n'est pas sans poser question et justifie *a posteriori* le recours à une évaluation standardisée pour déterminer la proportion de maîtrise du socle. Quant aux résultats obtenus à partir de la méthode des « zones de

jugement », ils ont le plus souvent permis d'éclairer les discussions des experts lors de la phase de recherche de consensus de la méthode des marque-pages.
Concernant la pratique d'une langue vivante étrangère au collège (compétence 2), dont le taux de maîtrise est plus faible, la méthode des marque-pages a aussi été appliquée à d'autres données d'évaluation ▸ **Encadré**. De manière générale, il est notable que, malgré la variabilité dans les attentes des enseignants et des experts, l'item retenu comme point de césure dans toutes les compétences correspond au passage de l'explicite à l'implicite.

DÉTERMINATION DU SEUIL DE MAÎTRISE POUR LA COMPÉTENCE 2 AU COLLÈGE

Pour la compétence 2 au collège, l'indicateur a été établi pour la première fois en 2012 à partir des réponses apportées par un échantillon d'élèves à des épreuves standardisées portant sur leurs compétences en anglais (compréhension orale et compréhension écrite).
Pour la compréhension orale, le pourcentage de maîtrise retenu (26,9 %) est très proche de celui observé dans les évaluations de l'Étude européenne sur les compétences en langues (ESLC) en fin de scolarité obligatoire (jeunes de 14 à 16 ans) [Bessonneau et Verlet, 2012]. Les résultats de cette enquête montrent en effet que, pour la compréhension orale, seuls 26 % des élèves français maîtrisent au moins le niveau A2[6]. En revanche, pour la compréhension écrite, on observe un écart significatif : 40,4 % à partir des épreuves du socle contre 22,8 % dans ESLC.
Au regard de ces résultats, il a été demandé au groupe d'experts de participer à une nouvelle séance de mise en application de la méthode des marque-pages aux items de compréhension écrite de l'épreuve ESLC. Même si cette évaluation ne repose pas sur les mêmes protocoles et méthodologies, elle comporte un plus grand nombre d'items que celle du socle, permettant ainsi d'affiner la détermination du point à partir duquel s'opère le basculement entre maîtrise et non-maîtrise. Ce nouveau seuil a permis d'aboutir à un taux de validation de la compétence 2 assez proche de celui observé à partir de l'épreuve du socle (43 %), ce qui témoigne d'une cohérence dans le jugement apporté par les experts. Ce résultat demeure nettement supérieur à celui observé dans ESLC où la définition des seuils de maîtrise a été effectuée par des experts de différents pays[7], méthode susceptible d'introduire des biais liés aux limites de la comparabilité des niveaux de difficultés des items entre pays.
L'indicateur de maîtrise de la compétence a ensuite été construit en faisant la moyenne entre la compréhension écrite et la compréhension orale. Ainsi, en 2012, 35,1 % des élèves maîtrisent la compétence 2 en anglais en fin de collège. En 2013, en appliquant le même seuil, ils sont 36,6 % ▸ **Tableau 2 p.155**.

6. Sur l'échelle européenne (CECRL), le niveau A2 – exigé pour la validation du socle commun de connaissances et de compétences – correspond à la mention « utilisateur élémentaire de l'anglais ». À ce niveau, l'utilisateur élémentaire peut comprendre des phrases isolées et des expressions fréquemment utilisées en relation avec des domaines immédiats de priorité (par exemple, informations personnelles et familiales simples, achats, etc.) ; peut communiquer lors de tâches simples et habituelles ne demandant qu'un échange d'informations simple et direct sur des sujets familiers et habituels ; peut décrire avec des moyens simples son environnement immédiat et évoquer des sujets qui correspondent à des besoins immédiats.

7. Les méthodologies utilisées dans le cadre de ESLC sont décrites dans le rapport technique du consortium : https://crell.jrc.ec.europa.eu

ANALYSE DES RÉSULTATS ET PERSPECTIVES

Les démarches entreprises lors de la détermination des seuils de maîtrise avaient pour objectif d'obtenir des résultats fiables et destinés à être comparés dans le temps, mais aussi de contribuer à la réflexion méthodologique sur le sujet.
En 2013, selon la procédure présentée ci-dessus, en fin de CM2, 79,8 % des élèves maîtrisent la compétence 1 et 70,9 % la compétence 3 ▸ **Tableau 2**. En fin de troisième, ils sont respectivement 79,2 % et 78,3 %. À l'école, les garçons sont moins nombreux à maîtriser la compétence 1 que les filles (77,1 % contre 82,6 %, l'écart étant significatif à un seuil inférieur à 1/1000). La différence s'accentue au collège (72,3 % contre 85,9 %).
Pour la compétence 3, la différence selon le sexe (là aussi, statistiquement significative) s'inverse légèrement à l'école (72,5 % des garçons contre 69,3 % des filles), mais les filles devancent les garçons au collège (80,5 % des filles contre 76,2 % des garçons). Il peut être souligné que les disparités entre filles et garçons avaient déjà été obtenues lors des épreuves réalisées en 2012, mais la dimension plus faible des échantillons ne permettait pas de conclure systématiquement à des différences significatives. La maîtrise plus fréquente des compétences du socle dans les disciplines scientifiques (mathématiques, physique-chimie, sciences de la vie et de la Terre, technologie) en fin de collège par les filles se retrouve également sur les notes au contrôle continu du diplôme national du brevet (DNB). En 2013, la moyenne des filles était supérieure à celle des garçons de 0,2 point en physique-chimie, de 0,4 point en mathématiques, de 0,7 point en technologie et de 0,8 point en sciences de la vie et de la Terre. De tels écarts étaient également constatés à la session 2012 du DNB. Sauf à considérer qu'il existe un biais de notation important en faveur des filles, les résultats de maîtrise du socle selon le sexe des élèves sont corroborés par les résultats exhaustifs au contrôle continu du DNB. Pour la compétence 2 (pratique d'une langue vivante étrangère), les performances des filles sont aussi supérieures à celles des garçons, que ce soit à l'école ou au collège. Ces résultats confirment les observations issues des évaluations Cedre [Bessonneau, Beuzon, Boucé *et alii*, 2012 ; Bessonneau, Beuzon, Daussin *et alii*, 2012]. En revanche, elles sont proportionnellement moins nombreuses à maîtriser la compétence 5 (culture humaniste) à l'école et au collège. Ici aussi, les résultats sont à rapprocher des analyses effectuées à partir des évaluations Cedre en histoire-géographie et en éducation civique. Ceux-ci indiquent que les garçons sont plus nombreux parmi les élèves de haut niveau. En revanche, les proportions d'élèves dans les groupes de bas niveau sont pratiquement les mêmes chez les filles et les garçons [Garcia et Pastor, 2013 ; Garcia et Krop, 2013]. Les résultats des deux évaluations ne se recoupent donc que partiellement.

Les élèves en retard représentent en moyenne 12 % des élèves de fin de CM2 interrogés. En fin de troisième, ils sont un peu plus d'un quart. Que ce soit en fin d'école ou en fin de collège, la proportion d'élèves qui maîtrisent les compétences évaluées est nettement moins importante parmi les élèves « en retard » que parmi les élèves « à l'heure ». La différence entre les deux groupes d'élèves est particulièrement marquée à l'école où elle se situe autour de 40 points de pourcentage pour les compétences 1 et 3 et autour de 30 points pour la compétence 2. Au collège, les différences sont un peu moins élevées, mais l'écart entre les deux groupes reste tout de même important.

Tableau 2 Proportion d'élèves qui maîtrisent les compétences du socle commun en 2013 (en %)

Compétence	Public hors EP	RRS	Éclair	Privé	Public	Garçons	Filles	« En retard »	« À l'heure »	Ensemble
C1 école	81,8	69,8	62,5	87,4	78,6	77,1	82,6	46,2	84,7	79,8
C3 école	74,2	56,5	47,3	79,2	69,6	72,5	69,3	33,1	76,3	70,9
C2 école	81,6	61		85,9	78,1	75,0	83,8	54,3	82,8	79,3
C5 école	70,9	59,1	53,5	78,9	70,6	74,9	69,1	35,5	76,6	72,1
C1 collège	80,6	70,1	56,7	87,9	78,7	72,3	85,9	55,6	86,5	79,2
C3 collège	80,4	67,7	51,5	88,1	77,9	76,2	80,5	52,7	86,6	78,3
C2 collège	36,1	20,9		48,3	33,4	33,2	39,9	15,4	41,9	36,6
C5 collège	67,0	55,1	39,3	79,5	66,6	71,1	68,4	40,7	77,9	69,8

Lecture : en fin de CM2, 81,8 % des élèves des établissements public hors éducation prioritaire maîtrisent la compétence 1.
Note : comme toutes les estimations réalisées sur échantillons, les résultats du tableau présentent des intervalles de confiance. Selon les compétences et la variable étudiée, les intervalles de confiance sont compris dans une fourchette allant de +/- 1,2 à +/- 4,1.
Champ : France métropolitaine + DOM.
Source : MENESR-DEPP.

Les données permettent aussi de comparer les niveaux de maîtrise selon les secteurs d'enseignement et, plus particulièrement, entre les établissements publics relevant de l'éducation prioritaire et les autres. On constate ainsi qu'en fin de CM2, comme en fin de troisième, les élèves scolarisés en Éclair maîtrisent moins bien que les autres élèves les compétences 1, 3 et 5 du socle commun. Par exemple, si 62,5 % des élèves de CM2 des écoles du programme Éclair maîtrisent la compétence 1 du socle, ils sont 69,8 % dans les écoles RRS et 81,8 % dans les écoles publiques hors éducation prioritaire.

L'origine sociale de l'élève influence également la maîtrise des compétences du socle. Les enfants de cadres ou de personnes exerçant une profession intellectuelle supérieure maîtrisent très souvent les compétences 1 et 3 du socle (respectivement 91,3 % et 92,3 % pour les compétences 1 et 3 au collège). Cette proportion de maîtrise est plus faible pour les enfants ayant une origine sociale défavorisée[8] (75,8 % maîtrisent la compétence 1 ; 73,9 % maîtrisent la compétence 3).

La situation de la grande majorité des élèves de troisième à la rentrée scolaire suivant l'évaluation a pu être déterminée à partir du système d'information du ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche : 70,5 % d'entre eux étaient en seconde générale et technologique, 22,3 % en seconde professionnelle, 4,4 % en CAP et 2,7 % de nouveau en classe de troisième.
Il s'avère que les résultats aux évaluations présentent un caractère prédictif intéressant. En effet, 9 élèves sur 10 ayant poursuivi en seconde générale et technologique ont été évalués positivement à l'évaluation en troisième (90,7 % pour la compétence 1 ; 91 % pour la compétence 3). Les élèves qui sont toujours en classe de troisième à la rentrée suivant l'évaluation maîtrisent moins fréquemment la compétence 1 ou la compétence 3 (72,8 % pour la compétence 1 ; 71,8 % pour la compétence 3). La maîtrise de ces compétences est encore moins fréquente

8. Ouvriers, retraités ouvriers et employés, inactifs (chômeurs et inactifs n'ayant jamais travaillé).

pour les élèves orientés dans l'enseignement professionnel (66,5 % en seconde professionnelle et 53 % en CAP pour la compétence 1 ; 65 % en seconde professionnelle et 44,3 % en CAP pour la compétence 3). Ces résultats sur le devenir des élèves confèrent une forme de validité aux évaluations standardisées administrées par la DEPP.

Les résultats présentés dans cet article rappellent aussi tout l'intérêt de la confrontation des méthodes de définition des seuils de maîtrise. Parmi elles, c'est la méthode des marque-pages qui s'est révélée la plus adaptée. De ce fait, elle sera privilégiée dans les prochains travaux portant sur la détermination des seuils de maîtrise qui vont de nouveau être entrepris par la DEPP sur les compétences 1 et 3. Ils porteront tout d'abord sur le palier 1 pour lequel des évaluations se sont tenues en fin de CE1 en mai 2014. Ensuite, ils s'appuieront sur des évaluations qui seront effectuées respectivement en début de sixième (2015) et en fin de troisième (2016).

BIBLIOGRAPHIE

Bessonneau P., Verlet I., 2012, « Les compétences en langues étrangères des élèves en fin de scolarité obligatoire. Premiers résultats de l'Étude européenne sur les compétences en langues 2011 », *Note d'information*, n° 12.11, MEN-DEPP.

Bessonneau P., Beuzon S., Boucé S., Daussin J.-M., Garcia É., Levy M., Marchois C., Trosseille B., 2012, « L'évolution des compétences en langues des élèves en fin de collège de 2004 à 2010 », *Note d'information*, n° 12.05, MENJVA-DEPP.

Bessonneau P., Beuzon S., Daussin J.-M., Garcia É., Levy M., Marchois C., Trosseille B., 2012, « L'évolution des compétences en langues des élèves en fin d'école de 2004 à 2010 », *Note d'information*, n° 12.04, MENJVA-DEPP.

Brennan R., Kolen M., 2004, *Test Equating, Scaling, and Linking. Methods and Practices,* 2nd edition, New York, Springer, 548 p.

Bressoux P., Pansu P., 2003, *Quand les enseignants jugent leurs élèves*, Paris, Presses universitaires de France, 190 p.

Bunch M., Cizek G., 2007, *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests,* London, Thousand Oaks, Sage Publications, 352 p.

Daussin J.-M., Rocher T., Trosseille B., 2010, « L'attestation de la maîtrise du socle commun est-elle soluble dans le jugement des enseignants ? » *Éducation & formations*, n° 79, MENJVA-DEPP, p. 45-58.

Garcia É., Pastor J.-M., 2013, « CEDRE 2012 histoire-géographie et éducation civique en fin d'école primaire : grande stabilité des acquis depuis six ans », *Note d'information*, n° 13.10, MEN-DEPP.

Garcia É., Krop J., 2013, « CEDRE 2012 histoire-géographie et éducation civique : baisse des acquis des élèves de fin de collège depuis six ans », *Note d'information*, n° 13.11, MEN-DEPP.

Grégoire J., Laveault D., 2002, *Introduction aux théories des tests en psychologie et en sciences de l'éducation,* 2e édition, Bruxelles, De Boeck, 377 p.

L'état de l'École, 2012, Paris, MEN-DEPP.

Le Donné N., Rocher T., 2010, « Une meilleure mesure du contexte socio-éducatif des élèves et des écoles. Construction d'un indice de position sociale à partir des professions des parents », *Éducation & formations,* n° 79, MENJVA-DEPP, p. 103-115.

Mons N., 2009, « Effets théoriques et réels des politiques d'évaluation standardisée », *Revue française de pédagogie,* n° 169, p. 99-140.

Rocher T., Chesné J.-F., Fumel S., 2008, « Méthodologie de l'évaluation des compétences de base en français et en mathématiques en fin d'école et en fin de collège », *Note d'information,* n° 08.37, MEN-DEPP.

UNE ÉVALUATION SOUS FORME NUMÉRIQUE EST-ELLE COMPARABLE À UNE ÉVALUATION DE TYPE « PAPIER-CRAYON » ?

Pascal Bessonneau
MENESR-DEPP, bureau de l'évaluation des actions éducatives et des expérimentations

Philippe Arzoumanian et Jean-Marc Pastor
MENESR-DEPP, bureau de l'évaluation des élèves

Aujourd'hui, la place prépondérante prise par l'informatique questionne l'école sur la transition d'un environnement dominé par le papier vers un environnement dominé par le support numérique. Cette transition est en marche dans le domaine des évaluations standardisées. Cependant, la question de la comparabilité de la mesure est posée. L'hypothèse sous-jacente d'une transition naturelle et sans contrainte d'un support à l'autre doit en effet être interrogée.
De nombreux articles tirés de la littérature scientifique comparent les performances des élèves à des évaluations proposées sur support papier et sur support électronique. Ces études indiquent des résultats divergents. Les tests sont parfois plus faciles, parfois plus difficiles, ou de même difficulté.
Une question se pose : un item peut-il être proposé aux élèves à l'identique dans les deux supports sans influencer sa difficulté et sans provoquer de modification des compétences mises en jeu ?
L'article présente les résultats de deux expériences menées sur ce thème dans le cadre des évaluations standardisées conduites par la DEPP. La première cherche à identifier les différences de difficulté des items entre le support papier et le support numérique, à partir d'une évaluation des compétences de base en français et en mathématiques conduite en fin de primaire. La seconde expérience tente de dégager les variables explicatives de ces différences sur la base d'une étude menée en mathématiques en fin d'école et en fin de collège dans le cadre de Cedre.

La DEPP a pour mission de concevoir et de développer des évaluations pour mettre à disposition des responsables du système éducatif des informations rigoureuses et objectives, tant sur l'évolution des connaissances et des

compétences des élèves que sur leur développement conatif [Trosseille et Rocher, dans ce numéro, p. 15]. De plus en plus, ces évaluations sont réalisées sur support numérique. Se pose donc la question de la comparabilité des épreuves sur les deux supports, notamment dans le cas de reprise d'épreuves afin d'établir des comparaisons temporelles. Les items peuvent-ils être proposés aux élèves à l'identique sans changer la difficulté et les compétences mises en jeu ?

De nombreux tests institutionnels et commerciaux sont proposés sous forme papier et/ou numérique en parallèle, ou bien exclusivement sous forme numérique. Pour citer des enquêtes internationales récentes, PISA (*Programme for International Student Assessment*) a interrogé en 2009 une partie des élèves sous forme numérique [OCDE, 2011] et pour la session 2015, les évaluations seront entièrement proposées sous forme numérique. L'enquête ESLC *(First European Survey on Language Competences)* a interrogé, en 2011, dans certains pays une partie des élèves sous format numérique (Commission européenne, 2012). De même, l'enquête Piaac (*Programme for the International Assessment of Adult Competencies*) a également évalué les individus selon les deux modalités, papier-crayon et numérique [voir Murat et Rocher, dans ce numéro, p. 83].

Ces enquêtes sont illustratives des problèmes de comparabilité rencontrés. En 2009, PISA a interrogé les élèves sur du matériel créé spécialement pour le support informatique. Il s'agissait d'une évaluation spécifique de la lecture dans un environnement électronique : ERA (*Electronic Reading Assessment*). En 2015, le mode d'interrogation sous forme numérique sera généralisé, à tous les domaines, précédemment évalués sous forme papier. Les responsables de PISA devront donc s'assurer de la comparabilité des résultats avec les vagues antérieures, notamment la session 2006, alors que les modalités d'interrogation sont différentes. C'était d'ailleurs l'objet d'une étude dans le cadre de l'expérimentation de PISA 2015, qui a eu lieu en 2014, et dont les résultats ne sont pas connus au moment où le présent article est rédigé. Les enquêtes ESLC et Piaac interrogent quant à elles les élèves ou les adultes d'un même pays, soit sur informatique soit sur papier, avec le même matériel d'évaluation, en supposant que les deux modalités sont comparables.

La première question, posée dans le cadre d'ERA, est celle de l'équivalence des compétences de lecture mises en jeu selon les deux modalités d'interrogation. Outre les questions ergonomiques, les concepteurs utilisent dans ce type de test la navigation hypertextuelle, les onglets, de nouvelles formes d'items, etc. Se pose alors la question de savoir si on ne mesure pas une nouvelle compétence distincte de la lecture sur support papier.

La comparaison réalisée dans le cadre de l'évaluation ERA repose sur l'analyse de la corrélation des scores. Le coefficient de corrélation observé entre l'épreuve papier de compréhension de l'écrit de PISA et l'épreuve numérique construite dans le cadre d'ERA est élevé, d'une valeur de 0,83 sur l'ensemble des 16 pays ayant participé à ERA. Il est cependant intéressant de comparer cette corrélation avec celle observée entre la compréhension de l'écrit sous forme papier et les mathématiques et les sciences, elles aussi évaluées sous forme papier ▸ **Tableau 1**.

Tableau 1 **Corrélations entre les épreuves de PISA**

	Littératie (papier)	Littératie (électronique)
Littératie (papier)	1	
Littératie (électronique)	0,83	1
Mathématiques	0,83	0,76
Sciences	0,88	0,79

Note de lecture : le tableau indique la corrélation entre les scores des différentes épreuves pour les 16 pays de l'OCDE participant à l'expérimentation.
Source : OCDE, 2011.

Il apparaît que la corrélation entre les épreuves papier et électronique est du même ordre de grandeur que la corrélation entre la compréhension de l'écrit sous forme papier et d'autres compétences telles que les mathématiques. La lecture sur support électronique apparaît donc bien comme un domaine distinct de la lecture sur papier, au même titre que les mathématiques ou les sciences. Enfin, il faut noter que la corrélation papier/électronique varie de manière importante selon les pays : de 0,71 à 0,89. Ce dernier point peut soulever un problème de comparabilité internationale, qui doit très certainement prendre en compte la familiarité des élèves avec le support numérique.

Pour l'enquête ESLC, le rapport n'indique pas, quant à lui, de différences majeures entre les difficultés des items issus du papier et du test électronique [Commission européenne, 2012]. Les données ne sont présentées que pour l'anglais en compréhension écrite et en compréhension orale : il semble ressortir que les niveaux de difficulté observés sur les deux supports, papier et électronique, sont moins comparables en compréhension de l'écrit qu'en compréhension orale.

Au niveau national, la DEPP a développé une application « Lecture sur support électronique » (LSE) afin d'évaluer spécifiquement la lecture sur support informatique. La comparaison entre les résultats à ce test et à un test de maîtrise de la langue issu du cycle Cedre [Colmant, Daussin, Bessonneau, 2011] a abouti pour des élèves de CM2 à des corrélations entre les tests relativement basses, de l'ordre de 0,6 [Bessonneau, 2012 ; Dierendonck, 2014]. En outre, elle mettait en avant de meilleurs résultats relatifs sur le format électronique pour les garçons et pour les élèves en zone d'éducation prioritaire notamment.

Au-delà des enquêtes d'évaluation, des revues de littérature comparant les performances sur papier et sur support électronique abondent. Par exemple, la revue de Wang et Shin [2009] fait état de résultats divergents. Les tests sont parfois plus faciles, plus difficiles ou de même difficulté. Ces études sont pour la plupart anglo-saxonnes et portent fréquemment sur des épreuves comportant un certain enjeu pour les élèves, car faisant partie intégrante de leur cursus scolaire. Or, un facteur important de différence pourrait avoir trait à la motivation des élèves face à la situation d'évaluation. Des enquêtes telles que PISA ou LSE sont des évaluations sans enjeu pour les élèves de l'échantillon, et cette absence d'enjeu pourrait expliquer certains écarts observés entre les différents supports. Une autre hypothèse sur la diversité des résultats peut porter sur le fait que les études sont menées à partir de logiciels spécifiques, qui adoptent des ergonomies différentes les uns par rapport aux autres.
Les résultats obtenus à la DEPP à travers les évaluations nationales et internationales, ainsi que les résultats contrastés rapportés par la littérature scientifique,

nous ont amené à conduire des expériences spécifiques pour apprécier la comparabilité des deux supports.

Ainsi, cet article présente les résultats de deux expériences :
– la première cherche à identifier les différences de difficulté des items entre le support papier et le support numérique, à partir d'une évaluation des compétences de base en français et en mathématiques conduite en fin de primaire, et selon un plan d'expérience (*design*) original ;
– la seconde expérience concerne les résultats d'un lot d'items identiques proposés sur support numérique et sur support « papier » dans le cadre des évaluations Cedre mathématiques.

Les résultats de ces expériences mettent en avant l'interaction entre la forme des items, la charge cognitive pour l'élève et les restrictions liées à l'un ou l'autre des supports.

EXPÉRIENCE 1 : ÉVALUATION DES COMPÉTENCES DE BASE EN FRANÇAIS ET EN MATHÉMATIQUES EN FIN DE PRIMAIRE

De 2007 à 2012, pour alimenter les indicateurs de performance du système éducatif attendus par la LOLF (loi organique des lois de finances), une évaluation annuelle de la DEPP au primaire (CM2) et au collège (troisième) réalisée sur échantillon a permis d'évaluer le niveau de maîtrise des compétences de base en français et en mathématiques. La description de la création de cette épreuve a fait l'objet d'une *Note d'information* [Rocher, Chesné, Fumel, 2008].

Concernant le CM2, l'épreuve finale était composée de 75 items de français et de 68 items de mathématiques. Ces items provenaient d'une large expérimentation d'items dont la création impliquait la DEPP, des enseignants, des conseillers pédagogiques, des inspecteurs de l'éducation nationale (IEN, IA-IPR, IGEN). La qualité des résultats de cette évaluation reposait sur sa conception, sur l'évaluation de larges échantillons d'élèves et sur l'utilisation d'outils psychométriques pour l'analyse des résultats [Laveault et Grégoire, 2002].

Les indicateurs attendus par la LOLF étaient les proportions d'élèves maîtrisant les compétences de base. Les scores des élèves devaient donc être utilisés pour différencier ces deux populations en créant un score seuil départageant ces deux populations d'élèves. La détermination de ce score seuil est établie selon des procédures dites de « *standard settings* » [Bunch et Cizek, 2007]. Ce travail permet de croiser les exigences et les attentes pédagogiques avec les résultats psychométriques en vue de déterminer un score seuil faisant consensus. Depuis 2011, de nouvelles épreuves ont été conçues par la DEPP pour évaluer les compétences du socle commun, avec les mêmes soucis méthodologiques que les épreuves évaluant les compétences de base [Miconnet et Vourc'h, dans ce numéro, p. 141]. Mais afin de s'assurer notamment de la consistance des résultats, les deux évaluations ont coexisté durant quelques années.

Les deux épreuves se caractérisent par la nécessité de produire des résultats fiables et comparables dans le temps. Or, pour des exigences matérielles, ces nouvelles épreuves du socle sont confrontées à terme à la possibilité d'être dématérialisée, c'est-à-dire que les cahiers d'évaluation pourraient être remplacés par une évaluation sur support numérique. C'est pour étudier les conséquences d'une éventuelle transition du papier vers le numérique qu'une expérimentation *ad hoc* a été proposée.

Description de l'expérimentation

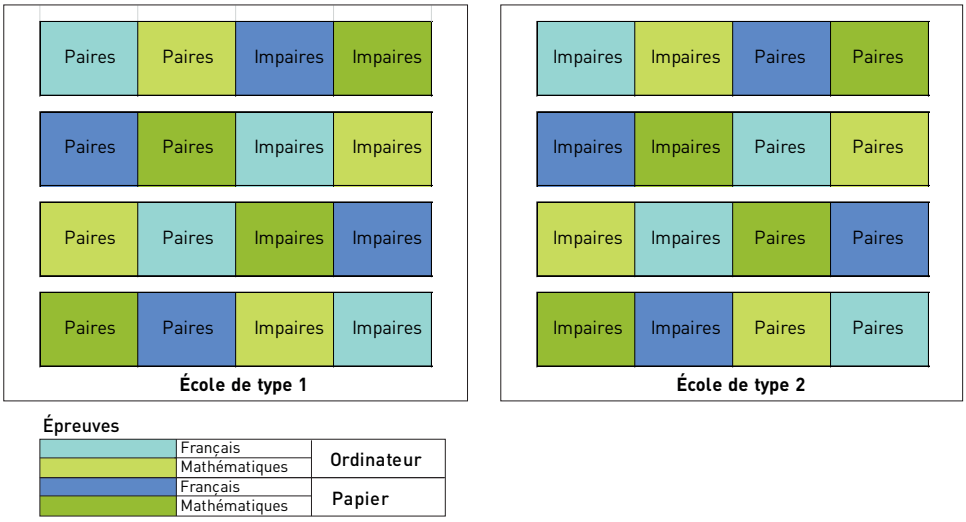
L'épreuve de compétences de base a été divisée en deux : une épreuve dite « paire » et une épreuve dite « impaire ». Les items pairs ont été inclus dans l'épreuve paire et les items impairs dans l'épreuve impaire. Cette alternance pair/impair permet de garder la position relative des items dans chaque épreuve.

Toutefois, les textes et supports longs présents dans l'épreuve n'ont pu être divisés. En effet, si on avait divisé les items en gardant les supports longs, l'épreuve aurait été :
– trop longue pour les élèves ;
– les élèves auraient retrouvé les mêmes textes en partie paire et impaire ;
– le décloisonnement des items pour chaque texte aurait nui à la comparabilité en cas de dépendance entre items.

L'élaboration du plan d'expérience (*design*) s'est appuyée sur plusieurs contraintes. Le but était de contrôler différents paramètres concernant l'ordre de passation des épreuves, à savoir le support de passation (informatique ou papier), l'épreuve (paire ou impaire) ainsi que la discipline (français ou mathématiques). Afin que les élèves ne puissent pas communiquer sur les items, la passation se fait en parallèle sur les mêmes items.
Ainsi, dans un premier type d'écoles dit « paire puis impaire » la moitié des élèves passait la première épreuve « paire » sur support informatique tandis que l'autre moitié des élèves passait l'épreuve « paire » sur papier. Chacun de ces deux groupes d'élèves était divisé en deux : l'un passant d'abord le français, l'autre passant d'abord les mathématiques.
La même organisation était appliquée dans les écoles du deuxième type dit « impaire puis paire », mais l'ordre de passation des épreuves était inversé (d'abord les épreuves impaires puis les paires).
Le *design* est donc complètement équilibré puisque chaque discipline, chaque série d'items et chaque support se retrouvent dans chaque position. Ce *design* est schématisé par la **figure 1 p. 164**.

Concernant le contenu de l'évaluation, toutes les questions sont des QCM avec 2, 3 ou 4 choix possibles. La présentation des items était presque identique : les items provenaient d'une même banque d'items et leur génération sur papier et sur électronique était automatique. L'épreuve impaire de français était composée de 36 items et celle de mathématiques de 33 items. L'épreuve paire est, quant à elle, composée de 33 items de français et de 32 items de mathématiques.
L'échantillon était composé d'écoles se répartissant en deux catégories :
– des écoles volontaires situées en province (un tiers de l'effectif) ;
– des écoles parisiennes tirées au hasard sur les écoles ne participant pas à une autre opération de la DEPP en CM2 (deux tiers de l'effectif).

Figure 1 Plan de rotation des blocs électronique et papier



Note de lecture : pour une école de type 1, quatre cas sont possibles. Dans les deux premiers cas, le français est vu avant les mathématiques alors que dans les deux derniers cas, les mathématiques sont passées en premier. Dans les cas 1 et 3, le test électronique est passé avant le test papier-crayon. Dans les cas 2 et 4, c'est l'inverse : le test papier-crayon est passé avant le test électronique.

Tous les élèves de CM2 des écoles sélectionnées participaient à l'expérience. Au total, nous avons collecté des résultats pour 44 écoles et environ 800 élèves.
Selon les résultats recueillis, il s'avère qu'une partie des écoles n'a pas respecté le plan de rotation. En cause notamment la difficulté pratique pour la mettre en place au sein d'une école. En outre, le respect des conditions de passation était parfois particulièrement difficile au regard du matériel informatique disponible. Ainsi, du point de vue du *design*, si les cahiers sont équilibrés, le respect de l'ordre de passation n'a pas été tout à fait respecté.

Des questions de « contrôle » étaient proposées sur la partie informatique telles que : « As-tu déjà passé l'épreuve papier ? », « Quel était l'identifiant de ton cahier ? », etc. En outre, il était demandé à l'élève d'indiquer son mois de naissance, son année de naissance et son sexe sur les deux supports. Le but de ces questions était de s'assurer lors de l'analyse que les résultats sur les deux supports correspondaient à un seul et même élève. Pour près de 200 élèves, les réponses à ces questions indiquaient qu'il s'agissait probablement d'un élève différent : les réponses étaient discordantes entre papier et support électronique. Ces élèves ont été supprimés pour l'analyse. L'échantillon final portait donc sur 632 élèves dans 39 écoles.

Des statistiques sur les établissements et les élèves ont été récoltées :
– nombre de filles et de garçons ;
– école de secteur privé ou école publique ;
– niveau de l'école aux épreuves nationales de CM2 en 2011 ;
– ordre de passation des épreuves.

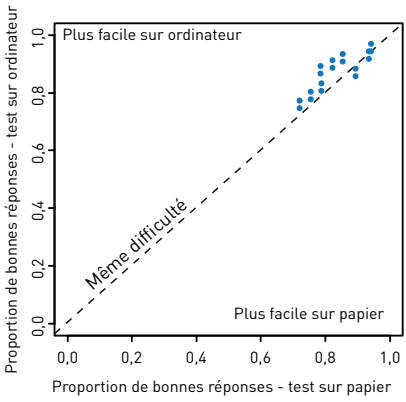
Ces statistiques ont été utilisées pour calculer des poids dans le but de redresser l'échantillon. En incluant les poids, chaque cahier présente la même répartition pour chacune des statistiques citées. Le calcul des poids a été réalisé par un calage sur marge [Sautory, 1993].

Résultats

En premier lieu, il s'agit d'identifier des différences entre la difficulté des items sur le support papier et le support numérique qui serait un indicateur important de comparabilité des deux supports. Dans un second temps, une analyse des scores selon les supports était envisagée. Toutefois étant donné la faible longueur de chacune des parties du test et la grande facilité de l'épreuve, une proportion importante d'élèves obtenait le score maximal à chaque partie du test. La distribution des scores pour plusieurs épreuves étant tronquée, l'analyse des scores n'était pas envisageable.

Comme les différents échantillons ont été équilibrés, les premières analyses portent sur la comparaison des proportions de bonnes réponses aux questions (items) entre leur version sur papier et leur version sur support électronique. En français, les proportions de réussite sont très voisines pour les items d'orthographe ▸ **Figure 2**. Inversement, des différences non négligeables apparaissent pour les items de compréhension de textes.

▸ **Figure 2** **Proportion comparée de bonnes réponses pour l'orthographe**

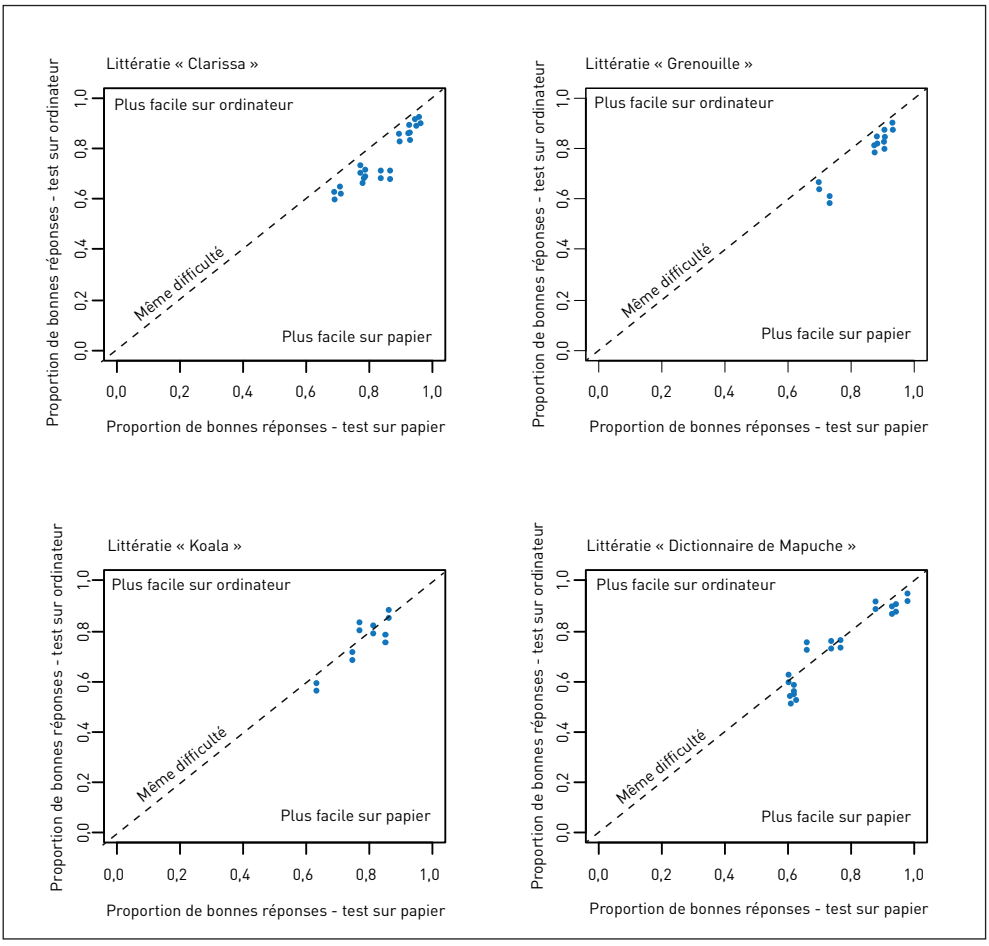


Note de lecture : chaque point représente un item. L'axe des abscisses représente le taux de réussite des items en version électronique, l'axe des ordonnées le taux de réussite des mêmes items dans leur version papier/crayon. La droite qui partage ce graphique indique un niveau de réussite identique sur les deux supports. Un éloignement de cette droite correspond soit à une réussite plus grande sur le papier (au-dessus de la droite) soit à une réussite plus grande sur le support numérique (au-dessous de la droite).

Globalement les items sont mieux réussis sur papier que sur ordinateur. Les différences sont notables pour les textes « Clarissa » et « Grenouille » tandis qu'elles sont moindres pour les textes « Koala » et « Dictionnaire de Mapuche » ▸ **Figure 3**.
Pour « Clarissa » et « Grenouille », on peut poser l'hypothèse que, les supports étant longs, le défilement du document sur ordinateur a rendu la tâche plus difficile. En outre pour « Grenouille », les questions sont en regard du texte sur la version papier, pas sur la version informatisée. Ce qui a conduit à diminuer la difficulté des items sur papier.
Pour « Koala », l'hypothèse inverse peut être formulée. Le texte assez court ne nécessite pas de manipulation supplémentaire (ascenseur vertical) pour lire le texte dans son intégralité, d'où une différence de difficulté moindre selon le support.

Pour le « dictionnaire de Mapuche », le document est long mais, comme pour « Grenouille », les questions se retrouvant en regard du support, elles sont plus faciles sur papier.
De manière générale, il apparaît une cohérence globale en termes de hiérarchie de difficulté des items, à l'exception de quelques items, notamment pour le texte « Clarissa », dont l'écart de difficultés entre les deux versions est plus prononcé que les autres items.

Figure 3 Proportions comparées de bonnes réponses pour les textes

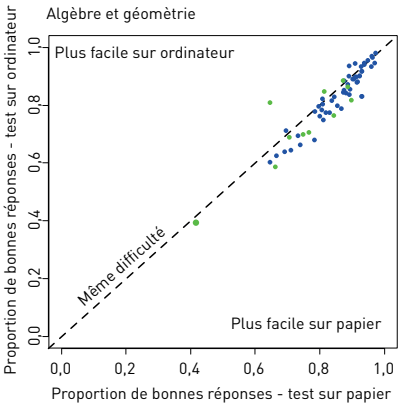


Note de lecture : chaque point représente un item. Un éloignement de la droite qui partage le graphique correspond soit à une réussite plus grande sur le papier (au-dessous de la droite) soit à une réussite plus grande sur le support numérique (au-dessus de la droite).

Pour les mathématiques, à part pour les questions les mieux réussies, l'épreuve apparaît plus facile sur papier que sur ordinateur ▸ Figure 4.
Nous notons une exception toutefois : un item est largement plus facile, car il s'agissait d'identifier un parallélépipède et le redimensionnement de l'image sur ordinateur rendait la tâche plus aisée.

Pour les mathématiques, une recherche a été menée sur certaines questions difficiles à réaliser sans brouillon. Nous avons recherché l'utilisation du cahier comme brouillon, mais sur les images des cahiers numérisés, aucune trace d'écriture n'était visible.

Figure 4 Proportions comparées de bonnes réponses pour les items de mathématiques



Note de lecture : chaque point représente un item. L'axe des abscisses représente le taux de réussite des items en version électronique, l'axe des ordonnées le taux de réussite des mêmes items dans leur version papier/crayon. La droite qui partage ce graphique indique un niveau de réussite identique sur les deux supports. Un éloignement de cette droite correspond soit à une réussite plus grande sur le papier (au-dessous de la droite) soit à une réussite plus grande sur le support numérique (au-dessus de la droite).

Conclusion de l'expérience 1

Les différences de difficulté des items semblent dépendre de la nature du support, du « contexte » de l'exercice sur le papier et de la nature des tâches.
Plus que des différences entre les supports, les différences entre numérique et papier mettent en lumière des différences qui pourraient exister entre différentes mises en page papier : la proximité du texte avec les questions, sa taille, etc. La version numérique y apparaît comme une mise en page de l'item parmi d'autres.
En mathématiques comme en orthographe, peu de différences entre les taux de réussite selon les supports sont observées. Toutefois cela est à relativiser car les supports de ces items sont tous de faible taille.

EXPÉRIENCE 2 : LES ACQUIS DES ÉLÈVES DANS LE CADRE DE L'ÉVALUATION CEDRE MATHÉMATIQUES

Les enquêtes du cycle d'évaluations disciplinaires réalisées sur échantillons (Cedre) sont réalisées tous les ans sur une discipline différente. En 2013, en vue de l'évaluation de 2014, des items de mathématiques ont été expérimentés en fin de collège et en fin de CM2 afin d'écarter ceux apportant peu d'informations sur la compétence des élèves. Cette sélection porte sur une analyse psychométrique et pédagogique [Rocher, dans ce numéro, p. 37].
Les techniques de l'information et de la communication viennent modifier aussi bien les pratiques pédagogiques que les compétences devant être mises en œuvre par les élèves.

Il est important d'observer ce que ces techniques numériques garantissent comme continuité par rapport aux techniques traditionnelles ; ce qu'elles proposent comme nouveautés et ce qu'elles ne peuvent pas remplacer.

Dans le cadre de l'évaluation des élèves et plus spécifiquement du protocole Cedre qui a pour but d'observer les évolutions temporelles, il est primordial de savoir si le passage d'une évaluation papier-crayon à une évaluation numérique permet d'assurer une mesure comparable. Comme pour la première expérience, nous avons établi un protocole de « dématérialisation », qui consiste à reproduire à l'identique un item sur le support « papier » et sur le support « numérique ». Les élèves de l'échantillon passent les items en partie sous une forme ou en partie sous l'autre.

Cette deuxième expérience a pour objectif d'analyser les différences entre les deux formats selon plusieurs critères liés au domaine des mathématiques :
– les champs mathématiques ;
– les compétences mises en jeu ;
– les facteurs de complexités liés à l'énoncé, à la connaissance mathématique et à la tâche demandée à l'élève.

L'analyse des résultats, à travers les taux de réussite observés selon le support, doit nous conduire à répondre aux questions suivantes :
– si des écarts sont constatés en fonction d'un support, peut-on quantifier ces écarts ?
– l'écart est-il toujours dans le même sens ?
– quelles sont les pistes explicatives ?

Notons tout d'abord qu'il existe des différences « intrinsèques » entre les deux modes :
– les écrans, même en haute résolution, n'offrent pas le confort de lecture du support écrit ; ils introduisent parfois des déformations. Ceci peut s'avérer un biais important pour les items de géométrie. L'écran impose une lecture verticale moins confortable que celle proposée par la lecture d'un texte posé sur un bureau ;
– la structure d'une page numérisée d'information n'utilise pas les mêmes référents que la page imprimée. Sur un support numérique, la segmentation de l'information est spécifique : les pages sont plus courtes, les repères graphiques multiples, les liens hypertextes renvoient à de nouveaux contenus. Tous les éléments ne sont donc pas accessibles d'un seul coup d'œil par le lecteur. Les pages ne sont pas numérotées, mais plutôt disposées en réseau ;
– des dispositifs particuliers permettent l'accès à ces pages (liens, fenêtres multiples, onglets, etc.). Le lecteur doit maîtriser ces structures spécifiques. Sur support numérique, l'accès aux contenus demande à l'élève de s'approprier le mode de navigation et les actions spécifiques de déplacements entre les pages numériques ;
– à la capacité de lecture-compréhension du message s'ajoute celle de l'accès à l'information. On ne déploiera pas les mêmes procédures pour accéder à un contenu donné dans un livre ou dans un site Web. Dans certains cas la transposition d'un support « papier » à un support « numérique » n'est pas simple.
Ces différences impliquent très certainement une variabilité en termes de procédures de résolution, de stratégies, de processus cognitifs. Cela peut expliquer que les taux de réussite des items soient dépendants du mode d'interrogation.
Dans le cadre de l'évaluation Cedre, nous trouvons cette différence. Chaque unité de

l'évaluation correspond à un ensemble d'exercices sur un même thème. Les exercices sont disposés « classiquement » sur le support papier. L'élève peut appréhender l'ensemble de la situation d'un seul coup d'œil. Sur le support numérique, nous avons procédé à une segmentation de l'unité qui donne à voir à l'élève un seul exercice à la fois.

Première partie : Cedre école

Lors de l'expérimentation de 2013, les élèves ont été évalués à la fois sur support papier et sur support numérique en mathématiques. L'échantillon comptait un peu plus de 4 500 élèves de CM2 répartis dans 172 écoles. Les réponses de 3 841 élèves répartis dans 149 écoles ont été analysées pour la partie papier et les réponses de 2 575 élèves répartis dans 118 écoles pour la partie numérique. Le faible pourcentage de retour de l'enquête numérique s'explique par la grande disparité et la quantité de matériel disponible dans les écoles, ce qui constitue aujourd'hui une première limite importante en matière de comparabilité des deux modes d'interrogation[1].

Le matériel d'évaluation était constitué de six cahiers pour les items en version papier-crayon et de douze modules pour les items en version numérique. Parmi l'ensemble des items expérimentés, 56 items ont été évalués sur les deux supports (papier et numérique). Ils recoupent les champs mathématiques de la connaissance des nombres entiers naturels, des nombres décimaux, de la géométrie, des grandeurs et mesures, de l'organisation et gestion de données.

Ces items étaient répartis dans les six cahiers et dans les douze modules numériques. Chaque élève se voyait attribuer, de manière aléatoire, un des six cahiers et un des douze modules. La correspondance entre cahiers et modules a été élaborée de manière à ce qu'un élève ayant répondu à un item sur le support « papier » ne retrouve pas le même item sur le support « numérique ».
L'évaluation est séquencée par une présentation de l'activité conduite par l'enseignant ; une phase d'entraînement et l'évaluation de mathématiques que l'élève réalise en complète autonomie.

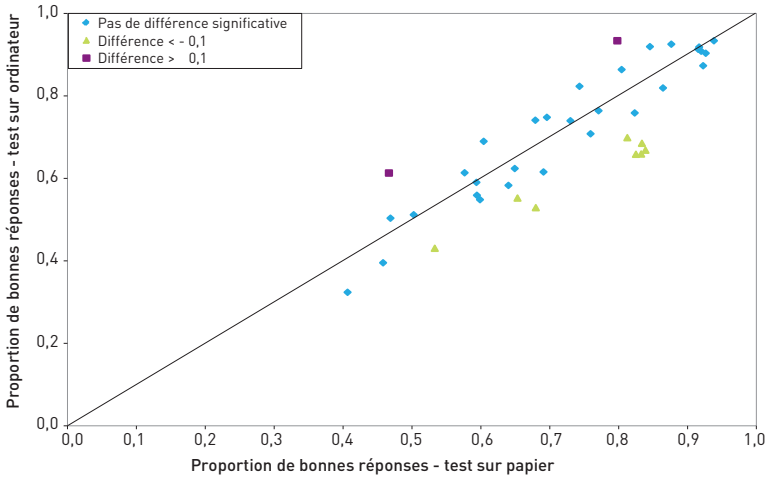
Sur le support papier, l'élève doit appréhender les formats de questions ; sur le support numérique s'ajoute le nécessaire apprentissage de la navigation.
À la fin de l'évaluation, les cahiers nous sont retournés pour le traitement des résultats ; pour la partie numérique, les résultats sont directement sauvegardés dans la base de données de la DEPP et immédiatement consultables.

Analyse globale

Trente-neuf items présentant des qualités psychométriques satisfaisantes ont été retenus pour l'analyse. Pour chaque item, le taux de réussite sur les deux supports a été calculé ▸ **Figure 5 p.170**.

1. Nous observons tous types de systèmes d'exploitation (Windows, MacOs, Linux) ou d'architectures (monoposte, réseau, terminaux, netbooks, tablettes). Les ordinateurs sont regroupés en salle informatique ou répartis dans les classes de l'école. Nous constatons l'impossibilité de faire une passation lorsqu'il y a trop peu d'ordinateurs (un ordinateur en fond de classe par exemple).

Figure 5 Taux de réussite aux items selon le support



Note de lecture : chaque point correspond à un item.
L'axe des abscisses représente le taux de réussite des items en version papier-crayon, l'axe des ordonnées le taux de réussite des mêmes items dans leur version numérique. La droite qui partage ce graphique indique un niveau de réussite identique sur les deux supports. Un éloignement de cette droite correspond soit à une réussite plus grande sur le papier (au-dessous de la droite) soit à une réussite plus grande sur le support numérique (au-dessus de la droite).

Une première approche montre que :
– 8 items sont mieux réussis sur le support papier. Ces items comportent des textes (consignes ou propositions de réponses) longs et nécessitant une lecture fine. Parfois, l'élève utilise la possibilité d'écrire sur le document pour marquer des repères, faire des annotations ou utiliser un outil de mesure lui permettant de vérifier une longueur, par exemple ;
– 2 items sont mieux réussis sur le support numérique. Ces items comportent des textes courts. La réponse portée est facilitée par un bouton radio (la réponse ne peut être vraie et fausse à la fois). Les graphiques semblent mieux mis en valeur sur ce support, mais les performances dépendent de la tâche demandée : lecture directe du graphique, sans de multiples prises d'indices ;
– 29 items ne présentant pas d'écart de réussite en fonction du support. Il est difficile de dégager des caractéristiques saillantes communes à ces items. Cependant, il nous a semblé pertinent de les regrouper selon quatre catégories liées au niveau de lecture ainsi qu'au graphisme :
T1 – item dont les consignes et/ou les propositions impliquent une lecture directe sans difficulté majeure ▸ **Figure 6** ;
T2 – item dont les consignes et/ou les propositions impliquent une lecture fine et attentive. La réponse n'est pas directe, il y a une nécessaire appropriation du message avant de porter la réponse ▸ **Figure 7** ;
T3 – item constitué d'un graphisme (ce peut-être une courbe, un organigramme, un tracé géométrique, etc.). La tâche de l'élève consiste à observer ce graphisme et d'en déduire directement la réponse ▸ **Figure 8** ;

Figure 6 Item dont les consignes et/ou les propositions impliquent une lecture directe sans difficulté majeure.

Pour chaque nombre de cette liste, indique s'il est multiple de deux :

	Vrai	Faux
5	☐ 1	☐ 2
14	☐ 1	☐ 2
25	☐ 1	☐ 2
40	☐ 1	☐ 2
33	☐ 1	☐ 2
124	☐ 1	☐ 2
250	☐ 1	☐ 2

Figure 7 Item dont les consignes et/ou les propositions impliquent une lecture fine et attentive. La réponse n'est pas directe, il y a une nécessaire appropriation du message avant de porter la réponse.

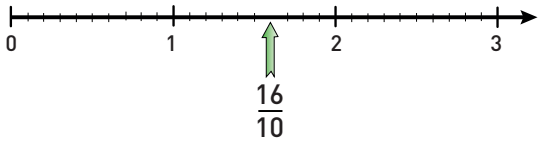
		Vrai	Faux
1	Le quotient est le résultat d'une addition	1 ☐	2 ☐
2	Le produit est le résultat d'une division	1 ☐	2 ☐
3	Le quotient est le résultat d'une division	1 ☐	2 ☐
4	Le produit est le résultat d'une multiplication	1 ☐	2 ☐

T4 – item constitué d'un graphisme. La tâche de l'élève consiste à relever plusieurs éléments avant de pouvoir porter sa réponse **Figure 9**.

Cette typologie montre que :
– les items proposant une lecture directe (présentés sous formes de losanges) sont systématiquement mieux réussis sur un support numérique **Figure 10**. Pour décrire la tâche demandée à l'élève, nous pourrions dire : l'élève est focalisé sur la lecture à l'écran, il clique avec sa souris, il n'a pas besoin d'étape intermédiaire[2] ;

2. Étape intermédiaire : l'élève doit utiliser un instrument, ou il doit prendre des notes, ou il doit mémoriser des éléments (faire des inférences).

Figure 8 **Item constitué d'un graphisme. La tâche de l'élève consiste à observer ce graphisme et d'en déduire directement la réponse**

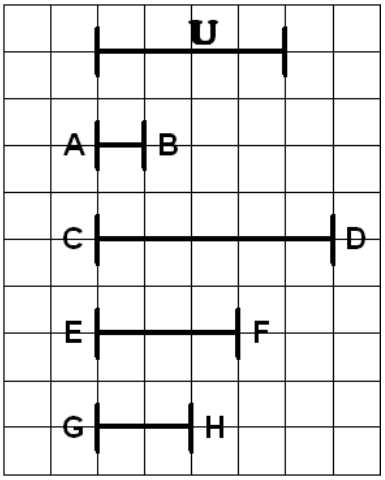


1 ☐ Oui

2 ☐ Non

Figure 9 **Item constitué d'un graphisme. La tâche de l'élève consiste à relever plusieurs éléments avant de pouvoir porter sa réponse**

Observe les segments.



Observation 1

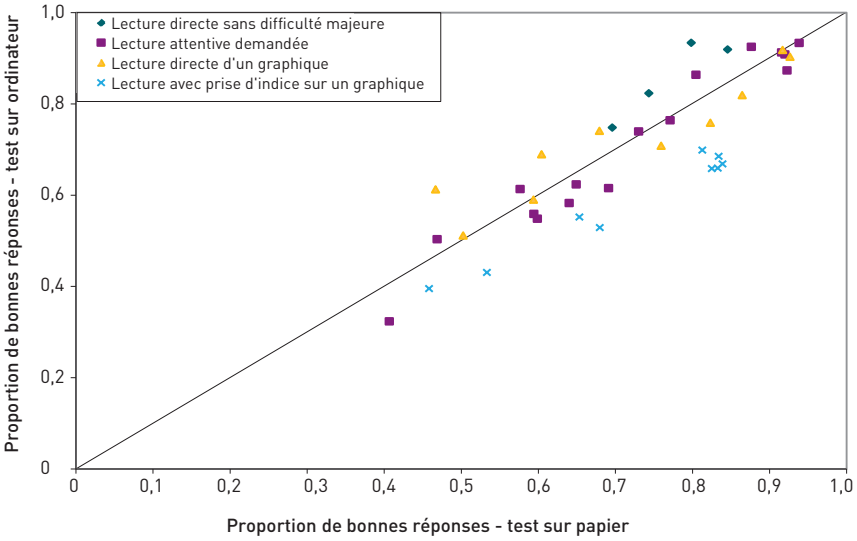
La longueur du segment [AB] est :

1	☐	$\frac{1}{2}U$
2	☐	$\frac{4}{1}U$
3	☐	$\frac{1}{8}U$
4	☐	$\frac{1}{4}U$

– à l'opposé, les items proposant une lecture avec prise d'indice sur un graphique (croix bleue) sont systématiquement mieux réussis sur un support papier. La tâche demandée à l'élève implique une étape intermédiaire ;
– entre ces deux pôles, nous trouvons les items réclamant une lecture attentive (carrés) et la lecture directe d'un graphique (triangles). Ces items se répartissent de part et d'autre de la droite indiquant une réussite identique entre les deux supports.

Il est sans doute difficile d'être catégorique après une étude ne reposant que sur 39 items. Néanmoins, nous pouvons dégager des pistes explicatives et répondre partiellement aux questions initialement posées. Concernant l'écart de difficulté entre les deux supports, il n'est pas possible d'y répondre simplement en faveur

Figure 10 Comparaison des niveaux de difficulté selon le type d'item



Note de lecture : chaque point correspond à un item.
L'axe des abscisses représente le taux de réussite des items en version papier-crayon, l'axe des ordonnées le taux de réussite des mêmes items dans leur version numérique. La droite qui partage ce graphique indique un niveau de réussite identique sur les deux supports. Un éloignement de cette droite correspond soit à une réussite plus grande sur le papier (au-dessous de la droite) soit à une réussite plus grande sur le support numérique (au-dessus de la droite).

de l'un ou de l'autre médium. En revanche, il est possible de caractériser les items en fonction de leur difficulté sur l'un ou l'autre de ces supports.
Pour rappel, la difficulté mesurée de l'item est la résultante de l'énoncé, de la connaissance mathématique mise en jeu et de la tâche à accomplir. Lorsque l'on propose des items qui réclament une lecture longue, de multiples inférences ou présentant une structure syntaxique riche, ils sont mieux réussis sur le support papier. Lorsque l'on propose un graphisme en tant que document, dans la mesure où la consigne concerne une prise d'indice directe, le résultat sur support numérique est meilleur. En d'autres termes, si l'élève peut agir directement et rapidement sans recourir à des outils ou à des étapes intermédiaires, le support numérique lui est favorable ; à l'opposé, s'il doit utiliser un instrument de mesure ou procéder à un brouillon, c'est le support papier qui est le plus efficace.

Seconde partie : Cedre collège

Lors de l'expérimentation de 2013, les élèves de troisième échantillonnés ont été évalués à la fois sur support papier et sur support numérique. L'échantillon comptait un peu plus de 5 000 élèves de troisième générale répartis dans 199 classes. Les réponses de 4 176 élèves répartis dans 194 collèges ont été analysées pour la partie papier ainsi que celles de 3 204 élèves répartis dans 148 classes pour la partie numérique.
Le matériel d'évaluation était constitué de huit cahiers pour les items en version papier-crayon et de huit modules pour les items en version numérique. Parmi l'ensemble des items expérimentés, cent trente-neuf items ont été évalués sur les

deux supports (papier et numérique). Chaque élève se voyait attribuer de manière aléatoire un des huit cahiers et un des huit modules. La correspondance entre cahiers et modules a été élaborée de manière à ce qu'un élève ayant répondu à un item sur le support «papier» ne retrouve pas le même item sur le support « numérique ».

Analyse globale

En fin de collège, l'analyse a porté sur 109 items dont les qualités psychométriques étaient satisfaisantes. Pour chaque item, le taux de réussite sur les deux supports a été calculé. La **figure 11** montre de manière évidente que les items sont nettement mieux réussis sur support « papier » que sur support « numérique ».

L'analyse détaillée des résultats par items tend à montrer que :

- l'item est moins bien réussi sur support numérique lorsque le type de tâche demandé relève d'une ou plusieurs des catégories suivantes :
  – il induit un raisonnement à plus d'une étape ;
  – il nécessite le recours à une schématisation de la situation ;
  – il nécessite le recours à des instruments de mesure ;
  – il nécessite des calculs intermédiaires ;
- l'item a des taux de réussite équivalents sur les deux supports lorsque le type de tâche demandé relève du calcul automatisé. Ce constat ouvre la voie à la création d'items d'activités mentales dans le cadre d'une éventuelle évaluation adaptative afin d'utiliser la plus-value apportée par l'environnement numérique ;
- l'item est mieux réussi sur support numérique lorsque le type de tâche demandé relève d'une méthode d'apprentissage.

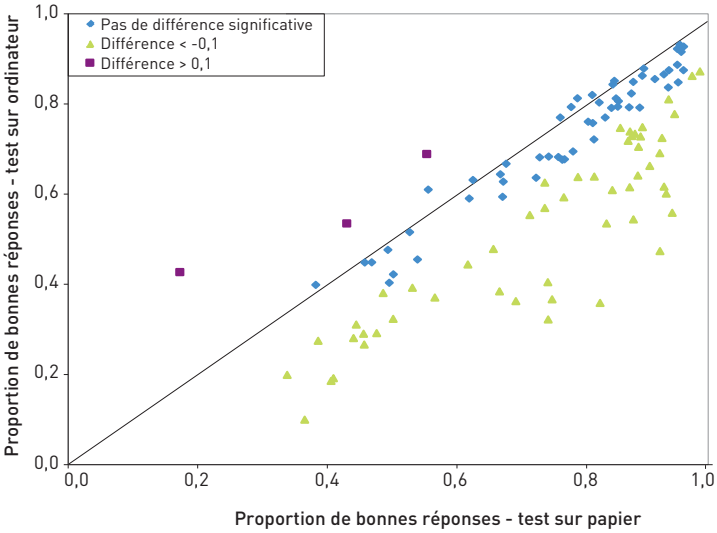
Nous illustrons des premiers constats en présentant un exemple d'item, proposé dans l'évaluation, par type de tâche ▸ **Figure 12**.

Sur support papier, l'élève a la possibilité d'exclure certaines figures lors de son raisonnement en les barrant par exemple, tandis que sur support « numérique », il est contraint de garder en mémoire l'ensemble des étapes de sa procédure de résolution. Ainsi pour cet item la différence de réussite est très importante : 74 % sur support papier et 32 % sur support numérique.

Dans l'exemple suivant, le constat est encore plus marqué ▸ **Figure 13**. Sur support « papier », l'élève peut colorier, barrer, cocher. Le support « numérique » contraint l'élève à repérer la case B3 pour chaque nouvelle question. Il amorce son raisonnement à partir des questions posées et teste les solutions proposées. Pour cet item également la différence de réussite est très importante : 91 % sur support papier et 47 % sur support numérique.

Sur support « papier », il est possible de faire une représentation imagée du problème (représenter les sacs puis chaque bille dans le sac par un point, etc.) ▸ **Figure 14**. Sur support « numérique », le raisonnement reste mental. On observe des différences de taux de réussite importantes entre les supports : 87 % sur papier et 54 % sur numérique.

Figure 11 Comparaison des taux de réussite selon le support (papier/ordinateur) et selon le type d'item



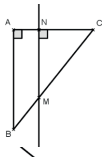
Note de lecture : chaque point correspond à un item. L'axe des abscisses présente le taux de réussite sur le support « papier », l'axe des ordonnées, le taux de réussite sur le support « numérique ». La diagonale tracée indique un taux identique sur les deux supports. Plus un point (un item) se rapproche de cette droite, plus les taux sont proches sur les deux supports. À l'inverse, un éloignement correspond à un taux plus important sur support « numérique » (au-dessus de la droite) ou à un taux plus important sur support « papier » (au-dessous de la droite).

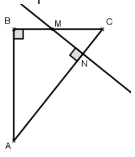
Figure 12 Item dont le type de tâche induit un raisonnement à plus d'une étape

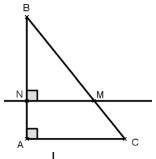
On donne le programme de construction suivant :

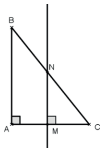
Tracer un triangle ABC rectangle en A.
Placer un point M sur le segment [BC]
Tracer la perpendiculaire à la droite (AC) passant par M.
Noter N son point d'intersection avec le segment [AC].

Parmi les constructions suivantes, cocher celle qui correspond à l'énoncé ci-dessus.

1 ☐ 

2 ☐ 

3 ☐ 

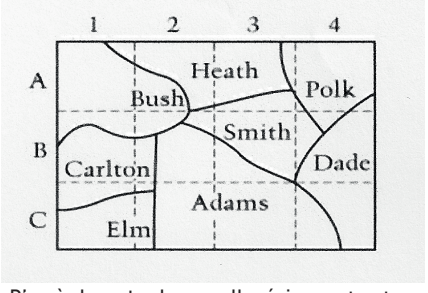
4 ☐ 

Certaines questions supposent une rotation de la figure (d1 et d5 par exemple) afin de trouver une réponse ▸ **Figure 15**. Ceci est impossible sur support numérique.
Il est également difficile de poser une équerre sur l'écran alors que les outils de géométrie sont autorisés. Il n'est pas non plus possible de prolonger les droites dans le cadre de la recherche du parallélisme. Ce type d'items au contenu en phase avec les programmes officiels ne peut être testé sans avoir recours à un logiciel de géométrie dynamique. Pour cet item, le taux de réussite sur support papier est de 93 % et sur support numérique de 56 %. Sur support papier, les élèves prennent un brouillon pour effectuer les calculs intermédiaires. Sur support numérique, les élèves sont moins enclins à utiliser le brouillon de par l'environnement numérique et une certaine partie de l'activité de l'élève consiste à manipuler la souris. L'effort de mémorisation est donc plus important et source d'erreur. Pour l'item de la **figure 16**, le taux de réussite sur support papier est de 94 % et sur support numérique de 78 %. Le type de tâche de la **figure 17** correspondant à l'étude de la représentation graphique d'une fonction est régulièrement étudié en classe en utilisant le support numérique. Les élèves le retrouvent donc dans son environnement habituel. Les taux de réussite sont de 43 % sur support numérique et de 17 % sur support papier.

Bien que les items soient moins bien réussis sur support numérique, il ne faut pas en déduire la nécessité d'évaluer uniquement sur support papier. L'évaluation sur support numérique en mathématiques serait pertinente à condition de prendre en compte l'environnement numérique dans la construction des items, de mettre à disposition des élèves des outils (tableur, logiciel de géométrie dynamique, grapheur, etc.) pour résoudre des problèmes selon les méthodes préconisées par les programmes officiels et de les utiliser dans le contexte habituel de la classe.

▸ **Figure 13** **Second item dont le type de tâche induit un raisonnement à plus d'une étape**

Sur la carte ci-dessous sont indiquées 8 régions.
La plus grande ville sur la carte se situe dans la case B3.



D'après la carte, dans quelle région peut se trouver cette ville ?
Cocher la bonne réponse.

1 ☐ Adams ou Carlton
2 ☐ Adams ou Smith
3 ☐ Carlton ou Elm
4 ☐ Dade ou Polk
5 ☐ Polk ou Smith

Figure 14 **Item dont le type de tâche nécessite le recours à une schématisation de la situation**

On dispose de trois sacs de tailles différentes :

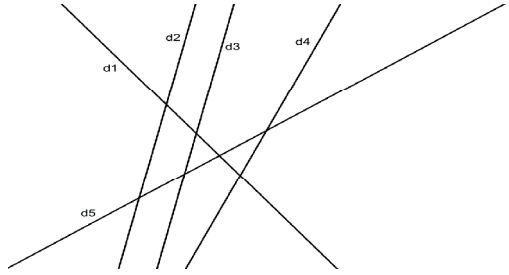
- Le plus petit sac contient 5 billes,
- Le sac de taille moyenne contient 50 billes,
- Le grand sac contient 500 billes.

Dans chaque sac, il n'y a qu'une seule bille noire. Sans regarder et au hasard, on prend une bille de chacun des sacs.
Quel sac doit-on choisir pour avoir le plus de chance de tirer une bille noire ?

1 ☐ **Le sac contenant 5 billes.**
2 ☐ **Le sac contenant 50 billes.**
3 ☐ **Le sac contenant 500 billes.**
4 ☐ **Il n'y a aucune différence.**

Figure 15 **Item dont le type de tâche nécessite le recours à des instruments de mesure**

On donne la figure suivante :



Pour chaque ligne du tableau, cocher la bonne réponse.

		Parallèles	Sécantes mais non perpendiculaires	Perpendiculaires
1	d1 et d2 semblent	☐ 1	☐ 2	☐ 3
2	d2 et d3 semblent	☐ 1	☐ 2	☐ 3
3	d4 et d5 semblent	☐ 1	☐ 2	☐ 3
4	d3 et d4 semblent	☐ 1	☐ 2	☐ 3

Conclusion de l'expérience 2

À l'issue de cette étude, les items semblent se répartir selon trois grandes catégories :
– les items « dématérialisables » ;
– les items spécifiques au support papier ;
– les items spécifiques au support numérique.

Lorsqu'un item permet une réponse directe, les taux de réussite constatés sont identiques sur les deux supports.

Les items spécifiques sur support papier sont ceux qui font appel à un raisonnement « papier-crayon » avec les outils associés (tracés géométriques, mesures, schématisation d'une situation, etc.).
Les items spécifiques sur support numérique sont ceux qui mettent en œuvre les fonctionnalités des outils numériques (géométrie dynamique, calculs avec tableur, utilisation d'un grapheur, etc.).

Les items spécifiques à chacun des deux supports peuvent être complémentaires pour un même objet d'étude. Par exemple, construire une médiatrice à la règle et au compas sur support papier ne mobilise pas les mêmes connaissances que construire la même médiatrice à l'aide d'un logiciel de géométrie dynamique.

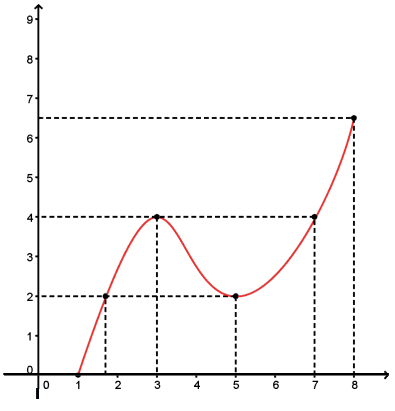
Figure 16 **Item dont le type de tâche nécessite des calculs intermédiaires**

Manon pense à un nombre, elle le double, puis ajoute 10. Elle trouve 60.
Le nombre auquel Manon a pensé est...

1 ☐ 20
2 ☐ 25
3 ☐ 35
4 ☐ 140

Figure 17 **Item dont le type de tâche demandé relève d'une méthode d'apprentissage**

On a représenté ci-dessous la courbe représentative d'une fonction *f* définie pour tous les nombres compris entre 1 et 8.



	Vrai	Faux
1 a pour image 0 par la fonction *f*.	☐ 1	☐ 2
7 est un antécédent de 4 par la fonction *f*.	☐ 1	☐ 2
3 est un antécédent de 4 par la fonction *f*.	☐ 1	☐ 2
$f(3) = 4$	☐ 1	☐ 2
$f(2) = 5$	☐ 1	☐ 2

CONCLUSION

Cette expérience permet de mettre en évidence que la transition entre support « papier » et support « numérique » n'est pas sans conséquence.
Trois variables influent particulièrement sur la réussite aux items :
– la structure de l'item (la longueur des textes proposés, le nombre de documents, le type de documents, la mise en page et l'ergonomie intrinsèque) ;
– le type de tâches mises en jeu (raisonnement nécessitant des étapes intermédiaires et capacité à « naviguer » dans le support numérique) ;
– les contraintes liées à la spécificité du support (utilisation d'outils différents : le brouillon, le tableur, le grapheur, etc.).
Dans l'optique d'une évaluation des acquis des élèves, il est nécessaire de prendre en compte les critères mis en évidence dans cette étude.

BIBLIOGRAPHIE

Bessonneau P., 2012, *Évaluation de la compréhension de l'écrit sur support informatique et comparaison avec des épreuves de type papier-crayon*, 24e colloque de l'Admée Europe, Luxembourg.

Bunch M. B., Cizek G. J., 2007, *Standard Settings,* Londres, Sage Publications, 352 p.

Colmant M., Daussin J.-M., Bessonneau P., 2011, « Compréhension de l'écrit en fin d'école, Évolution de 2003 à 2009 », *Note d'information,* n° 11.16, MENJVA-DEPP.

Commission européenne, 2012, *First European Survey on Language Competences – Technical Report.*

Dierendonck C., Loarer E., Rey B., 2014, *L'évaluation des compétences en milieu scolaire et en milieu professionnel*, Bruxelles, De Boeck, 359 p.

Garcia E., Krop J., 2013, « Cedre 2012 histoire-géographie et éducation civique : baisse des acquis des élèves de fin de collège depuis six ans », *Note d'information*, n° 13.11, MEN-DEPP.

Laveault D., Grégoire J., 2002, *Introduction aux théories des tests en psychologie et en sciences de l'éducation*, Bruxelles, De Boeck, 336 p.

OCDE, 2011, *Résultats du PISA 2009 : Élèves en ligne – Technologies numériques et performance*, vol. 6, PISA, Éditions OCDE.

Rocher T., Chesné J.-F., Fumel S., 2008, « Méthodologie de l'évaluation des compétences de base en français et en mathématiques en fin d'école et en fin de collège », *Note d'information*, n° 08.37, MEN-DEPP.

Sautory O., 1993, *La macro CALMAR – Redressement d'un échantillon par calage sur marges*, Paris, Insee.

Wang H., Shin C. D., 2009, "Computer-Based and Paper-Pencil Test Comparability Studies", *Test, Measurement and Research Services Bulletin*, No. 9, Pearson Education.

Analyses et résultats des évaluations

LES COMPÉTENCES DES ÉLÈVES FRANÇAIS EN ANGLAIS EN FIN D'ÉCOLE ET EN FIN DE COLLÈGE

Quelles évolutions de 2004 à 2010 ?

Sylvie Beuzon, Émilie Garcia
et Corinne Marchois
MENESR-DEPP, bureau de l'évaluation des élèves

Les évaluations Cedre de 2004 et de 2010 en anglais nous ont permis de mesurer l'évolution des acquis des élèves de fin de CM2 et de fin de troisième à six ans d'intervalle. Ces évaluations, portant sur trois des cinq activités langagières – compréhension de l'oral, compréhension de l'écrit et expression écrite – montrent que les résultats des élèves de fin de CM2 sont en hausse significative, alors que ceux des élèves de fin de troisième affichent une tendance inverse. Après avoir rappelé le contenu de ces évaluations, nous en présenterons les résultats et formulerons des hypothèses pour expliquer l'évolution des performances des élèves entre 2004 et 2010. À l'école, il semble que trois facteurs aient joué un rôle prépondérant : les efforts de formation continue pour les enseignants d'une part, un contact plus fréquent des élèves avec la langue anglaise en dehors du cadre scolaire d'autre part et enfin la volonté grandissante des parents de voir leur enfant mieux maîtriser l'anglais. Au collège, la baisse des résultats est davantage multifactorielle. À peine deux élèves sur dix ont une perception positive de leurs performances en anglais ; ils ne se sentent pas encouragés par leurs enseignants. Les nouvelles technologies sont encore trop rarement utilisées en classe et contrairement à l'école, l'exposition à la langue en dehors du cadre scolaire reste trop faible.

Le niveau des élèves en langues étrangères, et plus particulièrement en anglais, est un sujet de préoccupation croissante tant des pouvoirs publics et des parents que des médias.

Mesurer le niveau des élèves est l'une des missions de la direction de l'évaluation, de la prospective et de la performance (DEPP). À cette fin, elle a mis en place depuis 2003 un cycle d'évaluations disciplinaires réalisées sur échantillons (Cedre) [Rocher, dans ce numéro, p. 37 ; Trosseille et Rocher, dans ce numéro, p. 15].

Dans cet article, nous avons choisi de nous intéresser plus particulièrement à l'évolution temporelle des résultats des élèves en anglais, en fin de CM2 et en fin de troisième issus de cette enquête. Nous présenterons tout d'abord les principaux résultats en compréhension de l'oral et de l'écrit, à l'école puis au collège. Nous dégagerons ensuite les tendances prégnantes de l'évolution des acquis des élèves, éclairées d'exemples de situations d'évaluation puis tenterons d'analyser les raisons qui peuvent expliquer cette évolution. Enfin, toujours à partir de l'enquête Cedre, nous présenterons une méthode permettant d'estimer le pourcentage d'élèves qui maîtrisent le niveau A1 en fin de CM2 et le niveau A2 en fin de troisième, en référence au cadre européen commun de référence pour les langues (CECRL) du Conseil de l'Europe [2001] ▸ **Encadré**.

Le cadre européen commun de référence pour les langues (CECRL) est le fruit de plusieurs années de recherche linguistique menée par des experts des états membres du Conseil de l'Europe.
Il constituait en 2001, date de sa publication, une approche innovante ayant pour but de repenser les objectifs et les méthodes d'enseignement des langues et, surtout, de fournir une base commune pour la conception de programmes, de diplômes et de certificats. En ce sens, il avait également pour vocation de favoriser la mobilité éducative et professionnelle.
Le CECRL est fondé sur une approche dite « actionnelle » : l'usage de la langue n'est pas dissocié des actions accomplies par celui qui est à la fois locuteur et acteur social.
Le cadre introduit **six niveaux de compétence** (de A1 à C2) :

– **niveau A :** utilisateur élémentaire (= scolarité obligatoire), lui-même subdivisé en niveau introductif ou de découverte (A1) et intermédiaire ou usuel (A2) ;
– **niveau B :** utilisateur indépendant (= lycée), subdivisé en niveau seuil (B1) et avancé ou indépendant (B2) ;
– **niveau C :** utilisateur expérimenté, subdivisé en C1 (autonome) et C2 (maîtrise).

Le cadre découpe la compétence communicative en activités langagières :
– **la réception :** écouter, lire ;
– **la production :** s'exprimer oralement en continu, écrire ;
– **l'interaction :** prendre part à une conversation ;
– **la médiation** (notamment activités de traduction et d'interprétation).

CEDRE ET L'ÉVALUATION EN LANGUES VIVANTES

Le dispositif national d'évaluation Cedre permet, d'une part, de dresser un état des lieux des compétences des élèves au regard des programmes et, d'autre part, d'en mesurer l'évolution dans le temps. Pour ce faire, les élèves sont positionnés à partir de leur score sur des échelles de performances découpées en cinq ou six niveaux de compétences hiérarchisés ; des questionnaires « de contexte » destinés aux élèves, aux enseignants, aux directeurs d'école et chefs d'établissement ont pour but d'éclairer les résultats bruts des élèves en les rapprochant des caractéristiques personnelles et d'environnement. En ce qui concerne les langues vivantes, la première évaluation a eu lieu en 2004, la deuxième en 2010 ; cette dernière reprend un certain nombre de situations d'évaluation de 2004, ce qui permet, pour la première fois, de mesurer l'évolution des performances des élèves.

L'évaluation a été proposée dans trois des cinq activités langagières définies dans le CECRL : la compréhension de l'oral, la compréhension de l'écrit et l'expression écrite. Les compétences propres à l'expression orale en continu et en interaction n'ont pas pu être évaluées jusqu'à présent dans le cadre des évaluations Cedre, mais l'expérimentation d'un protocole destiné à l'évaluation de ces compétences est à l'étude depuis 2013.

À l'école, les élèves ont été évalués en anglais ou en allemand. Au collège, ils l'ont été en anglais, en allemand ou en espagnol. Les résultats de ces évaluations ainsi que l'analyse de l'évolution des performances des élèves entre 2004 et 2010 ont donné lieu à la publication de notes synthétiques [Bessonneau, Beuzon, Boucé *et alii*, 2012 ; Bessonneau, Beuzon, Daussin *et alii*, 2012] et de dossiers [Beuzon, Boucé *et alii*, 2013 ; Beuzon, Garcia, Marchois, 2013a, 2013b ; Beuzon, Garcia, Keskpaik *et alii*, 2013].

DES RÉSULTATS EN HAUSSE À L'ÉCOLE ET EN BAISSE AU COLLÈGE

À l'école, en compréhension de l'oral comme en compréhension de l'écrit, les résultats des élèves sont meilleurs en 2010 qu'en 2004 et les écarts s'accroissent

En compréhension de l'oral, le score moyen des élèves de CM2 augmente de 18 points, soit 40 % d'écart-type, ce qui est considérable. Par ailleurs, on constate un plus grand étalement de la répartition des élèves. Ils sont plus nombreux dans les groupes de haut niveau et moins nombreux dans les groupes de niveau intermédiaire.
En compréhension de l'écrit, à l'instar de la compréhension de l'oral, le score moyen des élèves de fin de CM2 a nettement augmenté (+ 22 points). Ces résultats peuvent paraître surprenants au regard de la priorité sans cesse réaffirmée donnée à l'oral dans les textes officiels[1] [MEN, 2007b, p. 4], conformément au CECRL et à la politique linguistique du Conseil de l'Europe.
Par ailleurs, quelle que soit l'activité langagière, on observe une hétérogénéité des résultats selon le sexe. Ainsi, les garçons demeurent beaucoup plus nombreux que les filles aux plus bas niveaux de l'échelle. C'est l'inverse à l'autre extrémité de l'échelle où les filles se démarquent encore plus nettement.

Au collège, baisse des performances en compréhension de l'oral et accroissement des écarts en compréhension de l'écrit

En 2010, on a observé un score moyen en baisse significative de 14 points par rapport à 2004 en compréhension de l'oral. Le pourcentage d'élèves situés dans les niveaux de performances les plus faibles a augmenté tandis qu'à l'autre extrémité, dans les niveaux les plus élevés, le pourcentage d'élèves a diminué de manière significative. En compréhension de l'écrit, le score moyen reste stable. Cependant, l'écart se creuse entre les élèves les moins habiles et les très bons.

1. Préambule commun : « *À l'école élémentaire l'enseignement d'une langue vivante a trois objectifs prioritaires* [...] *lui* [l'élève] *faire acquérir dans cette langue des connaissances et des capacités, prioritairement à l'oral* ».

Comme en fin d'école, les garçons restent plus nombreux que les filles aux plus bas niveaux de l'échelle [Beuzon, Boucé *et alii*, 2013, p. 171].

LES ACQUIS DES ÉLEVES DE FIN DE CM2

L'enquête Cedre de 2010 évalue la première génération d'élèves à avoir vécu la réforme de l'enseignement des langues. En effet, le plan de rénovation des langues, qui s'appuie sur le CECRL, a été introduit en 2005 au collège [MEN, 2005] et en 2007 à l'école [MEN, 2007b]. La France est le premier pays d'Europe à avoir inscrit des références directes au CECRL dans ses programmes de langues [Tardieu, 2008, p. 41] ▸ **Encadré**.
Selon les programmes en vigueur, le niveau A1 est requis en fin d'école primaire, le niveau A2 en fin de classe de cinquième et en fin de troisième, les élèves doivent atteindre un niveau de compétence *« tendant vers B1 »*. Le niveau A1 étant exigé à la fin du CM2 et le niveau A2 pour la validation du socle commun à la fin de la scolarité obligatoire, ce sont ces niveaux qui ont été prioritairement visés pour l'élaboration des items de 2010 en CM2 et en troisième.

PLAN DE RÉNOVATION DE L'ENSEIGNEMENT DES LANGUES

Le ministère de l'Éducation nationale a lancé en 2005 un plan de rénovation de l'enseignement des langues vivantes étrangères qui concerne tous les élèves de l'école élémentaire au lycée. L'objectif de ce plan est d'améliorer le niveau des élèves dans deux langues étrangères dans un contexte d'ouverture européenne et internationale, notamment en renforçant les compétences orales des élèves et en s'appuyant sur le CECRL [MEN, 2006].
Les nouveaux programmes de langues vivantes étrangères à l'école et au collège ont été mis en conformité avec les orientations du CECRL. Ils privilégient l'apprentissage de l'oral au cours de la scolarité obligatoire et une entrée dans les apprentissages par les contenus culturels.

Qu'avons-nous évalué en compréhension de l'oral et de l'écrit en 2010 ?

Afin de mieux appréhender les résultats, il convient de revenir tout d'abord sur le contenu de l'évaluation.
En **compréhension de l'oral**, l'évaluation Cedre a permis de mesurer l'aptitude des élèves à mobiliser des connaissances afin de vérifier qu'ils sont capables dans un message sonore de « connaître et reconnaître des éléments connus » (mots ou expressions lexicalisés) et de « dégager les principales informations d'un document sonore », conformément aux programmes[2]. En conséquence, les situations proposées visaient à évaluer, d'une part, le champ des connaissances lexicales, la maîtrise des expressions figées, des expressions de la vie courante et, d'autre part, la compréhension de courts dialogues, de descriptions.

2. Bulletin officiel hors série n° 8 du 30 août 2007, p. 5 : *« Les élèves doivent acquérir des éléments de base des thèmes culturels et champs lexicaux proposés au niveau A1 : la personne, la vie quotidienne, l'environnement géographique et culturel ».*

Les compétences requises pour la compréhension de l'oral étant, pour la plupart d'entre elles, également pertinentes pour l'**écrit,** nous avons évalué les compétences suivantes : « connaître et reconnaître le lexique », « se faire une idée du contenu d'un texte », « comprendre des phrases, des textes avec ou sans aide visuelle ». Les situations proposées visaient ainsi à évaluer la reconnaissance du lexique, d'expressions de la vie courante, de courtes phrases concernant l'environnement proche, ainsi que la compréhension de courts textes, de courriels ou de cartes postales.

Les compétences des élèves ont-elles progressé de la même manière dans chaque activité langagière ?

Les résultats aux items communs aux deux années permettent de mesurer l'évolution des performances des élèves. Il s'en dégage une tendance majeure : de meilleurs résultats pour la compétence « connaître et reconnaître des éléments connus (mots ou expressions lexicalisés) » que pour la compétence « dégager les principales informations », ceci à l'oral comme à l'écrit. Comme en 2004, les élèves réussissent mieux dans la reconnaissance du lexique que dans la compréhension des principales informations d'un texte, à l'oral comme à l'écrit ▸ **Tableau 1**.

▸ **Tableau 1** **Taux de réussite par compétence aux évaluations de 2004 et de 2010**

	Compréhension de l'oral		Compréhension de l'écrit	
	2004	2010	2004	2010
Connaître et reconnaître	63,6 %	77,7 %	67,6 %	69,7 %
Dégager les principales informations	69,4 %	72,9 %	53,5 %	56,0 %

Que savent faire les élèves ? Quels items réussissent-ils le mieux ?

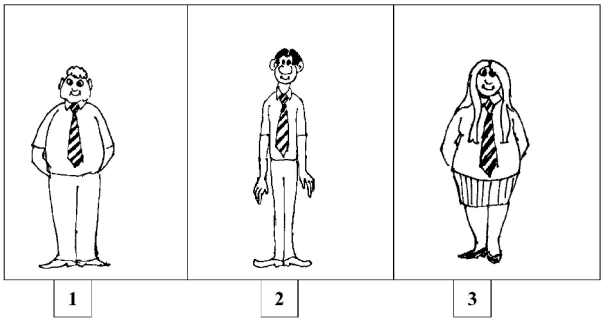
Le classement des items selon le taux de réussite montre que les élèves savent reconnaître, à l'oral comme à l'écrit, un lexique simple et des expressions figées mémorisées très usuelles. Ils comprennent les nombres, les couleurs, le matériel scolaire courant, la nourriture (essentiellement celle concernant le petit-déjeuner) ainsi que des éléments de la description physique. En revanche, il apparaît clairement que lorsque le lexique est un peu moins usuel, ou que le mot porteur de sens est un pronom personnel à la troisième personne ou encore lorsqu'il s'agit d'établir des liens, les élèves réussissent beaucoup moins bien. On remarque que les items les mieux réussis à l'oral sont ceux qui abordent les thèmes de l'anniversaire ou de l'école, thèmes étudiés à maintes reprises au primaire.

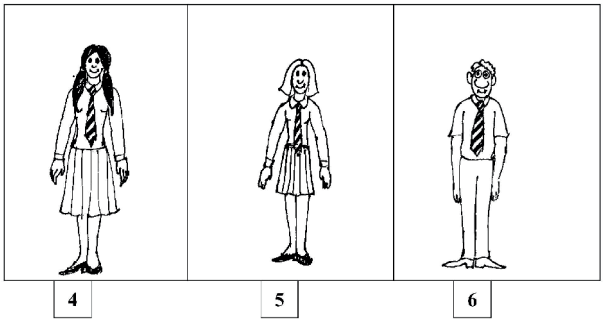
Malgré les bons résultats des élèves aux évaluations de 2010, nous pouvons néanmoins regretter que le corpus lexical maîtrisé par la majorité des élèves soit limité à un nombre réduit de champs lexicaux : les nombres, les couleurs, le matériel scolaire, quelques animaux, le lexique du petit-déjeuner, de la météorologie, et des éléments de description, essentiellement du visage.
L'exemple de la **figure 1 p. 188** illustre la maîtrise différenciée du lexique à l'oral.

Figure 1 Comprendre une description physique à l'oral

Consigne

Vous devez retrouver qui sont Lucy, Tommy, Harry et Mary parmi ces six dessins.
Avant d'écouter leurs descriptions, observez bien tous les petits détails de ces dessins.





1 ☐ Lucy 3 ☐ Mary

2 ☐ Tommy 4 ☐ Harry

Texte de l'enregistrement donné à l'écoute :
« Qui est qui ?
Vous devez retrouver qui sont Lucy, Tommy, Harry et Mary parmi ces six dessins.
Avant d'écouter leurs descriptions, observez bien tous les petits détails de ces dessins.

Écoutez et écrivez le numéro du dessin qui correspond à Lucy à côté de son prénom.
Lucy has got blonde hair. She is very thin.

Écoutez et écrivez le numéro du dessin qui correspond à Tommy à côté de son prénom.
Tommy has got a big nose and black hair.

Écoutez et écrivez le numéro du dessin qui correspond à Mary à côté de son prénom.
Mary has got long black hair. She is tall.

Écoutez et écrivez le numéro du dessin qui correspond à Harry à côté de son prénom.
Harry is short and fat. »

Dans cette situation, il est tout d'abord demandé aux élèves d'observer les six illustrations, puis d'écouter quatre courtes descriptions pour pouvoir associer chacune d'elles à l'illustration correspondante. On évalue ici l'aptitude à repérer des informations explicites, à reconnaître et à mettre en relation des éléments du lexique concernant la description : le visage, les cheveux, la taille et la corpulence.

Pour identifier Lucy, il fallait comprendre *"blonde hair"* et *"thin"*. Si 63 % des élèves y sont parvenus, il est intéressant de constater que presque 22 % d'entre eux ont choisi la troisième illustration, en s'appuyant uniquement sur la compréhension de *"blonde hair"*. Le choix de ce distracteur[3] atteste qu'ils n'ont pas compris la deuxième caractéristique *thin*. La difficile identification de Harry (*"Harry is short and fat"*) pose question : le mot *fat*, et plus généralement les champs lexicaux « discriminants », tels la corpulence, l'aspect physique, sont-ils enseignés à l'école ? Seuls 53 % ont su identifier le portrait de Harry (portrait n° 1) ; 30 % ont choisi le portrait n° 6, le seul personnage masculin (avec le n° 1) restant disponible à ce stade du questionnement.

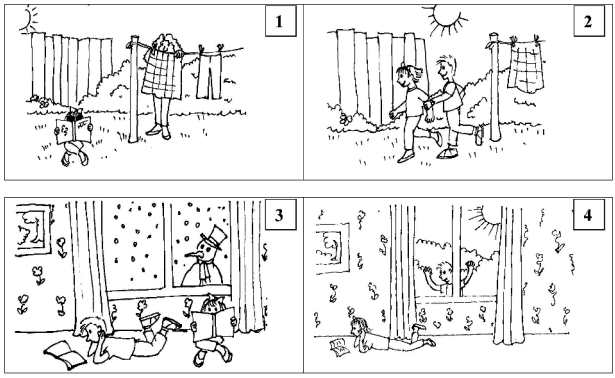
L'exemple de la **figure 2** illustre la compétence la plus difficile à maîtriser pour des élèves de primaire : la prise en compte de plusieurs indices en compréhension de l'oral.
Les élèves doivent ici observer les quatre illustrations, puis écouter deux courtes histoires afin de pouvoir associer chacune d'elles à l'illustration correspondante. On évalue ici l'aptitude des élèves à repérer des informations explicites sur lesquelles ils devront s'appuyer pour construire le sens. Pour parvenir à cet objectif, ils doivent repérer, reconnaître et mettre en relation des éléments du lexique concernant la météorologie, les membres de la famille et les activités. La situation présentée a été réussie par les élèves situés dans les deux groupes supérieurs de l'échelle de performance [Beuzon, Garcia, Keskpaik *et alii,* 2013, p. 108], soit 37 % de l'ensemble. Nous constatons que lorsqu'il est demandé d'identifier hors contexte des éléments du lexique concernant la météorologie, les membres de la famille ou les activités, les résultats sont meilleurs que lorsque cette information est donnée dans un court message oral comme dans la situation p. 190.

Pour identifier l'illustration correspondant à la première histoire, il fallait repérer la situation météorologique (*"It's sunny"*), le lieu (*"in the garden"*) et l'action (*"playing with John"*). Plus de 75 % des élèves ont réussi cet item. Pour ce qui concerne l'illustration correspondant à la deuxième histoire, il fallait repérer la situation météorologique (*"It's sunny"*), le lieu (*"in the garden"*), ainsi que l'action (*"reading"*), mais également la présence de la mère (*"with her mother"*). Il est intéressant de constater que presque 28 % des élèves ont opté pour la quatrième illustration. Le choix de ce distracteur atteste qu'ils n'ont pas pris en compte l'ensemble des indices, notamment la présence de la mère. Le taux de réussite à cet item (55 %), inférieur au précédent, indique que les élèves sont certes capables de prendre en compte plusieurs indices, sans parvenir toutefois à considérer l'ensemble

3. Un distracteur est une réponse proposée, attractive, plausible, mais fausse. Il renseigne sur le type d'erreur ou le cheminement incorrect susceptible d'être suivi par l'élève.

Figure 2 Comprendre à l'oral une scène de la vie quotidienne

Observe bien ces quatre dessins.
Tu vas entendre deux histoires A et B. Elles seront dites deux fois.
Chacune d'elle correspond à un des dessins.
Tu dois repérer le numéro du dessin qui correspond à chacune des histoires.



Histoire A

Écoute une première fois l'histoire A.
Écoute à nouveau et écris dans la case le numéro du dessin qui correspond à l'histoire A.

L'histoire A correspond au dessin n° : ☐ E5VOT190101

Histoire B

Écoute une première fois l'histoire B.
Écoute à nouveau et écris dans la case le numéro du dessin qui correspond à l'histoire B.

L'histoire B correspond au dessin n° : ☐ E5VOT190201

Texte de l'enregistrement donné à l'écoute :
« Observe bien ces quatre dessins. Tu vas entendre deux histoires, A et B. Elles seront dites deux fois. Chacune d'elle correspond à l'un des dessins. Tu dois repérer le numéro du dessin qui correspond à chacune des histoires. Écoute une première fois l'histoire A.
It's sunny today. Lucy's in the garden. She's playing with John.
Écoute à nouveau et écris dans la case le numéro du dessin qui correspond à l'histoire A.
Écoute une première fois l'histoire B :
It's sunny today. Lucy's in the garden with her mother. She's reading a book.
Écoute à nouveau et écris dans la case le numéro du dessin qui correspond à l'histoire B. »

des informations pertinentes. Encore une fois, l'étendue du champ des connaissances est souvent limitée aux groupes nominaux et aux blocs lexicalisés les plus courants, comme l'atteste la difficulté des élèves à identifier les racines de mots (*read* → *reading*). Nous pouvons nous demander si l'entraînement à une pratique raisonnée de la langue est suffisant à l'école.

Nous constatons à l'écrit le même type d'acquis qu'à l'oral. Associer une expression mémorisée à son illustration (exemple : *Happy birthday !* à un gâteau d'anniversaire, *What time is it ?* à une horloge) est considéré comme très facile par les élèves de CM2, ce qui traduit un entraînement régulier en classe. En revanche, l'exemple de la **figure 3** illustre la difficulté des élèves à repérer un lexique usuel dans un contexte précis.
Il est demandé aux élèves de lire un court texte, puis de déterminer si les thèmes proposés y sont abordés on non. On évalue ici l'aptitude des élèves à dégager les principales informations d'un texte court. Pour parvenir à cet objectif, les élèves peuvent s'appuyer sur des mots porteurs d'indices dans les différents thèmes proposés. Comme à l'oral, lorsqu'il s'agit de repérer un lexique simple, fréquemment étudié en classe (comme celui des vêtements, des animaux, des fruits), les résultats sont nettement plus élevés que lorsque le repérage est complexe ou s'appuie sur un lexique plus élaboré ▸ **Tableau 2**. L'item 1 en est une illustration. Pour identifier la présence de deux amis, les élèves pouvaient s'appuyer sur leurs connaissances lexicales de *Mr* et *Mrs*, du mot *friend*, ainsi que sur les indices grammaticaux que constituent les deux pronoms personnels *he* et *she*.

▸ **Figure 3** **Se faire une idée du contenu d'un texte**

Consigne

Monsieur Smith.

Mr Smith is at the market.
He is wearing a black jacket, a pullover and brown shoes.
He has got apples and bananas in his basket.
He is with his friend Mrs Wilson. She has got potatoes, carrots and tomatoes in her basket.
Today she is wearing a blue shirt and a yellow T-shirt because the weather is fine.

Ce petit texte parle-t-il...

	Oui	Non	
de 2 amis.	☐ 1	☐ 2	E5VET710101
de vêtements.	☐ 1	☐ 2	E5VET710102
d'animaux.	☐ 1	☐ 2	E5VET710103
de fruits.	☐ 1	☐ 2	E5VET710104
d'un restaurant.	☐ 1	☐ 2	E5VET710105

▸ **Tableau 2** **Taux de réussite de la situation 3 illustrée à la figure 3, en 2004 et en 2010**

	2004	2010
	Réponse exacte	Réponse exacte
Item 1	42 %	50 %
Item 2	74 %	86 %
Item 3	64 %	84 %
Item 4	57 %	76 %
Item 5	53 %	66 %

La situation montre certes que les résultats sont en hausse entre 2004 et 2010, mais que les items « difficiles » en 2004 le demeurent en 2010. Ce constat nous amène en conséquence à considérer la hausse des résultats entre 2004 et 2010 avec prudence, notamment au regard de la maîtrise du niveau attendu à la fin de l'école primaire[4], à savoir le niveau A1 du CECRL[5] d'une part, et des observations faites dans les classes par les inspecteurs et inspecteurs généraux d'autre part [Inspection générale de l'éducation nationale, 2013, p. 26 à 35].

LES ACQUIS DES ÉLÈVES EN ANGLAIS EN FIN DE TROISIÈME

Comme nous l'avons évoqué précédemment, les nouveaux programmes de langues vivantes étrangères au collège privilégient l'apprentissage de l'oral. Ils encouragent à renforcer l'exposition des élèves à la langue à travers de nombreux dispositifs tels que l'utilisation de ressources audiovisuelles étrangères, les échanges à distance avec des établissements étrangers, pour ne citer qu'eux. Or, les résultats de 2010 indiquent une tendance contraire aux attentes liées à ces préconisations et pointent même une nette baisse des performances des élèves de troisième **en compréhension de l'oral** [Beuzon, Boucé *et alii*, 2013, p. 52-53]. Quant à la **compréhension écrite**, les résultats sont stables, même si l'on note une dispersion accrue des élèves, plus nombreux en 2010 aux deux extrémités de l'échelle de performances.

Qu'a-t-on évalué en compréhension de l'oral en 2010 ? Quels objectifs d'évaluation ont posé le plus problème aux élèves de troisième ?

En compréhension de l'oral, plusieurs compétences ont été évaluées : « percevoir », « identifier » et « construire le sens »[6]. « Percevoir » renvoie aux capacités de discrimination des sons propres à la langue, ceux-là mêmes qui posent le plus de problèmes aux apprenants, tels les phonèmes-voyelles (voyelles simples, brèves et longues comme [i] en opposition avec [i:]). Cette compétence est fondamentale dans l'accès au sens du message puisque la non-discrimination de certains sons peut conduire à des interprétations erronées, notamment dans le cas de paires minimales [Ginesy, 1995, p. 34][7].
L'exemple de la **figure 4**, tiré de l'évaluation Cedre 2010, illustre l'incidence de la discrimination sur la compréhension.

Les élèves devaient cocher parmi trois propositions la phrase entendue. À la question E, les élèves entendaient l'enregistrement suivant : *"I don't want you to slip"* (« Je ne veux pas que vous glissiez »). Que ce soit en 2004 ou en 2010, le choix des élèves s'est majoritairement porté (à 60 %) sur la réponse C (*"I don't want you to sleep"* :

4. Se référer à la partie « Utilisation de l'évaluation Cedre pour estimer le pourcentage d'élèves maitrisant le niveau A1 en fin d'école et A2 en fin de collège », p. 205 de cet article.
5. Se référer à la partie « Le cadre, le socle, les programmes et le positionnement de Cedre », p. 204 de cet article.
6. Certaines des compétences dites de « réception », comme « identifier » et « construire le sens », sont communes à la compréhension de l'oral et à celle de l'écrit.
7. Une paire minimale désigne, en phonologie, deux mots qui ne se distinguent que par un seul phonème. Les linguistes distinguent les phonèmes lorsqu'une différence de sens existe entre deux mots qui forment une paire minimale (ex. : *ship/sheep*).

Figure 4 **Discriminer des sons proches**

Vous allez entendre des phrases.
Cochez à chaque fois parmi les 3 propositions, la phrase que vous entendez.
Attention chaque phrase sera lue tout d'abord une seule fois.

A.	1 ☐	I wake up early every day.
	2 ☐	I work early every day.
	3 ☐	I walk early every day.
B.	1 ☐	Oh my god, she's living in London !
	2 ☐	Oh my god, she lives in London !
	3 ☐	Oh my god, she's leaving London !
C.	1 ☐	Though you said it was true...
	2 ☐	So you said it was true...
	3 ☐	Thought you said it was true...
D.	1 ☐	Here is another pest.
	2 ☐	Here is another pace.
	3 ☐	Here is another piece.
E.	1 ☐	I don't want you to slip.
	2 ☐	I don't want you to split.
	3 ☐	I don't want you to sleep.

« Je ne veux pas que vous dormiez ») et non sur la réponse A, la réponse attendue. Le taux de réussite a même baissé pour cet item entre 2004 et 2010 : d'une réussite de 29 %, on atteint à peine 24 % six ans plus tard. Les élèves n'ont manifestement pas perçu la différence entre la voyelle brève (dans *"slip"*) et la longue (dans *"sleep"*) et ils n'auraient donc pas été en mesure de comprendre le message sonore dans une situation de la vie réelle. Les résultats des autres items de la situation présentée confirment cette tendance à la baisse.

Quant à la seconde compétence visée par l'enquête, « identifier », elle recouvre le repérage et la reconnaissance d'éléments lexicaux comme les nombres, essentiels eux aussi, car véhiculant des informations permettant de comprendre un prix, un nombre de jours, un horaire de train, etc.
Enfin, dernière compétence visée par l'enquête Cedre 2010 : « construire le sens ». Il s'agit d'évaluer la capacité à comprendre l'information contenue dans un message, qu'elle soit explicite ou implicite. Dans ce type de situations, l'élève devait mettre en œuvre des stratégies complexes, dites « de haut niveau » [Portine, 2008] ou « ascendantes » (*bottom-up*) [Vandergrift, 2002] comme établir des liens entre divers éléments (lexicaux, grammaticaux), inférer ou synthétiser l'information pour aboutir à une hypothèse conclusive pertinente.
Prenons un exemple de situation proposée en 2004 et en 2010, qui illustre la

combinatoire des compétences ici visée ▸ **Figure 5**. Dans cette situation, l'élève écoute un dialogue entre deux personnages et doit comprendre pourquoi la jeune fille est triste.

▸ Figure 5 Comprendre l'implicite dans un message oral

DIALOGUE 1

Vous allez entendre un dialogue entre Mike et Jane.
Écoutez attentivement et dites pourquoi Jane est triste.

Cochez la case correspondant le mieux à ce que vous avez compris.

Vous entendrez le dialogue 2 fois.

A - Pourquoi Jane est-elle triste ?

1 ☐ Parce que sa mère est malade.
2 ☐ Parce qu'elle a un problème avec son petit ami.
3 ☐ Parce que son petit ami en préfère une autre.
4 ☐ Parce que son ami Mike est déprimé.

Texte de l'enregistrement donné à l'écoute :

— *Hello Jane, how are you?*
— *So so...*
— *Would you like a cup of coffee?*
— *No thanks, Mike.*
— *Why do you feel so depressed?*
— *(sigh)*
— *Not him again...*
— *You can't understand, I felt so good with him.*
— *Another girl, is it?*
— *I'm afraid so.*

C'est la mise en réseau de plusieurs types d'indices qui permettait de répondre à la question posée. L'existence d'une « autre fille » (*"another girl"*) dans la vie de son petit ami signifie probablement pour Jane la fin de leur relation (traduite par l'emploi du passé dans *"I felt so good with him"*) qui a par ailleurs connu des hauts et des bas (*"not him again"*). La compétence évaluée dans cette situation (comprendre l'implicite) est la plus difficile à maîtriser. Seuls les élèves assez habiles en 2010 ont été capables de repérer les indices pertinents d'une part, puis de les mettre en relation pour pouvoir déduire le sens, d'autre part. En 2004, le taux de réussite était de 64 %, contre 54 % en 2010, soit une baisse de 10 points entre les deux cycles d'évaluation.

Les performances des élèves ont baissé de manière différenciée selon la compétence de l'oral évaluée en 2010

Les résultats dans les trois domaines de compétences accusent une baisse depuis 2010 ; cette tendance est davantage marquée pour « identifier » (- 5 points de pourcentage) : en 2010, les élèves ont eu plus de difficulté à repérer et à identifier un vocabulaire pourtant censé être connu ▸ **Tableau 3**. Doit-on voir là un déficit de réels acquis lexicaux ou une difficulté à mettre en œuvre des stratégies efficaces dans ce domaine, telles que le repérage et l'identification des mots porteurs d'accent, donc de sens, la capacité à opérer une segmentation efficace des énoncés oraux dans une chaîne parlée et la mise en réseau d'indices ?

▸ **Tableau 3** **Taux de réussite en 2004 et en 2010 aux items communs**

Compétences	Taux de réussite		
	Élèves de 2004 sur les items communs 2004-2010	**Élèves de 2010** sur les items communs 2004-2010	**Élèves de 2010** sur tous les items de 2010
Reconnaître, percevoir	63 %	60 %	60 %
Reconnaître, identifier	73 %	68 %	66 %
Construire le sens	59 %	56 %	50 %
Ensemble	67 %	64 %	57 %

Autre constat, à six ans d'intervalle : « construire le sens » demeure la compétence la plus délicate pour les élèves (50 % de réussite contre 60 % et 66 % pour les deux premières compétences). Impliquant la mise en place des opérations complexes évoquées plus haut, la compréhension fine d'un message sonore n'est accessible qu'à des élèves très performants, c'est-à-dire situés dans les groupes supérieurs de l'échelle de performance, les groupes 4 et 5, soit 16 % de l'ensemble. On peut donc parler en fin de troisième d'une forte hétérogénéité des élèves, dont les compétences sont disparates et inégales, en 2010 comme en 2004 [Bessonneau, Beuzon, Boucé *et alii*, 2012, p. 2][8].

Résultats stables en compréhension de l'écrit, mais accroissement des écarts entre les élèves

Dans le domaine de l'écrit, la maîtrise de compétences en jeu à l'oral a également été évaluée, à savoir « reconnaître/identifier » et « construire le sens ». A ainsi été mesurée l'aptitude des élèves à reconnaître le genre d'un document ou les différents registres de langue, à retrouver une information explicite, à inférer la situation d'énonciation à partir d'indices très divers ou encore à synthétiser l'information, comme l'illustre l'exemple de situation proposée à l'écrit en 2010 ▸ **Figure 6**. Les élèves devaient prendre connaissance des avis de quatre lecteurs d'un journal à propos des expérimentations médicales sur les animaux. Ils devaient lire l'opinion

8. En compréhension de l'oral, *« la courbe de répartition des élèves selon les niveaux de l'échelle se décale vers la gauche : le pourcentage d'élèves situés dans les niveaux de performances les plus faibles (groupes 0 et 1) augmente, passant de 15 % à 20,4 % ; à l'autre extrémité de l'échelle, le pourcentage d'élèves dans les niveaux les plus élevés (groupes 4 et 5) diminue de manière significative, évoluant de 23,9 % à 15,7 % ».*

de chacun des lecteurs, puis, pour chaque personne interrogée, indiquer sa position sur le sujet. Pour atteindre cet objectif, les élèves devaient croiser différents éléments d'ordres lexical, syntaxique ou grammatical pour aboutir à une hypothèse synthétique (« pour », « contre », « mitigé »).
Si l'on compare les taux de réussite obtenus à six ans d'intervalle, on constate une certaine stabilité ▸ **Tableau 4**.

Les deux items les mieux réussis (3 et 4) illustrent les compétences des élèves de fin de troisième : ils sont capables de construire le sens en s'appuyant sur un lexique transparent ou sur une information explicite facilement identifiable ("*Nothing can justify*", par exemple, n'a posé aucun problème de compréhension).

▸ Figure 6 Synthétiser les informations dans un document écrit

Consigne : quatre lecteurs d'un journal donnent leur avis à propos de l'expérimentation sur les animaux. Lisez chaque opinion. Puis classez les différents avis en cochant la case qui correspond le mieux.

Is animal testing a good thing?

Kevin Greenwood: *"On the one hand, I think that using animals for medical research is necessary. On the other hand, I'm against animals suffering for our benefit..."*

Helen Cummings: "If testing were not carried out on animals, diseases would still kill thousands of people around the world. People who are against testing should think about the consequences."

Jason Smith: "Animal testing is necessary in certain fields like vaccination. Don't forget that children's lives can be saved and people can live longer."

Sally Archer: *"As a member of the R.S.P.C.A, how could I accept any kind of cruelty to animals? They are living beings and they deserve to be treated as such. Nothing can justify experimenting on monkeys or dogs, not even medical research."*

Kevin Greenwood :

1 ☐ Pour l'expérimentation 2 ☐ Contre l'expérimentation 3 ☐ Avis mitigé (ni pour, ni contre)

Helen Cummings :

2 ☐ Pour l'expérimentation 2 ☐ Contre l'expérimentation 3 ☐ Avis mitigé (ni pour, ni contre)

Jason Smith :

3 ☐ Pour l'expérimentation 2 ☐ Contre l'expérimentation 3 ☐ Avis mitigé (ni pour, ni contre)

Sally Archer :

4 ☐ Pour l'expérimentation 2 ☐ Contre l'expérimentation 3 ☐ Avis mitigé (ni pour, ni contre)

Tableau 4 **Taux de réussite en 2004 et 2010**

	Taux de réussite 2004	Taux de réussite 2010
Item 1	43 %	41 %
Item 2	26 %	23 %
Item 3	65 %	68 %
Item 4	62 %	61 %

Cependant, les deux premiers items (1 et 2) sont nettement moins réussis que les autres. L'item 1 est intéressant, car il illustre bien les compétences des élèves de troisième « moyens ». En effet, plus de 40 % des élèves ont choisi « pour l'expérimentation » ; ils n'ont donc pris en compte qu'un seul indice : la phrase initiale (*"using animals* [...] *is necessary"*) sans la mettre en relation d'une part avec l'avis contrasté exprimé dans *"I'm against animals suffering"* et d'autre part avec les deux expressions *"on the one hand/on the other hand"*, caractéristiques des prises de position mitigées. Seuls quatre élèves sur dix sont parvenus à la réponse attendue.

Cet exemple est à l'image des résultats relevés dans le domaine de l'écrit : la véritable compréhension du message n'est maîtrisée que par les élèves les plus performants : ceux-là mêmes qui sont capables de relever puis d'établir des liens entre les indices, de synthétiser des informations et de bâtir du sens. Or, si entre 2004 et 2010, le pourcentage de ces élèves a augmenté [Bessonneau, Beuzon, Boucé *et alii*, 2012 ; Beuzon, Boucé *et alii*, 2013], on est en droit de s'interroger sur ce que comprennent les élèves les plus faibles (qui sont eux aussi plus nombreux qu'en 2004), dans un document écrit en anglais. Les résultats montrent en effet qu'ils sont seulement capables de reconnaître des éléments épars rencontrés en classe (vocabulaire des consignes écrites, lexique de base), d'associer certaines formes et valeurs, mais il s'agit là de repérages ponctuels, bien distincts de l'accès au sens véritable d'un message.

COMMENT EXPLIQUER LA HAUSSE DES RÉSULTATS EN FIN D'ÉCOLE ENTRE 2004 ET 2010 ?

Évolution institutionnelle de l'enseignement des langues au primaire

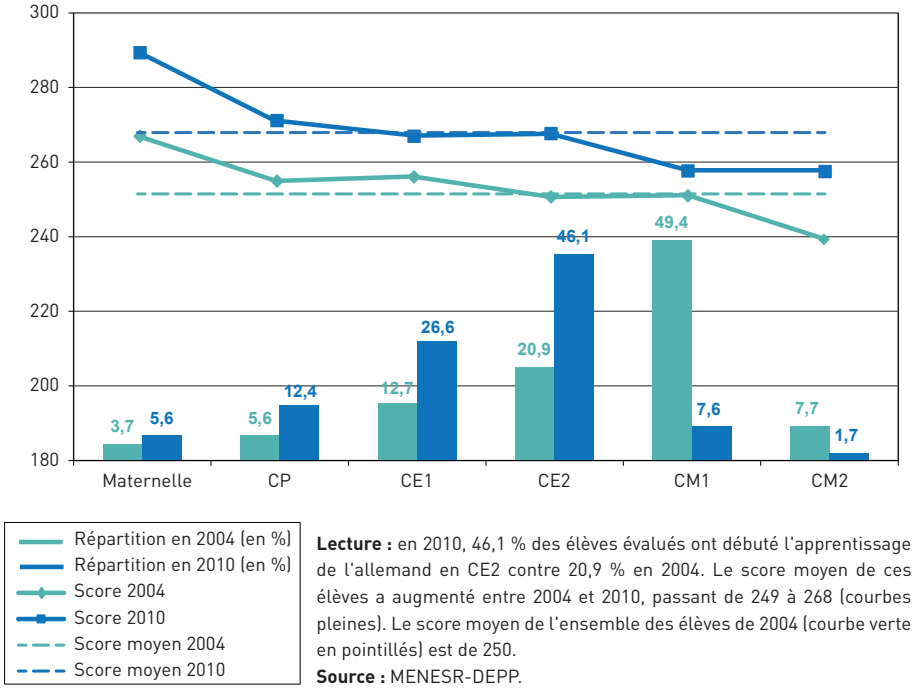
Les enquêtes sur lesquelles nous nous appuyons pour mesurer l'évolution des compétences des élèves ont été effectuées en 2004 et en 2010. Entre ces deux dates, de nombreuses réformes ont été mises en œuvre dans l'enseignement des langues à l'école primaire, notamment le caractère rendu obligatoire de cet enseignement pris en charge de manière plus régulière par les enseignants qui appliquent les programmes de manière plus rigoureuse, comme le souligne Philippe Claus [Bessonneau, Beuzon, Daussin *et alii*, 2012, p. 6].

Rappelons que l'enseignement d'une langue vivante étrangère à l'école élémentaire a été rendu obligatoire en 2002, soit deux années seulement avant la première

évaluation[9] [Azzam-Hannachi, 2005, p. 96-126 ; MEN, 2002]. En conséquence, les élèves évalués en 2004 avaient un parcours en langues très hétérogène, certains d'entre eux n'ayant bénéficié que d'une ou deux années d'apprentissage de l'anglais[10].
On peut donc affirmer que les élèves évalués en 2010 ont bénéficié d'un nombre d'années d'apprentissage de l'anglais plus important que les élèves de 2004, d'une organisation plus régulière et constante de cet enseignement et d'un entraînement plus intensif à la communication, à l'oral essentiellement.
Les scores des élèves selon la classe de début d'apprentissage confortent notre hypothèse ▸ **Figure 7**.

En 2010, les élèves de CM2 ont commencé l'étude de l'anglais plus précocement : plus de 90 % l'ont débutée en CE2 ou avant. C'était le cas d'un peu moins de 43 % d'entre eux en 2004. D'une manière générale, plus cet apprentissage est précoce, meilleurs sont les résultats en fin de CM2.
Cette tendance s'observe pour les élèves évalués en 2004 comme pour les élèves évalués en 2010.
Pour une même classe de début d'apprentissage, le score moyen des élèves de 2010

▸ Figure 7 Scores en compréhension de l'oral en anglais en fin de CM2 selon la classe de début d'apprentissage



Lecture : en 2010, 46,1 % des élèves évalués ont débuté l'apprentissage de l'allemand en CE2 contre 20,9 % en 2004. Le score moyen de ces élèves a augmenté entre 2004 et 2010, passant de 249 à 268 (courbes pleines). Le score moyen de l'ensemble des élèves de 2004 (courbe verte en pointillés) est de 250.
Source : MENESR-DEPP.

9. La mise en place des langues vivantes au primaire a été pendant de nombreuses années une expérimentation contrôlée, puis une sensibilisation, et ensuite une initiation.
10. En 2004, 43 % des élèves avaient commencé à apprendre l'anglais au CE2 ou avant ; ils sont 90 % en 2010.

en **compréhension de l'oral** est plus élevé que celui des élèves de 2004 (figure 7). Mais, si le score moyen a augmenté de 18 points, 2 seulement sont liés à la classe de début d'apprentissage. L'évolution nette, c'est-à-dire celle qui est calculée en tenant compte de l'évolution de la répartition des élèves selon la classe de début d'apprentissage, est donc de 16 points. Il en est de même en compréhension de l'écrit où le score moyen des élèves a augmenté de 22 points dont 3 liés à la classe de début d'apprentissage, soit une évolution nette de 19 points.

On peut donc affirmer que les progrès des élèves entre 2004 et 2010 sont dus partiellement au nombre d'années d'apprentissage. Par conséquent, cette seule piste ne suffit pas pour expliquer la forte hausse des résultats à l'école.
Un deuxième facteur a évolué entre 2004 et 2010 : le statut des enseignants prenant en charge le cours d'anglais et la formation de ces derniers.

Prise en charge de l'enseignement de l'anglais par les enseignants du premier degré de mieux en mieux formés

Outre le nombre d'années d'apprentissage de l'anglais et une régularité de cet enseignement, une autre piste nous semble importante à explorer pour expliquer la forte hausse des résultats des élèves : la prise en charge de l'enseignement de l'anglais par les professeurs des écoles et les efforts considérables de formation pour ces derniers.
Il apparaît clairement qu'en 2010 davantage de professeurs des écoles prennent en charge les cours d'anglais. En effet, ils étaient 50 % à le faire en 2004, ils sont 79 % en 2010 (22 % d'entre eux le font dans le cadre d'un échange de service entre collègues). Parallèlement, le nombre de professeurs du second degré intervenant à l'école diminue fortement, passant de 19 % en 2004, à 10 % en 2010. Toutefois, en 2010, l'enseignement de l'anglais est encore assuré par des intervenants extérieurs[11] dans 17 % des classes.
On peut émettre l'hypothèse que l'offre de formation continue mise en place de manière accrue depuis 2002[12] pour répondre aux préconisations des instructions officielles a permis à bon nombre d'enseignants de se former à la spécificité de cet enseignement et de se (re)mettre à niveau linguistiquement pour le prendre en charge dans de meilleures conditions et en adéquation avec les nouvelles orientations. Malgré une formation linguistique proposée au sein des stages de formation continue, bien que les échanges se développent de plus en plus et que les stages à l'étranger soient plébiscités [MENESR, 2014], beaucoup d'enseignants déclarent ne pas se sentir à l'aise en anglais [Inspection générale de l'éducation nationale, 2013, p. 29]. Pourrait-on alors aller jusqu'à affirmer que les progrès des élèves entre 2004 et 2010 sont également dus à la meilleure formation des enseignants du premier degré à l'enseignement de l'anglais, et tout particulièrement à la didactique de cet enseignement ? Autrement dit, la manière d'enseigner ne serait-elle pas aussi – voire plus – importante que l'aisance dans la langue ? Donner confiance aux élèves,

11. On entend par intervenant extérieur toute personne recrutée et rémunérée localement par les directions académiques parmi les locuteurs natifs et les diplômés en langues ayant passé avec succès une procédure d'habilitation.
12. Parmi les enseignants du premier degré qui ont fait passer l'évaluation, 26 % déclarent avoir suivi une formation continue de plus de 30 heures en 2004 ; ils sont 29 % en 2010. Le questionnaire de contexte ne permet pas d'avoir de regard sur les stages d'une durée inférieure ou encore sur les animations pédagogiques. Or, les efforts de formation ces dernières années ont également pris la forme de stages « courts ».

les motiver en mettant en œuvre des situations variées [MEN, 2007b] porterait-il autant ses fruits que l'immersion une heure par semaine avec un assistant, très à l'aise dans sa langue, mais souvent beaucoup moins quand il s'agit de l'enseigner à de jeunes enfants ?

Nos propos rejoignent les conclusions des travaux de Derivry-Plard [2006] qui a montré qu'en formation des adultes, les enseignants locuteurs natifs obtiennent de moins bons résultats que les enseignants locuteurs non natifs (socialisés dans le pays de la langue étudiée et plus conscients des difficultés d'apprenants francophones). Elle met ainsi en évidence l'importance pour l'apprentissage du lien privilégié établi entre le professeur et sa classe : même si la compétence en langues des enseignants du premier degré est moindre que celle des spécialistes, ils ont un « meilleur » impact en tant qu'enseignant de la classe. Une conséquence directe ou non est la confiance acquise par les élèves dans l'utilisation d'une autre langue. Les élèves du primaire interrogés [Beuzon, Garcia, Keskpaik *et alii*, 2013, p. 106] estiment dans une large majorité (83 % en 2010 et 82 % en 2004) que le cours d'anglais est un moment agréable[13]. Ils sont plus nombreux en 2010 à oser prendre la parole en anglais, même s'ils font des erreurs (76 % contre 71 % en 2004). Ils sont également plus nombreux (65 % en 2010 contre 52 % en 2004) à ne pas être gênés lorsqu'il faut prendre la parole en anglais devant les autres élèves. L'analyse « toutes choses égales par ailleurs » [*id.*, 2013, p. 53] montre une relation positive entre intérêt et performance.

Les élèves qui manifestent le plus de plaisir et d'intérêt pour le cours d'anglais et pour les activités qui y sont proposées obtiennent un score supérieur de 9 points en compréhension de l'écrit et de 8 points en compréhension de l'oral. Enfin, l'analyse révèle que la motivation a une relation positive au score. Ceux qui disent apprécier beaucoup l'anglais ont un score supérieur de 22 points en compréhension de l'écrit et de 13 points en compréhension de l'oral, par rapport aux élèves qui disent ne pas aimer du tout cette langue. De même, ceux qui jugent important de connaître l'anglais ont un score plus élevé que ceux qui pensent que ce n'est pas important.

Les attentes grandissantes des parents vis-à-vis de l'enseignement de l'anglais

L'importante médiatisation de l'enseignement de l'anglais au primaire a entraîné un intérêt croissant des parents pour l'apprentissage de cette langue comme en témoigne le succès des *mini schools*, des centres de loisirs bilingues et des séjours linguistiques. De ce fait, les jeunes enfants sont de plus en plus en contact avec la langue anglaise en dehors du cadre scolaire, ce qui permet une acquisition certaine de compétences et de connaissances.
Les questionnaires de contexte[14] confirment nos hypothèses. En 2010, les élèves sont plus nombreux à déclarer avoir utilisé l'anglais en vacances dans un pays anglophone (29 % contre 23 % en 2004) et à avoir un correspondant anglais (16 % contre 12 % en 2004). Les élèves qui déclarent avoir été en vacances dans un pays où ils devaient utiliser l'anglais pour être compris ont obtenu un score plus élevé

13. Données recueillies grâce au questionnaire de contexte.
14. Les éléments de contexte sont issus de l'analyse des réponses des élèves, des enseignants et des chefs d'établissement, participant aux questionnaires qui leur étaient adressés.

(+ 6 points en compréhension de l'écrit et + 7 points en compréhension de l'oral) que ceux qui n'ont pas eu cette opportunité.
L'analyse « toutes choses égales par ailleurs » [Beuzon, Garcia, Keskpaik *et alii*, 2013, p. 53-54] montre également que la langue parlée à la maison influe sur le score en anglais. Les élèves qui déclarent avoir parlé l'anglais avec leurs parents lorsqu'ils étaient plus jeunes obtiennent un score plus élevé lors de l'épreuve écrite (+ 11 points) et lors de l'épreuve orale (+ 18 points) que ceux qui disent ne parler que le français ou une autre langue à la maison.

À l'école, il semble donc que la hausse des résultats des élèves puisse s'expliquer par l'augmentation du nombre d'années d'apprentissage, par la prise en charge de cet enseignement par des enseignants de mieux en mieux formés, par un entraînement aux compétences de l'oral plus systématique, par un intérêt croissant des familles pour cet apprentissage et en conséquence par une plus grande exposition à la langue en dehors de l'école. Il faut attendre la prochaine évaluation de 2016 pour confirmer ces hypothèses ou les compléter.

COMMENT EXPLIQUER LA BAISSE DES RÉSULTATS DEPUIS 2004 EN COMPRÉHENSION DE L'ORAL AU COLLÈGE ?

Il est difficile d'apporter une réponse univoque à cette question, en premier lieu parce que les élèves concernés par l'enquête de 2010 sont différents sur bien des plans de ceux évalués en 2004 [Beuzon, Boucé *et alii*, 2013]. Nous pouvons néanmoins avancer quelques pistes.

Les élèves ont une image plus négative de leurs performances en anglais

Les réponses aux questionnaires de contexte apportent un éclairage intéressant sur les résultats des élèves et leur perception de l'anglais, qu'ils étudient majoritairement au collège comme première langue[15]. Les réponses des élèves à ce questionnaire sont éloquentes : en 2010, comme en 2004, presque la moitié d'entre eux estiment leurs résultats scolaires en anglais « moyens », un quart les estiment « mauvais » et seulement environ 25 % d'entre eux les qualifient de « bons ». À peine deux élèves sur dix ont donc une perception positive de leurs performances en anglais.

Faut-il y voir une relation de cause à effet ? Comme en 2004, un élève sur deux déclare aimer « un peu » l'anglais. Ils ne sont qu'à peine un tiers à déclarer aimer « beaucoup » l'anglais : il est fort probable qu'il s'agisse de la catégorie d'élèves les plus performants, ayant jugé à 25 % leurs résultats « bons ».

La piste didactique et la relation au professeur

La relation au professeur d'anglais n'est pas à négliger dans l'analyse des résul-

15. La quasi-totalité des élèves évalués (97,7 %) suivait un enseignement « première langue ».

tats : selon cette même enquête contextuelle, elle est en effet perçue par les élèves plus négativement en 2010 que six ans plus tôt. En effet, ils sont moins nombreux à déclarer que leur professeur s'adresse à eux exclusivement en anglais (56 % contre 64 % en 2004) et qu'il s'efforce de faire participer tous les élèves presque tout le temps (51 % contre 61 % en 2004). Ils sont également plus nombreux à estimer que leur professeur ne privilégie ni n'encourage « jamais » ou seulement « de temps en temps » leurs productions orales, surtout s'ils font des erreurs (16 % contre 10 % en 2004). La mobilisation du temps de parole par le professeur demeure un point sensible, malgré une évolution positive depuis 2004. Pour un tiers des élèves (42 % en 2004), l'enseignant le mobilise « la moitié du temps ».

Et les nouvelles technologies ?

Curieusement, les cours d'anglais peinent à entrer dans « l'ère du numérique »[16] ainsi qu'en témoignent les résultats de notre enquête. Malgré le développement dans les collèges des matériels ou ressources TICE depuis 2004 (tableaux numériques interactifs, tablettes, outils de « baladodiffusion »[17], laboratoires de langue, pratique de visioconférence), selon une écrasante majorité d'élèves (85 % à 95 %), ces ressources ne servent jamais d'appui au cours d'anglais. Les cassettes audio, qui accompagnent encore parfois le manuel scolaire, sont quant à elles toujours bien présentes dans les classes (63 % en 2010). 70 % des élèves déplorent que les programmes informatiques ou Internet ne servent « jamais » en cours d'anglais. Ces déclarations sont par ailleurs confirmées par les enseignants eux-mêmes : « Les technologies de pointe sont encore très rarement utilisées par les enseignants interrogés (selon respectivement 90 %, 67 % et 97 % d'entre eux) » [Beuzon, Boucé *et alii*, 2013, p. 66]. Pourtant, toujours selon notre enquête contextuelle, 25 % des élèves considèrent Internet comme une véritable occasion d'être en contact avec l'anglais : ils étaient nettement moins nombreux à le penser en 2004 (18 %). Ainsi, les supports susceptibles d'entretenir, voire d'alimenter la motivation des élèves à apprendre l'anglais sont encore très peu utilisés en classe en 2010.

Les élèves fournissent-ils un travail personnel suffisant en anglais ?

Concernant leur travail personnel en anglais, là encore, le constat est sans appel. On ne note aucune évolution dans le temps dédié aux devoirs en anglais depuis 2004 : en 2010, plus d'un élève sur deux déclarent toujours passer moins de 15 minutes ou entre 15 et 30 minutes chaque semaine à faire leurs devoirs d'anglais. Les élèves passent donc entre 5 et 10 minutes à faire leurs devoirs d'anglais entre chaque cours de la semaine. Dans ces conditions, peut-on parler véritablement d'un entraînement et d'un développement des compétences orales et écrites ? Elles ne peuvent en effet pas être le seul apanage du travail en classe, compte tenu du nombre d'élèves, de la durée réelle d'une heure de cours si l'on soustrait le temps consacré à l'installation des élèves, au rituel d'accueil, à la gestion et la régulation de la vie de la classe et la mise au travail effective d'un groupe d'environ 30 élèves.

16. Référence à la campagne du ministre de l'Éducation nationale, Vincent Peillon, lancée en décembre 2012 : « Faire entrer l'école dans l'ère du numérique ».
17. Baladodiffusion : distribution de contenus audio et vidéo pour baladeur sous forme de fichiers, téléchargés à partir d'un ordinateur sur une clé USB, lecteur MP3 ou MP4, téléphones portables, smartphones, etc.

Une exposition à la langue étonnamment faible en 2010

Même si les élèves de 2010 ont davantage conscience de l'importance de savoir parler l'anglais, ils ne voient pas d'intérêt majeur à s'exposer à la langue en dehors des cours, ce qui est en parfaite contradiction avec ce que préconisent les programmes : *« Il va sans dire que l'exposition à la langue doit trouver des prolongements en dehors du cours et que doivent être encouragés les recours* [...] *au TIC »* [MEN, 2007a, p. 32].
Contrairement à une idée répandue liée à la popularité croissante des séries télévisées auprès des jeunes, le volet contextuel de l'enquête Cedre révèle que les élèves sont toujours aussi peu attirés en 2010 par les films, vidéos, séries en version originale : en 2004, seulement 2 % des élèves regardaient tous les jours des émissions de télévision ou des films en version originale sous-titrée. Ils ne sont pas plus de 5 % en 2010. On note toutefois un début d'inversion de cette tendance puisqu'ils sont plus de 20 % à en regarder une fois, voire plus d'une fois par semaine, alors qu'ils n'étaient que 8 % en 2004. L'exposition à la langue demeure trop faible : seuls deux élèves sur dix en 2010 sont en contact régulier avec un type de supports authentiques (c'est-à-dire non didactisés[18]) susceptibles de développer leurs compétences orales en anglais.

La compréhension orale : une activité langagière majeure

La compréhension de l'oral constitue l'une des étapes les plus fondamentales de la communication. Avant de pouvoir formuler et transmettre un message oral, l'élève doit avoir entendu et compris la langue étrangère. Cette activité langagière est l'un des rouages majeurs de l'apprentissage car « [c'est] *la compréhension orale qui conditionne la prise de parole des élèves »* [MEN, 2005]. La compréhension précède l'expression, indissociable de la réception du message en amont.
Au collège, les programmes du palier 2 rappellent qu'une *« place de choix doit être réservée à l'entraînement à la production orale en continu »* [MEN, 2007a, p. 29]. Les professeurs de collège doivent désormais l'évaluer comme les autres activités langagières plus traditionnelles (compréhension et expression écrites, notamment) afin de valider le niveau A2 et les compétences du livret personnel de compétences.

Cette nouvelle priorité – souvent chronophage – s'est-elle traduite dans les classes par un moindre entraînement aux activités de compréhension ? C'est la piste qu'avance François Monnanteuil, inspecteur général d'anglais, pour expliquer la baisse des performances des élèves de troisième en compréhension de l'oral [Bessonneau, Beuzon, Boucé *et alii,* 2012]. Pourtant les supports et ressources à la disposition des professeurs d'anglais ne manquent pas. Depuis quelques années, l'essor des banques de sons ou de sites Internet de ressources sonores (par exemple *audio-lingua*, *Elllo*, *Randall's ESL*, *Web Language Lab, Soundguide* pour ne citer qu'eux) facilitent désormais l'accès à des documents authentiques de qualité, libres de droit pour la plupart.
Encore faut-il les didactiser, c'est-à-dire les transformer en outils d'entraînement propres à développer chez l'élève de véritables stratégies d'écoute, transférables à d'autres supports. Il est en effet très difficile pour un apprenant de faire du sens face à un document oral présentant un débit naturel assez rapide, un ou des accents

18. *« La didactisation est* [donc] *le processus de transformation de documents authentiques à des fins pédagogiques »* [Hardy, 2005, p.19].

spécifiques[19] et des éléments paralinguistiques tels les bruits de fond notamment. Le professeur doit lui fournir explicitement, lors d'un apprentissage méthodique, les différentes clés d'accès au sens du message[20] et ainsi le rendre compétent ou autonome dans ce domaine.
La confusion possible entre vérification/évaluation et entraînement constitue un écueil important dans l'enseignement de la langue et sans doute une réelle difficulté pour les enseignants.
Les résultats des élèves de collège en anglais devraient bénéficier dans les prochaines années de deux avancées importantes. D'une part, la hausse des performances des élèves de l'école primaire en anglais[21], qui laisse espérer voir en 2016 (lors du prochain cycle d'évaluations Cedre en langues) cette tendance positive gagner du terrain au collège également ; d'autre part, l'introduction de nouvelles épreuves orales depuis 2013 (en compréhension et expression) dans les séries générales et technologiques du baccalauréat aura sans doute à terme un effet modélisant sur les pratiques des enseignants de collège (*washback effect*) [Heaton, 1990] et devrait contribuer à redonner à la compréhension de l'oral une place prépondérante dans l'apprentissage de l'anglais.

COMBIEN D'ÉLÈVES MAÎTRISENT LE NIVEAU A1 EN FIN DE CM2 ET A2 EN FIN DE TROISIÈME ?

Le cadre, le socle, les programmes et le positionnement de Cedre

Comme nous l'avons déjà dit, le CECRL définit six niveaux de compétence en langue, du plus bas, noté A1, au plus élevé, noté C2. Ces niveaux décrivent les aptitudes, compétences et connaissances que l'apprenant doit acquérir et auxquelles sont associés des types de tâches spécifiques[22]. Selon les programmes en vigueur, le niveau A1 est requis en fin d'école primaire, le niveau A2 en fin de classe de cinquième et en fin de troisième les élèves doivent atteindre un niveau de compétence « tendant vers B1 ». Le niveau A2 étant exigé pour la validation du socle commun à la fin de la scolarité obligatoire, c'est ce niveau qui a été prioritairement visé pour l'élaboration des items de Cedre 2010 en troisième. Mais, afin d'apprécier au mieux les différentes performances des élèves, il a aussi été proposé au collège des items de niveau A1 et de niveau B1.
Afin de déterminer le pourcentage d'élèves ayant atteint le niveau A1 en fin d'école et le niveau A2 en fin de collège, plusieurs sources peuvent être confrontées. Une première estimation a été réalisée à l'aide de l'enquête Cedre 2010. En 2012 et 2013,

19. L'anglais se caractérise par une grande diversité d'accents régionaux et nationaux, mais également liés à l'origine sociale.
20. Une clé peut être l'identification de la situation d'énonciation, le repérage puis la mise en relation d'éléments porteurs de sens et/ou d'accents, de schémas intonatifs ; la discrimination de phonèmes, etc.
21. Se référer à la partie : « À l'école, en compréhension de l'oral comme en compréhension de l'écrit, les résultats des élèves sont meilleurs en 2010 qu'en 2004 et les écarts s'accroissent », p. 185 de cet article.
22. *« Le nouveau contexte d'apprentissage des langues vivantes, avec l'adoption par décret en date du 22 août 2005 du Cadre européen commun de référence pour les langues (CECRL)* [...] *met en avant l'objectif de communication dans une variété de situations aussi proches que possible de celles de la vie réelle. »*

une évaluation de la compétence 2 du socle a permis de procéder à un *standard setting*, c'est-à-dire à une détermination de seuils de compétences [Miconnet et Vourc'h, dans ce numéro, p. 141] dont on présentera rapidement ici les résultats. Nous constaterons enfin que les résultats sont en partie cohérents avec *l'Étude européenne sur les compétences en langues de 2011* (ESLC) [Bessonneau et Verlet, 2012], mais, qu'en revanche, on observe des écarts importants avec les résultats des élèves recensés *via* le livret personnel de compétences.

Utilisation de l'évaluation Cedre pour estimer le pourcentage d'élèves maîtrisant le niveau A1 en fin d'école et A2 en fin de collège

Chaque item de l'évaluation est rattaché au niveau qu'il est censé évaluer (A1, A2, B1) en fonction des attendus du CECRL. Plusieurs cas de figure ont été établis, selon le seuil d'exigence auquel on fixe la probabilité moyenne de réussite des items de niveau A1 à l'école et A2 au collège.

À l'école, en compréhension de l'oral, le pourcentage d'élèves ayant atteint le niveau A1 se situe entre 58 % et 91 % en 2010, selon le degré d'exigence retenu ▸ **Tableau 5**.

En compréhension de l'écrit, ce pourcentage est compris entre 32 % et 73 % en 2010. Dans les deux domaines, on note une progression générale depuis 2004.

Au collège, en compréhension de l'oral, c'est entre 24 % et 67 %, que le pourcentage d'élèves ayant atteint le niveau A2 se situe en 2010 selon le degré d'exigence retenu ▸ **Tableau 6**.

En compréhension de l'écrit, pour la même année, le pourcentage d'élèves ayant atteint le niveau A2 se situe entre 26 % et 63 %.

Comparée avec 2004, la situation s'est dégradée pour la compréhension de l'oral. Pour la compréhension de l'écrit, il y a moins de cohérence dans l'évolution selon le seuil retenu. L'évaluation Cedre permet ainsi de définir une première approche du pourcentage d'élèves maîtrisant le niveau A1 en fin de CM2 et A2 en fin de troisième. Ce pourcentage a été estimé de manière plus précise à l'occasion de l'évaluation de la compétence 2 du socle en 2012 et en 2013.

▸ **Tableau 5** **Pourcentage d'élèves de CM2 maîtrisant le niveau A1 selon le degré d'exigence retenu**

Probabilité moyenne de réussite		Score seuil pour l'oral	Compréhension orale	Score seuil pour l'écrit	Compréhension écrite
50 % des items	2004	191	91 %	239	58 %
	2010		91 %		73 %
66 % des items	2004	224	69 %	269	30 %
	2010		75 %		47 %
75 % des items	2004	247	47 %	289	18 %
	2010		58 %		32 %

Lecture : en 2004, 58 % des élèves ont une probabilité moyenne de réussite aux items de niveau A1 de 50 % en compréhension écrite. Ils sont 73 % en 2010. Ces élèves ont un score supérieur ou égal à 239 en compréhension de l'écrit.
Source : MENESR-DEPP.

Tableau 6 Pourcentage d'élèves de troisième maîtrisant le niveau A2 selon le degré d'exigence retenu

Probabilité moyenne de réussite		Score seuil pour l'oral	Compréhension orale	Score seuil pour l'écrit	Compréhension écrite
50 % des items	2004	218	73 %	224	68 %
	2010		67 %		63 %
66 % des items	2004	245	48 %	257	39 %
	2010		36 %		40 %
75 % des items	2004	261	34 %	285	22 %
	2010		24 %		26 %

Lecture : en 2004, 68 % des élèves ont une probabilité moyenne de réussite aux items de niveau A2 de 50 % en compréhension écrite. Ils sont 63 % en 2010. Ces élèves ont un score supérieur ou égal à 224 en compréhension de l'écrit.
Source : MENESR-DEPP.

Évaluation de la compétence 2 du socle et détermination des standards minimaux

Dans le cadre de la LOLF (loi organique relative aux lois de finances), la DEPP fournit des indicateurs de maîtrise des compétences du socle. En 2012, ces indicateurs ont été calculés pour la première fois pour la compétence 2 (pratique d'une langue étrangère). L'évaluation consistait en une reprise de certains items de l'évaluation Cedre. Pour déterminer le point de bascule entre maîtrise et non-maîtrise du socle, plusieurs méthodes ont été utilisées et leurs résultats ont été confrontés. Pour plus de détails sur la détermination des standards minimaux et la méthodologie employée, le lecteur peut se référer à l'article de Miconnet et Vourc'h [dans ce numéro, p. 141].
Selon cette méthode, 73 % des élèves maîtrisent en fin d'école la compétence 2 du socle en compréhension de l'écrit ; ils sont 86 % pour la compréhension de l'oral ▸ Tableau 7.
Au collège, 44 % des élèves maîtrisent la compétence 2 du socle en fin de collège en compréhension de l'écrit ; ils sont à peine 30 % pour la compréhension de l'oral ▸ Tableau 8.

Tableau 7 Pourcentage d'élèves maîtrisant la compétence 2 du socle en fin d'école

	2012	2013
Compréhension de l'écrit	73,0 %	73,0 %
Compréhension de l'oral	86,1 %	85,5 %

Tableau 8 Pourcentage d'élèves maîtrisant la compétence 2 du socle en fin de collège

	2012	2013
Compréhension de l'écrit	43,3 %	44,2 %
Compréhension de l'oral	26,9 %	29 %

En compréhension de l'oral, ces résultats convergent avec ceux de l'étude européenne sur les compétences en langues (ESLC) réalisée en 2011 auprès d'élèves de 14 à 16 ans. Même s'il est difficile de comparer ces deux évaluations, qui ne reposent ni sur les mêmes protocoles ni sur la même méthodologie, l'enquête européenne révèle la faiblesse du niveau général des élèves français qui ne possèdent selon elle que des connaissances de base en anglais[23]. Seuls 26 % des élèves de 14 à 16 ans maîtriseraient au moins le niveau A2 en compréhension de l'oral, avec une proportion d'élèves de niveau pré-A1 très importante [Bessonneau et Verlet, 2012]. Ce chiffre rejoint également les 24 % d'élèves maîtrisant à 75 % les items A2 de l'évaluation Cedre (tableau 6).
Des écarts plus importants sont observés pour la compréhension de l'écrit (26 % pour ESLC contre 43 % pour le socle et 26 % pour Cedre). Le lecteur pourra se référer à l'article cité précédemment pour plus de détails.

Niveau A2 et livret personnel des compétences

Le livret personnel de compétences (LPC) atteste l'acquisition des connaissances et compétences du socle commun, de l'école primaire à la fin de la scolarité obligatoire. Depuis respectivement 2008 et 2009, il est généralisé à toutes les écoles primaires et à tous les collèges. Rappelons que depuis 2011, la validation du socle est obligatoire pour l'obtention du diplôme national du brevet (DNB)[24] et garantit notamment la maîtrise du niveau A2 dans une langue vivante étrangère.
Or, on note des écarts très importants entre le pourcentage d'élèves validés au niveau A2 par leurs enseignants en fin de troisième via le LPC (93 %) et les pourcentages d'élèves attestant d'une maîtrise de ce niveau, selon les résultats des trois enquêtes mentionnées ci-dessus.
Aussi, le poids de la contrainte institutionnelle, impliquant pour les élèves l'impossibilité d'obtenir le diplôme national du brevet en cas de non-validation de la compétence 2 du socle, soulève-t-il un problème majeur : la prise en compte des seules attestations des enseignants peut-elle constituer un indicateur robuste du niveau de compétence des élèves en anglais au sortir du collège ?
La validation massive du niveau A2 par les professeurs pose question alors que selon nos trois études standardisées, les élèves de troisième ne seraient qu'entre 24 % et 29 % seulement à maîtriser ce niveau à l'oral et aux alentours de 40 % à l'écrit [Miconnet et Vourc'h, dans ce numéro, p. 141].

23. À titre indicatif, en Suède, 96 % des élèves maîtrisent au moins le niveau A2 en compréhension de l'oral, 89 % en compréhension de l'écrit et 94 % en expression écrite [Bessonneau et Verlet, 2012].
24. Trois éléments sont pris en compte pour l'obtention du diplôme : la maîtrise du socle commun de connaissances et de compétences, palier 3, attestée par le « Livret personnel de compétences » ; les notes obtenues à l'examen du brevet ; les notes de contrôle continu, effectué tout au long de l'année.

CONCLUSION

Le Conseil de l'Europe et l'Union européenne ont érigé en combat commun l'amélioration des compétences en langues. La maîtrise d'au moins une langue étrangère est également une des préconisations du Sénat[25], le plan de rénovation des langues lancé en 2005 en étant une des manifestations. Or, les résultats de 2010 au collège affichent une baisse inquiétante à l'oral doublée de constats étonnants, tels la faible utilisation des nouvelles technologies dont les performances et les possibilités ne sont plus à démontrer, ou encore la moindre exposition à la langue en dehors du cadre scolaire.
À l'école, les résultats en hausse à l'oral comme à l'écrit laissent penser que l'enseignement de l'anglais bénéficie à la fois d'une plus grande fréquence, d'une régularité et d'une meilleure formation des enseignants.
Si les résultats de l'école sont teintés d'optimisme, il faut toutefois les relativiser. Cette forte hausse ne doit pas occulter le fait que le pourcentage d'élèves maîtrisant le niveau A1, niveau requis en fin de CM2, reste insuffisant[26].

Les résultats du collège vont à l'encontre des préconisations officielles (plan de rénovation des langues) et des investissements qui y ont été faits (plan numérique). Ils conduisent à s'interroger sur les difficultés à mettre en place de nouvelles réformes et sur la place de l'anglais, plus généralement des langues étrangères à la fois dans notre système éducatif et dans notre société. L'apprentissage des langues étrangères est trop souvent l'apanage de l'école au sens large et conduit à un apprentissage en milieu clos qui ne suscite pas toujours chez les élèves une très grande motivation. Or, l'incidence de la motivation et de la confiance en soi sur les performances des élèves n'est plus à démontrer. Ne pourrions-nous pas envisager qu'une place plus grande soit faite à la diffusion de films, de reportages ou de documentaires en version originale privilégiant le sous-titrage plutôt que leur doublage en français ?
Une deuxième piste serait une plus grande communication auprès des parents autour des programmes internationaux d'échanges existants (Comenius[27], Voltaire[28]), très prisés par les élèves des classes de sections européennes de lycée. Ces programmes ayant pour mission d'encourager la mobilité et l'ouverture internationale auraient un impact certain sur les performances des élèves de collège. Cependant, la participation à ce type de dispositifs s'inscrit-elle dans les habitudes culturelles des jeunes élèves français ?
Les résultats de Cedre 2016 seront très attendus : la hausse des performances des élèves de CM2 se poursuivra-telle ? Les résultats prometteurs de ceux qui ont quitté l'école primaire en 2010 seront-ils confirmés ou infirmés par les résultats de ceux qui seront en troisième dans deux ans ? L'enjeu est important : nous saurons alors si cette tendance aura eu des répercussions mesurables sur les performances des élèves qui seront soumis à l'enquête de 2016 en fin de troisième.

25. Site officiel du Sénat, www.senat.fr : *« Des expériences performantes et innovantes, porteuses d'un potentiel de renouvellement des pratiques pédagogiques, ont été menées ces dernières années aux niveaux européen, national ou local. Leur diffusion doit devenir une priorité, afin que ces actions au rôle pionnier, bien souvent encore trop confidentielles, concernent demain une majorité d'élèves. C'est la condition pour donner un sens concret à l'objectif de plurilinguisme. »* [Legendre J., 2003].
26. Voir tableau 5 p. 205.
27. L'action « Comenius de mobilité individuelle des élèves » permet aux élèves de l'enseignement secondaire, âgés de quatorze ans au moins, d'effectuer un séjour de trois à dix mois dans un établissement et dans une famille d'accueil à l'étranger.
28. Le programme Voltaire a été adopté par les deux gouvernements sur une idée de Brigitte Sauzay lors du sommet franco-allemand de Potsdam en 1998 et s'adresse à des élèves de troisième et de seconde en France et à des élèves de "*8., 9. et 10. Klasse*" en Allemagne. L'échange fonctionne sur le principe de la réciprocité.

BIBLIOGRAPHIE

Azzam-Hannachi R., 2005, *Évolution de l'enseignement des langues vivantes à l'école primaire en France : formation et représentation des enseignants du premier degré*, Thèse de doctorat, Université de Nancy 2, p. 96-126.

Bessonneau P., Beuzon S., Boucé S., Daussin J.-M., Garcia E., Lévy M., Marchois C., Trosseille B., 2012, « L'évolution des compétences en langues des élèves en fin de collège de 2004 à 2010 », *Note d'information*, n° 12.05, MEN-DEPP.

Bessonneau P., Beuzon S., Daussin J.-M., Garcia E., Lévy M., Marchois C., Trosseille B., 2012, « L'évolution des compétences en langues des élèves en fin d'école de 2004 à 2010 », *Note d'information*, n° 12.04, MEN-DEPP.

Bessonneau P., Verlet I., 2012, « Les compétences en langues étrangères des élèves en fin de scolarité obligatoire – Premiers résultats de l'Étude européenne sur les compétences en langues 2011 », *Note d'Information,* n° 12.11, MEN-DEPP.

Beuzon S., Boucé S., Garcia É., Keskpaik S., Marchois C., 2013, « L'évolution des compétences en anglais, en espagnol et en allemand des élèves en fin de collège », *Les Dossiers*, n° 204, MEN-DEPP.

Beuzon S., Garcia E., Keskpaik S., Marchois C., 2013, « L'évolution des compétences en anglais et en allemand des élèves en fin d'école », *Les Dossiers*, n° 203, MEN-DEPP.

Beuzon S., Garcia E., Marchois C., 2013a, *Anglais en fin de collège – L'évolution des compétences 2004-2010*, CNDP.

Beuzon S., Garcia E., Marchois C., 2013b, *Anglais en fin d'école primaire – L'évolution des compétences 2004-2010*, CNDP.

Conseil de l'Europe, 2001, *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*, Paris, Didier.

Derivry-Plard M., 2006, « Les enseignants "natifs" et "non-natifs" de langue(s) : catégorisation linguistique ou construction sociale ? », *Travaux de didactique du FLE,* n° 55, p. 100-108.

Ginésy M., 1995, *Mémento de phonétique anglaise*, Paris, Nathan Université, p. 34.

Hardy M., 2005, « La didactisation de documents authentiques pour l'enseignement des langues de spécialité : pourquoi et comment ? » *Les Langues Modernes*, n° 99, vol.1, p. 19-30.

Heaton J. B., 1990, *Classroom Testing*, Harlow, Longmann.

Inspection générale de l'éducation nationale, 2013, *Bilan de la mise en œuvre des programmes issus de la réforme de l'école primaire de 2008,* rapport n° 2013-066, p. 26-35.

LEGENDRE J., 2003, *L'enseignement des langues étrangères en France (2003-2004)*, Rapport d'information du Sénat, n° 63, p. 60.

MEN, 2007a, « Programmes de l'enseignement de langues vivantes étrangères au collège, palier 2 », *Bulletin officiel,* hors-série, n° 7.

MEN, 2007b, « Programmes de langues étrangères pour l'école primaire, mise en œuvre du cadre européen commun de référence pour les langues, mise en œuvre du socle commun de connaissances et de compétences », *Bulletin officiel*, hors-série n° 8.

MEN, 2006, « Enseignement des langues vivantes – Rénovation de l'enseignement des langues vivantes étrangères », *Bulletin officiel*, n° 23.

MEN, 2005, « Programme de l'enseignement des langues vivantes étrangères au palier 1 du collège », *Bulletin officiel*, hors-série, n° 6.

MEN, 2002, « Programme d'enseignement des langues étrangères ou régionales à l'école primaire », *Bulletin officiel*, hors-série, n° 4.

MENESR, 2014, « Échanges et actions de formation à l'étranger – année 2015-2016 », *Bulletin officiel*, n° 38.

PORTINE H., 2008, « Activités langagières, énonciation et cognition. La centration sur les apprentissages », *Recherches en didactique des langues – L'Alsace au cœur du plurilinguisme*, *Cahiers de l'ACEDLE,* vol. 5, n° 1, p. 109 à 128.

TARDIEU C., 2008, *La didactique des langues en 4 mots-clés : communication, culture, méthodologie, évaluation*, Paris, Ellipses, p. 41.

VANDERGRIFT L., 2002, "Listening: theory and practice in modern foreign language competence", *Subject Centre for Languages, Linguistics and Area Studies Good Practice Guide*. www.llas.ac.uk/resources/gpg/67

ÉVOLUTION DES ACQUIS COGNITIFS AU COLLÈGE AU REGARD DE L'ENVIRONNEMENT DE L'ÉLÈVE

Constat et mise en perspective longitudinale

Linda Ben Ali et Ronan Vourc'h
MENESR-DEPP, bureau de l'évaluation des élèves

Cet article s'appuie sur l'exploitation des données du panel d'élèves du second degré initié par la DEPP en 2007. Il s'intéresse aux évolutions des acquis cognitifs entre la sixième et la troisième en lien avec l'environnement socioculturel. Pour cela, sont mobilisés les résultats d'évaluations standardisées mises en œuvre lors de ces deux moments de mesure. L'objectif principal est de vérifier si les écarts sociaux qui peuvent être identifiés à la sortie de l'école primaire se creusent ou, au contraire, se réduisent pendant les années passées au collège. Lorsque l'on s'intéresse aux résultats des tests effectués en fin de troisième, il apparaît que la réussite dépend avant tout du niveau initial des collégiens – les élèves les plus compétents en sixième tendent à le rester en troisième – mais aussi de leurs caractéristiques sociales. En effet, les enfants de cadres et de diplômés de l'enseignement supérieur réussissent mieux tous les tests de fin de troisième que les enfants d'ouvriers ou ceux dont les parents sont dépourvus de diplôme. Quant aux analyses portant sur la progression des élèves en tenant compte de leur niveau initial, elles indiquent une stabilité des écarts sociaux, déjà marqués en sixième, pour les épreuves de compréhension et de raisonnement logique. En revanche, pour les compétences en mathématiques et en mémoire encyclopédique, les inégalités sociales se creusent entre la sixième et la troisième.

Dans la lignée des travaux entrepris dès les années 1960 dans la recherche en éducation sur les différences de réussite selon l'origine sociale, les analyses de suivi de cohortes menées à partir des données des différents panels de la DEPP (direction de l'évaluation, de la prospective et de la performance) ont privilégié l'étude des parcours scolaires. Ainsi, l'analyse des trajectoires d'élèves établies

grâce au panel d'élèves entrés en sixième en 1989 souligne combien celles-ci sont différenciées selon l'origine sociale [Lemaire, 2006 ; Caille et Lemaire, 2002] ou le niveau d'acquisition à l'entrée au collège [Caille, 1997]. L'effet de l'origine sociale sur les parcours se confirme et même s'amplifie au collège selon les travaux issus du suivi du panel d'élèves entrés en sixième en 1995 [Vanhoffelen, 2010 ; Lemaire, Guyon, Murat, 2007]. Dans ce contexte, l'évolution des acquis des élèves durant les années collège n'a pas souvent été abordée. On peut tout de même citer les travaux de Grisay [1997] et, plus récemment, ceux de Cayouette [2013] dont les résultats portant sur un échantillon localisé de collégiens indiquent des inégalités de progression selon l'origine sociale.

Pour la première fois, les données longitudinales issues du suivi du panel d'élèves du second degré recrutés en 2007 offrent la possibilité d'étudier l'évolution des acquis des élèves au collège en s'appuyant sur un échantillon représentatif au niveau national ▸ **Encadré**. En effet, les élèves entrés en sixième en 2007 ont participé, au moins à deux moments, à une épreuve d'évaluation standardisée mise en œuvre par la DEPP : la première fois en sixième, la seconde fois en troisième. À chaque fois, ces évaluations se sont déroulées à la fin de l'année scolaire concernée. Il s'agit de tests comparables dans le temps visant à évaluer la mémoire encyclopédique, la maîtrise phonologique, la compréhension de textes, les compétences en mathématiques et le raisonnement logique. Ces informations sont enrichies par l'apport de données détaillées portant sur les caractéristiques scolaires et familiales des élèves.

On peut dès lors s'intéresser à l'évolution des compétences de la sixième à la troisième en lien avec l'environnement social et chercher à voir si les inégalités se creusent pendant cette même période ou si elles sont figées dès l'entrée au collège. C'est ce que nous nous proposons d'étudier dans cet article. Pour cela, nous présenterons tout d'abord les instruments mobilisés pour l'analyse. Nous nous intéresserons ensuite aux facteurs explicatifs des scores obtenus aux tests en fin de troisième. Enfin, nous analyserons l'évolution des résultats aux tests standardisés des élèves entre la sixième et la troisième, en cherchant notamment à voir dans quelle mesure celle-ci est sensible aux caractéristiques sociales des élèves ainsi qu'à leur environnement culturel.

INSTRUMENTS MOBILISÉS POUR L'ANALYSE

Pour évaluer les compétences des collégiens du Panel 2007, on dispose, en plus des résultats scolaires, des données issues d'évaluations standardisées qui ont été mises en œuvre à trois reprises de façon à mesurer les évolutions des performances des élèves. Ce sont les résultats de ces évaluations qui sont mobilisés dans cet article. Elles ont été administrées dans des conditions similaires dans tous les établissements concernés. Le traitement des réponses des élèves était centralisé et standardisé, de manière à neutraliser tout biais de correction. Cette méthodologie permet de réduire les erreurs de mesure et l'influence d'éléments d'ordre contextuel [Bressoux et Pansu, 2003] qui peuvent affecter la notation des épreuves du brevet par exemple. La première prise d'information a eu lieu en mai 2008 et concerne l'ensemble des inscrits du panel en fin de sixième. Une nouvelle évaluation a eu lieu trois ans après, au printemps 2011, et ce quel que soit le niveau de

classe atteint. La plupart des élèves (près de 90 %) étaient alors en fin de troisième. Enfin, en 2012, les élèves de troisième ayant redoublé une fois pendant le collège ont de nouveau été évalués sur la base des mêmes épreuves que celles proposées en 2011. Dans cet article, la population étudiée est constituée des élèves du panel 2007 évalués en fin de sixième et en fin de troisième (qu'ils aient redoublé ou non) et dont les parents ont répondu à l'enquête famille en 2008[1]. Au total, près de 23 700 élèves sont concernés parmi les 35 000 élèves recrutés au départ dans le Panel 2007. Les données ont fait l'objet d'une repondération spécifique afin de tenir compte de cette attrition.

LE PANEL D'ÉLÈVES DE L'ENSEIGNEMENT DU SECOND DEGRÉ RECRUTÉS EN 2007

Le panel d'élèves du second degré recrutés en 2007 s'inscrit dans la tradition des panels mis en place par la DEPP depuis les années 1970. Son objectif principal est d'éclairer le système éducatif sur les parcours scolaires des élèves, leurs performances scolaires, le processus d'orientation des élèves, leur progression entre la sixième et la fin de la scolarité obligatoire.

Le panel étudié est un échantillon de 35 000 élèves entrés pour la première fois en classe de sixième en septembre 2007 dans un collège public ou privé sous contrat, en France métropolitaine ou dans un département d'outre-mer. Un tirage au hasard d'un entrant en sixième sur vingt-deux a été effectué dans les bases académiques afin de constituer une photographie représentative de l'ensemble de la population des nouveaux collégiens en septembre 2007. Les élèves scolarisés dans les établissements classés dans un réseau ambition réussite (RAR) ont été surreprésentés : un élève sur huit a été retenu.

La mise en œuvre d'un tel panel permet aussi de relier ces éléments de parcours au contexte d'enseignement ainsi qu'à des informations précises sur le milieu socio-économique et familial de l'élève, la représentation des parents et de l'élève sur sa scolarité et son devenir, l'environnement éducatif de l'élève dans et hors de l'école.

Pour cela, les familles des enfants sélectionnés dans l'échantillon ont répondu en 2008 à une enquête postale portant sur la situation scolaire antérieure des enfants, leurs attentes et leur implication par rapport à la scolarité de l'élève puis l'environnement familial. Cette enquête a été adaptée et soumise, dans les mêmes conditions, en 2011.

Les tests cognitifs proposés aux élèves lors des évaluations et dont les résultats sont utilisés dans cet article s'articulent autour des cinq épreuves suivantes[2] :

– **mémoire encyclopédique (Lexis)** : le test est issu de recherches montrant que la mémoire encyclopédique (corpus lexical et sémantiques des manuels de collège, hors vocabulaire courant) est très prédictive de la réussite scolaire [Lieury, 2012]. La mémoire encyclopédique est vue comme mesurant

1. Par construction, les redoublants de troisième évalués en 2011 n'ont pas été de nouveau interrogés en 2012. Ils ne sont pas pris en compte dans les analyses.
2. La durée totale des épreuves était d'environ trois heures. L'épreuve de phonologie proposée aux élèves n'a pas été retenue dans les analyses présentées dans cet article. En effet, les propriétés psychométriques des items ne permettaient pas la comparaison temporelle. Les items des tests de fin de sixième et ceux des tests de fin de troisième ne pouvaient pas être analysés sur une même échelle de difficulté.

des capacités fondamentales, comme la quantité de connaissances pouvant être apprises par un élève, ou le degré d'abstraction qui lui permet d'acquérir une grande variété de concepts d'un domaine donné ou des finesses sémantiques entre termes voisins. En s'appuyant sur le contenu des manuels scolaires, le test concerne le savoir enseigné l'année en cours en français, en mathématiques, en sciences et technologie, en sciences de la vie et de la Terre et en histoire-géographie. Il se présente sous la forme d'un QCM avec, pour chaque mot, une consigne simple « choisissez le mot le plus proche », afin d'éviter de donner des définitions trop longues. Le test comprend 48 items au total ;

– mathématiques : étant donné la grande diversité de domaines concernés, le test comprend 45 items répartis dans plusieurs situations correspondant à des épreuves variées des compétences mathématiques : calcul mental, problèmes, calculs d'horaires et d'unités, géométrie et logique. Par exemple, en calcul mental, il faut effectuer l'opération « 65-30 = [ ] » ;

– traitement de phrases lacunaires (TPL) : il s'agit d'une épreuve de compréhension basée sur la technique du texte à trous. Par exemple, il faut compléter la phrase *« Septembre ! C'est le mois ..... choisit l'hirondelle pour partir vers le sud du Sahara ..... elle peut passer l'hiver au chaud. »* Cette épreuve, qui comporte 20 items au total, met en jeu la mémoire sémantique, mais aussi la richesse du lexique ;

– lecture silencieuse (LS) : cette épreuve repose sur trois textes d'une cinquantaine de mots. Cinq questions sont posées après chaque texte, l'élève gardant le texte sous les yeux. Plus qu'un simple test de lecture, c'est un test de compréhension avec des questions nécessitant une inférence ;

– raisonnement sur cartes de Chartier (RCC) : cette épreuve vise à mesurer le raisonnement logique (type facteur G). L'élève doit trouver les caractéristiques (valeur et famille) d'une carte retournée afin qu'elle continue une suite proposée. Cette situation est composée de 30 exercices à résoudre en 20 minutes.

À l'exception de l'épreuve Lexis, les questions posées dans les évaluations ne portent pas sur les programmes d'une année donnée puisqu'elles peuvent concerner des élèves de niveaux scolaires différents comme ce fût le cas en 2011. De plus, pour appréhender la progression des performances dans le temps, il est nécessaire de disposer de résultats comparables. Pour ce faire, les épreuves proposées en 2011 et 2012 comprennent des items déjà utilisés en 2008 ainsi que de nouveaux items issus de mêmes familles de tests tenant compte de l'avancement en âge.

À l'issue des deux temps d'évaluation, des scores aux tests cognitifs ont été estimés selon le modèle de réponse à l'item [Rocher, dans ce numéro, p. 37]. Le score moyen en fin de sixième a été fixé par construction à 0 et l'écart-type à 1. Les scores de fin de troisième sont construits sur l'échelle des scores de sixième en conservant les paramètres du modèle pour les items repris à l'identique. Autrement dit, les scores de l'évaluation en fin de collège sont convertis dans la métrique des tests de sixième. Un des avantages de cette modélisation est d'assurer la comparabilité entre les deux années de référence. La progression au collège est mesurée en faisant la différence entre le score de troisième et celui de sixième.

Les évaluations standardisées permettent d'évaluer des dimensions cognitives variées. Certes, celles-ci sont liées les unes aux autres mais pas suffisamment pour justifier la création d'un score unique décrivant à lui seul les performances des

élèves. En fin de troisième, les coefficients de corrélation entre les scores issus des tests standardisés se situent tous autour de 0,6, voire même en dessous pour ce qui concerne l'épreuve de raisonnement sur cartes à jouer ▸ Tableau 1[3]. Celle-ci est moins corrélée aux autres épreuves à l'exception des mathématiques. Elle relève du facteur général d'intelligence (facteur g) et plus précisément de l'intelligence fluide (capacité générale à établir des relations entre des éléments), considérée comme étant relativement indépendante des connaissances acquises [Chartier, 2012]. L'analyse de ces corrélations souligne donc l'intérêt d'étudier isolément les dimensions couvertes par les évaluations afin de voir notamment si elles sont toutes sensibles aux mêmes variables de contexte.

▸ Tableau 1 Corrélations entre les scores des évaluations standardisées de fin de troisième

Scores standardisés en troisième	Traitement de phrases lacunaires (TPL)	Mathématiques	Lecture silencieuse (LS)	Mémoire encyclopédique (Lexis)	Raisonnement sur cartes à jouer (RCC)
Traitement de phrases lacunaires (TPL)	**1,00**	–	–	–	–
Mathématiques	**0,62**	**1,00**	–	–	–
Lecture silencieuse (LS)	**0,57**	0,50	**1,00**	–	–
Mémoire encyclopédique (Lexis)	**0,60**	**0,59**	**0,55**	**1,00**	–
Raisonnement sur cartes à jouer (RCC)	0,46	**0,58**	0,43	0,43	**1,00**

Lecture : le coefficient de corrélation entre les scores moyens en lecture silencieuse et en raisonnement sur cartes à jouer est de 0,43.
Source : MENESR-DEPP.

ANALYSE DES DÉTERMINANTS DU NIVEAU EN TROISIÈME

Des performances inégales selon l'environnement de l'élève

Une première approche descriptive des résultats obtenus aux évaluations standardisées passées par les élèves du Panel 2007 en fin de classe de troisième permet de constater que les performances varient selon l'origine sociale et l'environnement culturel. Ainsi, environ un tiers des enfants d'ouvriers ne dépassent pas les premiers quartiles de scores aux évaluations standardisées, c'est-à-dire figurent parmi le quart des élèves qui réussissent le moins bien ▸ Tableau 2. À titre de comparaison, cette proportion se situe autour de 10 % chez les enfants de cadres et professions intellectuelles supérieures pour toutes les épreuves sauf en raisonnement sur cartes à jouer où elle atteint 14,8 %.
Des différences sont aussi visibles lorsque l'on s'intéresse aux performances selon le niveau de diplôme du responsable de l'élève[4]. Plus le niveau de diplôme est élevé, plus la proportion d'élèves figurant dans les premiers quartiles de scores est faible.

3. Les coefficients de corrélation observés pour les évaluations standardisées passées en troisième sont comparables à ceux de sixième (figure non présentée).
4. Le responsable de l'élève est le père ou le conjoint de la mère. À défaut, il s'agit de la mère.

À titre d'exemple, pour l'épreuve standardisée de mathématiques, 9,4 % des élèves appartenant à une famille dont le responsable a effectué des études supérieures ne dépassent pas le premier quartile. Cette proportion s'élève à 44,7 % parmi les élèves dont le responsable est sans diplôme.
Il existe aussi des disparités de performances selon des variables liées à l'environnement culturel des élèves (tableau 2). Plus ils sont entourés de livres au quotidien, plus leurs performances sont élevées, quelle que soit la compétence concernée. Ainsi, pour l'épreuve visant à évaluer la mémoire encyclopédique, près des deux tiers des élèves déclarant ne pas avoir de livre à leur domicile obtiennent un score qui se situe dans le premier quartile. Pour les élèves qui disposent d'au moins 200 livres à leur domicile, cette proportion est de seulement 8,9 %. À l'inverse, le temps passé devant la télévision est lié négativement à la réussite, les performances des

Tableau 2 Proportions d'élèves qui figurent parmi le quart des élèves les plus faibles aux évaluations standardisées de fin de troisième (en %)

	Traitement de phrases lacunaires (TPL)	Mathématiques	Lecture silencieuse (LS)	Mémoire encyclopédique (Lexis)	Raisonnement sur cartes à jouer (RCC)
Catégorie socioprofessionnelle du responsable de l'élève					
Agriculteurs exploitants	22,7	18,2	23,4	21,2	18,8
Artisans, commerçants et chefs d'entreprise	22,7	21,8	23,9	22,9	23,4
Cadres et professions intellectuelles supérieures	11,1	9,2	12,6	9,5	14,8
Professions intermédiaires	19,5	20,8	21,4	19,6	22,9
Employés	30,1	33,1	28,1	29,9	30,5
Ouvriers qualifiés ou non-qualifiés	35,9	36,4	34,5	35,8	31,5
Sans profession	66,4	65,1	60,9	70,7	54,5
Diplôme du responsable de l'élève					
Aucun diplôme, CEP	43,9	44,7	39,9	45,1	38,0
BEPC	27,4	29,1	27,1	28,9	27,1
BEP ou CAP	28,9	30,0	28,9	27,9	27,2
Baccalauréat professionnel	17,4	16,6	20,4	17,2	21,7
Baccalauréat général ou technologique	21,2	23,4	23,5	22,2	22,2
Enseignement supérieur	11,1	9,4	12,9	9,7	15,4
Nombre de livres au domicile					
Aucun	63,4	58,5	55,4	62,0	50,4
de 1 à 29	45,0	44,4	40,7	45,3	38,6
de 30 à 99	27,5	27,9	27,7	27,2	26,6
de 100 à 199	17,3	17,0	19,3	16,9	19,4
200 et plus	10,4	10,9	12,1	8,9	15,2
Fréquence d'écoute de la télévision					
Régulièrement	27,3	27,4	27,1	26,3	25,5
De temps en temps	24,1	24,2	24,4	24,5	25,6
Presque jamais ou jamais	18,6	16,9	17,3	17,7	19,7

Lecture : 35,9 % des enfants d'ouvriers figurent parmi le quart des élèves qui présentent les scores les plus faibles à l'épreuve de traitement de phrases lacunaires.
Source : MENESR-DEPP.

élèves déclinant à mesure que la fréquence d'écoute augmente. Lorsque l'on prend en compte l'ensemble des épreuves, les élèves qui regardent le moins la télévision[5] sont, en moyenne, 17,3 % à figurer dans les premiers quartiles de scores. Alors qu'ils sont, en moyenne, 26,1 % parmi ceux qui la regardent régulièrement.

Le tableau 2 présente le lien entre les performances observées en fin de troisième et quelques variables prises séparément. Cependant, on peut se demander si les disparités observées selon ces variables relatives à l'origine sociale et à l'environnement culturel se maintiennent lorsque l'on raisonne sur des élèves rendus comparables au moyen de modèles de régression. En effet, les élèves qui présentent l'un ou l'autre de ces attributs favorables à la réussite possèdent également d'autres caractéristiques susceptibles d'être associées à de meilleures performances scolaires.
Pour chacune des compétences ayant fait l'objet d'une évaluation standardisée en fin de troisième, les scores ont donc été modélisés en fonction de certaines variables décrivant à la fois les caractéristiques individuelles de l'élève, son contexte de scolarisation ainsi que son environnement social et culturel ▸ **Tableau 3**. Le choix de ces variables, guidé par les problématiques soulevées dans cet article, repose aussi sur les apports de la littérature sur le sujet ainsi que sur les précédentes études réalisées sur le panel [par exemple, Caille, 1997]. De nombreuses variables ont tout d'abord été testées. Les plus pertinentes d'entre elles ont ensuite été retenues pour l'analyse et regroupées dans les catégories suivantes :

– **caractéristiques individuelles :** sexe, retard scolaire au collège[6] ;
– **contexte scolaire :** secteur de l'établissement (public ou privé) et appartenance à une zone d'éducation prioritaire ;
– **contexte social et familial :** catégorie sociale appréhendée par l'indice de position sociale du responsable de l'élève[7], diplôme du responsable de l'élève, lieu de naissance des parents, langue parlée par le responsable de l'élève, taille de la fratrie et structure de la famille ;
– **environnement culturel et matériel :** nombre d'activités extra-scolaires (inscription à un club de sport, à des cours de théâtre, au conservatoire, etc.), environnement matériel (ordinateur familial, ordinateur personnel, accès à Internet), fréquence d'écoute de la télévision, nombre de livres au domicile, partage ou non de la chambre de l'enfant.

La part de variance expliquée des scores de troisième par les variables présentes dans les modèles se situe autour de 30 % pour toutes les compétences à l'exception de l'épreuve de logique où elle s'élève à 13 %. En contrôlant les dimensions sociodémographiques et scolaires, on constate tout d'abord que la performance des élèves en fin de collège croît en fonction de l'indice de position sociale du parent responsable.

5. Dans le questionnaire famille, on demande à l'élève s'il regarde « régulièrement », « de temps en temps » ou « jamais ou presque jamais » la télévision à différents moments de la journée. S'il répond au moins une fois « régulièrement », il entre dans la première catégorie. Sinon, en répondant au moins une fois « de temps en temps », il entre dans la deuxième catégorie. À défaut, il est inclus dans la troisième catégorie.
6. Sont considérés « en retard » les élèves qui ont plus de 14 ans au 31 décembre de l'année de leur entrée en troisième.
7. L'indice de position sociale [Le Donné et Rocher, 2010] est construit à partir de plusieurs variables « mesurant la proximité au système scolaire du milieu familial de l'enfant » : caractéristiques sociales des parents, conditions de vie matérielles et financières, pratiques culturelles de l'enfant et de sa famille, implication des parents dans la scolarité, etc. Cet indice permet de mesurer la position socio-scolaire des élèves et peut se substituer à la PCS des parents dans le cadre d'études statistiques. De manière agrégée (niveau classe ou établissement par exemple), il permet d'appréhender le profil social de la structure étudiée. Dans ce cas-là, on parle d'indice de position sociale moyen.

Tableau 3 Relation entre les scores en troisième et les facteurs sociaux, scolaires et culturels (coefficients estimés)

Constante	
Retard scolaire *réf. = « À l'heure » en troisième*	En retard en troisième
Sexe *réf. = Fille*	Garçon
Secteur de l'établissement *réf. = Public hors éducation prioritaire*	Public en éducation prioritaire
	Privé
Indice social du responsable de l'élève	Indice social du parent responsable
Diplôme du responsable de l'élève *réf. = Enseignement supérieur*	Aucun diplôme, CEP
	BEPC
	BEP ou CAP
	Baccalauréat général ou technologique
	Baccalauréat professionnel
Structure de la famille *réf. = Famille nucléaire*	Garde alternée
	Monoparentale
	Recomposée
	Autres situations
Taille de la fratrie *réf. = Enfant unique*	1 frère ou 1 sœur
	2 frères ou sœurs
	3 frères ou sœurs
	Plus de 3 frères ou sœurs
Lieu de naissance des parents *réf. = Parents nés en France*	Un des deux parents né à l'étranger
	Parents nés à l'étranger
Langue parlée du responsable de l'élève *réf. = Français uniquement*	Autre langue uniquement
	Souvent le français, parfois une autre langue
	Souvent une autre langue, parfois le français
Activités extrascolaires *réf. = Trois activités ou plus*	Moins de deux activités
	Deux activités
Fréquence d'écoute de la télévision *réf. = Jamais ou presque jamais*	Régulièrement
	De temps en temps
Environnement matériel (ordinateur de famille, ordinateur personnel, Internet) *réf. = Plus de deux matériels*	Aucun
	Un matériel
	Deux matériels
Possession de livres au domicile *réf. = 200 et plus*	Aucun
	de 1 à 29
	de 30 à 99
	de 100 à 199
Chambre seul *réf. = Oui*	Non

*** significatif au seuil de 1 %, ** significatif au seuil de 5 %, *significatif au seuil de 10 %,
n.s. : non significatif au seuil de 10 %.

Traitement de phrases lacunaires (TPL)	Raisonnement sur cartes à jouer (RCC)	Lecture silencieuse (LS)	Mathématiques	Mémoire encyclopédique (Lexis)
0,91***	0,34***	0,58***	0,48***	0,78***
0,47***	- 0,45***	- 0,35***	- 0,5***	- 0,35***
- 0,22***	n.s.	- 0,08***	0,38***	0,24***
- 0,21***	- 0,13***	- 0,14***	- 0,22***	- 0,14***
0,12***	0,11***	0,12***	0,19***	0,17***
0,11***	0,03*	0,1***	0,09***	0,1***
- 0,32***	- 0,19***	- 0,22***	- 0,36***	- 0,37***
- 0,15***	- 0,1*	- 0,14***	- 0,25***	- 0,27***
- 0,22***	- 0,11***	- 0,14***	- 0,28***	- 0,27***
- 0,08*	n.s	- 0,06*	- 0,11***	- 0,11***
- 0,15***	- 0,08*	- 0,1*	- 0,19***	- 0,21***
0,13***	0,1*	0,13*	0,1*	0,07*
0,06*	n.s.	n.s.	- 0,04*	n.s.
n.s.	n.s.	n.s.	- 0,07*	- 0,05*
n.s.	- 0,16*	n.s.	- 0,27***	- 0,1*
- 0,04*	0,04*	n.s.	0,04*	- 0,12***
- 0,07*	0,06*	n.s.	n.s.	- 0,18***
- 0,12***	n.s.	- 0,07*	n.s.	- 0,23***
- 0,13***	n.s.	- 0,05*	n.s.	- 0,25***
- 0,05*	- 0,13***	- 0,05*	n.s.	- 0,11***
n.s.	- 0,09*	n.s	n.s.	n.s.
- 0,26***	n.s.	- 0,16*	n.s.	- 0,32***
- 0,05*	- 0,05*	- 0,07*	- 0,09*	- 0,06*
- 0,15***	n.s.	- 0,14*	n.s	- 0,17***
- 0,14***	- 0,1***	- 0,14***	- 0,16***	- 0,15***
- 0,06***	- 0,03*	- 0,07***	- 0,08***	- 0,07***
- 0,1***	n.s	n.s	n.s	- 0,05*
- 0,09*	n.s	- 0,07*	- 0,05*	- 0,08*
- 0,15***	- 0,1*	- 0,13***	- 0,15***	- 0,05*
n.s.	n.s.	n.s.	n.s.	0,06*
n.s.	n.s.	n.s.	n.s.	n.s.
- 0,55***	- 0,36***	- 0,51***	- 0,39***	- 0,69***
- 0,49***	- 0,28***	- 0,42***	- 0,4***	- 0,6***
- 0,29***	- 0,13***	- 0,26***	- 0,22***	- 0,38***
- 0,15***	- 0,08*	- 0,14***	- 0,12***	- 0,21***
0,05*	0,04*	0,07***	0,04*	0,06*

Lecture : le tableau présente les résultats de régressions linéaires de différentes caractéristiques sur les scores observés en troisième. En mathématiques, à caractéristiques comparables présentes dans le modèle, les garçons obtiennent de meilleurs résultats que les filles au test de mathématiques de fin de troisième puisque le coefficient estimé est positif (0,38) et significatif.
Source : MENESR-DEPP.

Les coefficients sont les plus élevés pour les compétences en mathématiques et celles relatives à la mémoire encyclopédique. Il en va de même lorsque l'on regarde les performances selon le diplôme du responsable de l'élève : les scores aux évaluations augmentent avec le niveau du diplôme, et ce même après avoir fixé les autres variables introduites dans le modèle.

La variation des acquis des collégiens s'explique aussi, en partie, par le pays de naissance des parents et la langue parlée à la maison. Ainsi, « toutes choses égales par ailleurs », les élèves dont les deux parents sont nés à l'étranger réussissent moins bien que les autres dans les épreuves mesurant la mémoire encyclopédique et la compréhension. Quelles que soient les compétences évaluées, les résultats sont aussi sensiblement moins favorables lorsque le parent responsable de l'élève ne parle pas uniquement le français au quotidien avec ses enfants.

Outre les caractéristiques des parents, la composition de la famille apporte des conditions plus ou moins favorables à la réussite à l'élève. Être un enfant unique semble constituer une condition propice pour l'acquisition de compétences telles que la mémoire encyclopédique et le traitement de phrases lacunaires. Ce n'est pas le cas, en revanche, pour les mathématiques et le raisonnement logique.

Les performances scolaires de l'enfant sont également liées aux activités effectuées en dehors des « bancs de l'école », telles que la lecture d'ouvrages, les sorties au théâtre ou au cinéma, la pratique d'un sport, etc. Le milieu familial et l'occupation du temps de l'élève en dehors du cadre scolaire sont très liés, créant un contexte plus ou moins propice à la réussite scolaire [O'Prey, 2004]. Les familles dans lesquelles les conditions sont réunies pour une meilleure réussite des enfants répondent à des exigences sociales et culturelles établies par l'école. On retrouve ici la notion de distance à la culture scolaire mise en lumière dans la littérature sociologique classique [Bourdieu et Passeron, 1964] ainsi que dans les travaux menés ces dernières années autour des pratiques culturelles des jeunes générations [Octobre, Détrez *et alii*, 2010]. Sur cet aspect, les modèles présentés dans le tableau 3 confirment les premières analyses descriptives. Ainsi, à caractéristiques comparables, plus les élèves possèdent de livres à leur domicile, plus leurs performances sont favorables, quelle que soit la compétence concernée. Pratiquer des activités extra-scolaires nombreuses constitue aussi un élément favorisant les performances scolaires. En revanche, cette analyse montre que l'effet lié au temps passé devant la télévision s'atténue lorsque les autres variables sont tenues constantes. Ce résultat indique que l'effet négatif de la télévision sur les scores identifié dans le tableau 2 viendrait surtout de l'association entre ce loisir et des caractéristiques du milieu familial.

Des performances contrastées selon le contexte scolaire

Les résultats des évaluations standardisées menées par la DEPP ont déjà mis en avant les différences de performances entre les élèves selon le secteur de scolarisation. Ainsi, la proportion d'élèves qui maîtrise les compétences 1 (maîtrise de la langue française) et 3 (principaux éléments de mathématiques et de la culture scientifique et technologique) du socle commun est significativement plus élevée

dans le secteur privé que dans le secteur public, aussi bien en fin d'école qu'en fin de collège. Par ailleurs, au sein du secteur public, les résultats sont moins favorables dans les zones d'éducation prioritaire [MENESR, 2014]. Le tableau 3 confirme ces résultats selon lesquels un accueil dans le secteur privé favoriserait une meilleure performance au collège. Au sein du secteur public, il montre aussi que les performances sont moins élevées dans les établissements relevant de l'éducation prioritaire. Néanmoins, ce constat est à nuancer car il intègre des effets d'ordres sociaux. Ainsi, lorsque l'on tient compte de variables d'ordres socioculturelles, comme le niveau social moyen des élèves de l'établissement, l'effet net du secteur de scolarisation a tendance à se réduire.

Le retard pris par l'élève pendant son parcours scolaire est un élément majeur dans l'interprétation des résultats obtenus aux évaluations standardisées de fin de troisième. Il s'agit de la variable dont les coefficients sont parmi les plus élevés quelle que soit la compétence concernée (tableau 3). Être en retard en troisième est lié négativement aux scores obtenus aux évaluations. La situation est plus défavorable parmi les élèves qui ont redoublé avant l'entrée au collège que parmi ceux qui ont redoublé uniquement au collège ▸ **Tableau 4**.

▸ Tableau 4 Proportions d'élèves qui figurent parmi le quart des élèves les plus faibles aux évaluations standardisées de fin de troisième selon le retard scolaire (en %)

	Traitement de phrases lacunaires (TPL)	Mathématiques	Lecture silencieuse (LS)	Mémoire encyclopédique (Lexis)	Raisonnement sur cartes à jouer (RCC)
Retard à l'entrée au collège	61,8	63,8	55,3	57,9	50,2
Retard au collège uniquement	39,6	45,7	41,8	38,4	41,4
« À l'heure » en troisième	15,9	14,5	16,7	16,5	17,6

Lecture : 61,8 % des élèves qui étaient en retard à l'entrée au collège figurent parmi le quart des élèves qui présentent les scores les plus faibles à l'épreuve de traitement de phrases lacunaires.
Source : MENESR-DEPP.

Détermination des scores en troisième selon les scores en sixième

À ce stade de notre étude, nous retrouvons les éléments de différenciation des performances, observés à un instant donné de la scolarité, de façon cohérente avec d'autres études de la DEPP déjà évoquées. Cependant, le niveau des acquis des élèves évalués en fin de sixième n'apparaît pas dans la modélisation. Or, les performances « initiales » constituent très certainement un critère majeur dans la détermination des performances en troisième.
Pour faire état du niveau à l'entrée du collège, nous utilisons, pour chaque dimension, les déciles des scores obtenus aux évaluations standardisées passées en classe de sixième. Cette information vient s'ajouter aux variables déjà présentes dans les modèles précédents ▸ **Tableau 5**. Le premier décile, représentant les 10 % d'élèves ayant les scores les plus élevés, est retenu comme étant la valeur de référence du niveau de l'élève au début du collège pour la compétence traitée. Dans cette nouvelle approche, la part de variance des scores en troisième expliquée par les variables présentes dans les modèles est plus élevée quelle que soit la compétence. Elle

Tableau 5 Relation entre les scores en troisième, les facteurs environnementaux et les scores en sixième

Constante	
Retard scolaire *réf. = « À l'heure » en troisième*	En retard en troisième
Sexe *réf. = Fille*	Garçon
Secteur de l'établissement *réf. = Public hors éducation prioritaire*	Public en éducation prioritaire
	Privé
Indice social du responsable de l'élève	Indice social du parent responsable
Diplôme du responsable de l'élève *réf. = Enseignement supérieur*	Aucun diplôme, CEP
	BEPC
	BEP ou CAP
	Baccalauréat général ou technologique
	Baccalauréat professionnel
Structure de la famille *réf. = Famille nucléaire*	Garde alternée
	Monoparentale
	Recomposée
	Autres situations
Taille de la fratrie *réf. = Enfant unique*	1 frère ou 1 sœur
	2 frères ou sœurs
	3 frères ou sœurs
	Plus de 3 frères ou sœurs
Lieu de naissance des parents *réf. = Parents nés en France*	Un des deux parents né à l'étranger
	Parents nés à l'étranger
Langue parlée du responsable de l'élève *réf. = Français uniquement*	Autre langue uniquement
	Souvent le français, parfois une autre langue
	Souvent une autre langue, parfois le français
Activités extrascolaires *réf. = Trois activités ou plus*	Moins de deux activités
	Deux activités
Fréquence d'écoute de la télévision *réf. = Jamais ou presque jamais*	Régulièrement
	De temps en temps
Environnement matériel (ordinateur de famille, ordinateur personnel,Internet) *réf. = Plus de deux matériels*	Aucun
	Un matériel
	Deux matériels
Possession de livres au domicile *réf. = 200 et plus*	Aucun
	de 1 à 29
	de 30 à 99
	de 100 à 199
Chambre seul *réf. = Oui*	Non
Scores aux épreuves spécifiques en sixième *réf. = Dernier décile 10 % les plus forts*	1ER décile
	2e décile
	3e décile
	4e décile
	5e décile
	6e décile
	7e décile
	8e décile
	9e décile

*** significatif au seuil de 1 %, ** significatif au seuil de 5 %, *significatif au seuil de 10 %,
n.s. : non significatif au seuil de 10 %

Traitement de phrases lacunaires (TPL)	Raisonnement sur cartes à jouer (RCC)	Lecture silencieuse (LS)	Mathématiques	Mémoire encyclopédique (Lexis)
1,43***	0,84***	0,98***	1,37***	1,5***
- 0,08***	- 0,18***	- 0,11***	0,06***	n.s
- 0,13***	0,05***	n.s	0,15***	0,04***
- 0,12***	- 0,06*	- 0,1***	- 0,09***	- 0,06***
0,07***	0,09***	0,1***	0,14***	0,11***
0,05***	0,02*	0,08***	0,03*	0,04***
- 0,14***	- 0,11***	- 0,13***	- 0,12***	- 0,18***
- 0,05*	- 0,06*	- 0,1*	- 0,09***	- 0,14***
- 0,08***	- 0,06*	- 0,07*	- 0,08***	- 0,12***
n.s.	n.s.	n.s.	- 0,04*	- 0,06*
- 0,05*	- 0,05*	n.s	- 0,04*	- 0,08*
0,08*	0,06*	0,09*	n.s	n.s.
n.s.	n.s.	n.s.	- 0,06*	n.s.
n.s.	n.s.	n.s.	- 0,05*	n.s.
- 0,09*	n.s.	- 0,12*	- 0,22***	- 0,12*
n.s	n.s	n.s	n.s.	- 0,06*
- 0,03*	n.s	n.s	n.s.	- 0,06*
- 0,06*	n.s	- 0,05*	n.s.	- 0,1***
- 0,05*	n.s	n.s	n.s.	- 0,09***
- 0,04*	- 0,09*	- 0,05*	n.s.	- 0,04*
n.s.	- 0,06*	n.s.	n.s.	n.s.
- 0,11*	n.s.	- 0,1*	- 0,08*	- 0,09*
n.s.	n.s.	- 0,06*	n.s.	n.s.
- 0,05*	n.s.	- 0,11*	n.s.	n.s.
- 0,06***	- 0,05*	- 0,1***	- 0,04***	- 0,06***
n.s.	n.s.	- 0,05*	- 0,03*	- 0,03*
- 0,06*	n.s.	n.s.	n.s.	n.s.
- 0,04*	n.s.	- 0,04*	n.s.	n.s.
- 0,12***	- 0,08*	- 0,11***	- 0,1***	- 0,06*
n.s.	0,04*	n.s.	n.s.	n.s.
n.s.	n.s.	n.s.	n.s.	n.s.
- 0,22***	- 0,22***	- 0,32***	- 0,12*	- 0,34***
- 0,2***	- 0,16***	- 0,26***	- 0,12***	- 0,24***
- 0,11***	- 0,06*	- 0,17***	- 0,06***	- 0,14***
- 0,05*	- 0,05*	- 0,09***	- 0,03*	- 0,08***
0,02*	0,03*	0,06*	n.s.	0,03*
- 2,11***	- 1,62***	- 1,35***	- 2,53***	- 2,21***
- 1,58***	- 1,18***	- 1,07***	- 2,08***	- 1,87***
- 1,44***	- 0,98***	- 0,92***	- 1,79***	- 1,68***
- 1,23***	- 0,83***	- 0,84***	- 1,58***	- 1,49***
- 1,11***	- 0,72***	- 0,75***	- 1,42***	- 1,31***
- 0,96***	- 0,59***	- 0,65***	- 1,23***	- 1,15***
- 0,82***	- 0,5***	- 0,56***	- 1,04***	- 0,99***
- 0,65***	- 0,38***	- 0,42***	- 0,78***	- 0,78***
- 0,42***	- 0,24***	- 0,25***	- 0,52***	- 0,5***

Lecture : le tableau présente les résultats de régressions linéaires de différentes caractéristiques sur les scores observés en troisième. À caractéristiques comparables, les garçons obtiennent de meilleurs résultats que les filles au test de mathématiques de fin de troisième puisque le coefficient estimé est positif (0,15) et significatif.
Source : MENESR-DEPP.

gagne jusqu'à 30 points notamment pour les compétences en mathématiques et celles relatives à la mémoire encyclopédique où elle dépasse 60 %. Pour rappel, le pouvoir explicatif des premiers modèles relatifs à ces compétences était respectivement de 29 % et 28 %. L'influence des scores à l'entrée au collège est donc plus forte que celle des autres variables du modèle et explique en grande partie la variabilité des résultats aux évaluations de troisième.

Dans ce modèle, il apparaît aussi que le niveau de compétences observé en sixième capte en partie l'influence du milieu familial de l'élève et de son capital social qui ont déjà joué leur rôle avant l'entrée au collège. On retrouve ici des mécanismes déjà identifiés lors de l'étude des progressions des élèves [voir notamment Caille et Rosenwald, 2006 ; Duru-Bellat, Jarousse, Minguat, 1993]. Ainsi, pour toutes les compétences évaluées, le coefficient de l'indice de position sociale du responsable de l'élève baisse lorsque l'on prend en compte le score des évaluations passées en sixième. Par ailleurs, le diplôme du responsable de l'élève ne joue plus un rôle aussi important : pour la compétence en mémoire encyclopédique, le fait d'avoir un parent sans diplôme joue deux fois moins sur la réussite de l'élève en troisième quand son niveau en sixième est pris en compte. L'influence de facteurs relatifs à l'environnement familial est, elle aussi, amoindrie. C'est le cas notamment des variables concernant l'origine des parents et la langue parlée par le responsable de l'élève ou encore la taille de la fratrie pour lesquelles on observe une baisse des coefficients associés.

Dans le même temps, l'ajout du niveau à l'entrée au collège réduit l'impact de l'environnement culturel et matériel sur les acquis des élèves en fin de collège. À titre d'exemple, le capital culturel mesuré par le nombre de livres présents au domicile joue beaucoup moins sur le score en troisième. En effet, les coefficients associés à cette caractéristique baissent de façon significative. C'est le cas plus particulièrement pour les compétences en mathématiques et en mémoire encyclopédique.
Ainsi, la prise en compte du niveau initial dans les modèles explicatifs des scores aux évaluations standardisées de fin de troisième réduit le poids associé aux autres variables présentes dans les modèles. Ces résultats tendent à valider l'hypothèse selon laquelle les écarts entre les catégories sociales seraient fixés en grande partie à l'entrée en sixième. Dès lors, on peut se demander si les facteurs sociaux ont malgré tout une influence lorsque l'on s'intéresse à l'évolution des performances des élèves au collège.

ÉVOLUTION DES SCORES ENTRE LA SIXIÈME ET LA TROISIÈME

Chaque compétence a été évaluée à la fois en sixième et en troisième, une sélection d'items communs permettant d'assurer la comparabilité des scores entre les deux moments de passation. Il est donc possible d'estimer l'ampleur des progressions des élèves dans les compétences évaluées et de les rapprocher des dimensions scolaires, sociales et culturelles. Le but de cet exercice est de vérifier si les écarts se creusent selon l'environnement de l'élève ou s'ils sont compensés durant le parcours au collège.

La question de l'influence de la dimension sociale apparaît au regard du croisement entre les progressions des acquis au collège et le statut social du responsable de l'élève. La différence de scores selon la catégorie sociale est ainsi présentée sur la **figure 1** qui compare l'évolution des scores des élèves en fonction de leur origine sociale[8]. Ce graphique permet de distinguer deux catégories de compétences. La première concerne trois des cinq épreuves (traitement de

Figure1 Évolution des scores entre la sixième et la troisième selon la catégorie sociale du responsable de l'élève

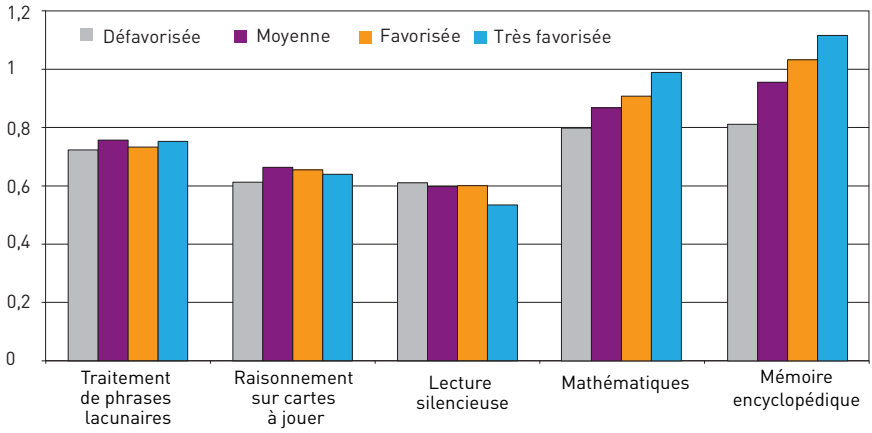
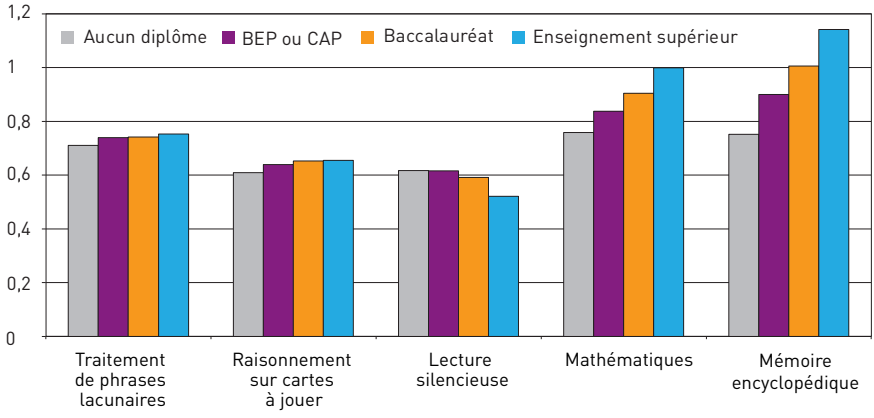


Figure 2 Évolution des scores entre la sixième et la troisième selon le niveau de diplôme du responsable de l'élève

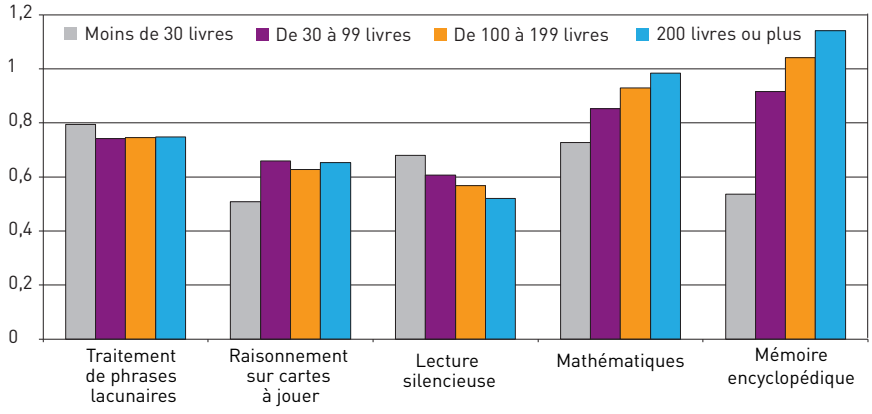


Lecture des figures 1 et 2 : les figures représentent la progression des scores moyens dans les différentes compétences évaluées selon des variables caractérisant l'environnement de l'élève. Par exemple, en mémoire encyclopédique, la progression est de 1,1 point pour les enfants d'origine sociale très favorisée contre 0,8 point pour les enfants d'origine sociale défavorisée.
Source : MENESR-DEPP.

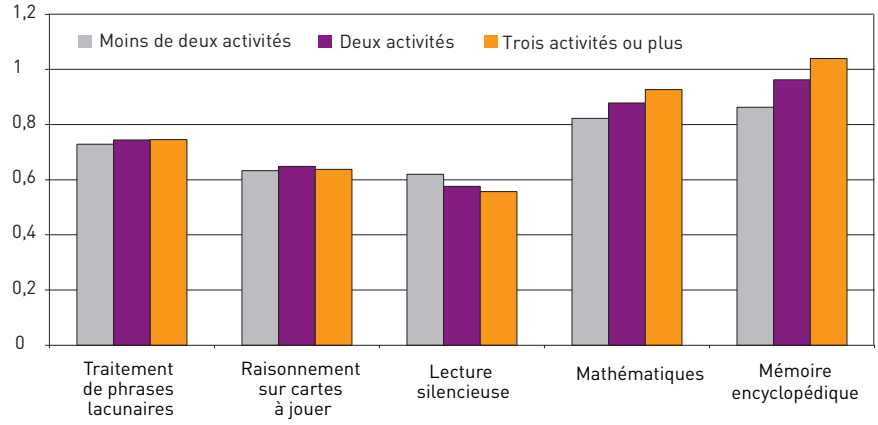
8. Un regroupement des catégories socioprofessionnelles du responsable de l'élève a été effectué de la façon suivante : très favorisées (cadres, professions intellectuelles supérieures, chefs d'entreprise de plus de dix salariés et enseignants), favorisées (professions intermédiaires), moyennes (agriculteurs exploitants, artisans, commerçants, employés) et défavorisées (ouvriers et inactifs).

phrases lacunaires, lecture silencieuse et raisonnement sur cartes à jouer) pour lesquelles les élèves semblent progresser à un degré comparable qu'ils soient enfants d'origine social favorisée ou défavorisée. Pour l'épreuve de lecture silencieuse, la progression est même sensiblement plus élevée parmi les enfants les plus défavorisés socialement. La seconde catégorie comprend deux épreuves (mathématiques et mémoire encyclopédique) pour lesquelles on observe des écarts de progression significatifs selon l'origine sociale. Le diplôme du responsable de l'enfant et des variables relatives à l'environnement culturel (nombre d'activités extra-scolaires, nombre de livres au domicile) exercent aussi une influence sur la progression des élèves au collège ▸ **Figures 2, 3 et 4**. Ici aussi, les deux mêmes catégories de compétences se distinguent.

▸ Figure 3 Évolution des scores entre la sixième et la troisième selon le nombre de livres au domicile de l'élève



▸ Figure 4 Évolution des scores entre la sixième et la troisième selon le nombre d'activités extra-scolaires



Lecture des figures 3 et 4 : les figures représentent la progression des scores moyens dans les différentes compétences évaluées selon des variables caractérisant l'environnement de l'élève. Par exemple, en mémoire encyclopédique, la progression est de 1,1 point pour les enfants qui disposent de plus de 200 livres contre 0,5 point pour les enfants qui en disposent moins de 30.
Source : MENESR-DEPP.

Pour certaines compétences, des progressions comparables entre la sixième et la troisième

Pour les trois épreuves pour lesquelles on observe des progressions similaires entre la sixième et la troisième, on peut considérer que, malgré les disparités de niveau induites par l'appartenance à un milieu social, le collège donnerait les mêmes chances aux élèves de maîtriser les éléments de connaissances requises à son issue. L'autre hypothèse peut tenir au maintien du niveau scolaire des élèves, quelle que soit leur origine sociale. Les élèves les plus forts/faibles resteraient au niveau le plus élevé/faible. Les méthodes de transmission des savoirs et des savoir-faire ne détérioreraient pas, pendant au moins quatre ans de scolarisation dans le secondaire, le degré d'acquisition par rapport à une situation de référence. En effet, les scores des enfants ont évolué de manière analogue quelle que soit l'origine sociale (figure 1) : progression de près de 0,6 point en compréhension de texte (lecture silencieuse) et en raisonnement sur cartes à jouer et de près de 0,7 point en traitement de phrases lacunaires. Pour ces compétences, les apprentissages ne permettraient pas de résorber des écarts dus au marquage social mais, dans le même temps, ils n'aggraveraient pas les différences.

Pour confirmer ces hypothèses, le travail de modélisation effectué précédemment est repris ici ▸ **Tableau 6 p. 228**. La variable d'intérêt porte maintenant sur l'évolution des scores entre la sixième et la troisième. Pour les trois compétences sus-citées, les variables de contexte social, culturel et scolaire ont un pouvoir explicatif très faible sur les évolutions de scores. À elles seules, ces variables n'ont plus ce rôle « prédicteur » sur les évolutions des acquis des collégiens.
On constate ainsi que les coefficients associés aux dimensions sociales ne sont pas significatifs ou sont très faibles (inférieurs à 0,1). Ceci est particulièrement vrai pour l'épreuve de raisonnement logique et celle de traitement de phrases lacunaires. Concernant l'épreuve de lecture silencieuse, on observe cependant un effet modéré de l'origine sociale et du niveau de diplôme du responsable de l'élève. Les coefficients relatifs à cette dernière variable indiquent même une progression plus prononcée des performances des élèves dont les parents ont un niveau de diplôme inférieur ou égal au baccalauréat.

Les variables caractérisant l'environnement culturel de l'élève ont un poids relativement négligeable dans l'explication de l'évolution des scores au collège. En revanche, les conditions de scolarisation semblent encore interférer sur la progression des scores de ces évaluations. La différenciation la plus marquée concerne le raisonnement sur cartes à jouer entre les établissements en zone d'éducation prioritaire et les autres (établissements publics hors éducation prioritaire et établissements privés). Paradoxalement, les coefficients associés au retard scolaire à l'entrée en troisième sont positifs. L'hypothèse principale permettant d'expliquer ce résultat repose sur un effet d'apprentissage. Les élèves ayant connu un redoublement au collège avant la troisième ont été évalués deux fois, en 2011 puis en 2012, sur la base des mêmes épreuves. C'est cette dernière évaluation qui a été prise en compte dans la construction des scores. Or, si l'on retire ces élèves du modèle, les coefficients liés au retard scolaire deviennent négatifs pour toutes les compétences à l'exception de celle évaluée par l'épreuve de lecture silencieuse.

Tableau 6 Relation entre l'évolution des scores au collège et les facteurs sociaux, scolaires et culturels (coefficients estimés)

Constante	
Retard scolaire *réf. = « À l'heure » en troisième*	En retard en troisième
Sexe *réf. = Fille*	Garçon
Secteur de l'établissement *réf. = Public hors éducation prioritaire*	Public en éducation prioritaire
	Privé
Indice social du responsable de l'élève	Indice social du parent responsable
Diplôme du responsable de l'élève *réf. = Enseignement supérieur*	Aucun diplôme, CEP
	BEPC
	BEP ou CAP
	Baccalauréat générale ou technologique
	Baccalauréat professionnel
Structure de la famille *réf. = Famille nucléaire*	Garde alternée
	Monoparentale
	Recomposée
	Autres situations
Taille de la fratrie *réf. = Enfant unique*	1 frère ou 1 sœur
	2 frères ou sœurs
	3 frères ou sœurs
	Plus de 3 frères ou sœurs
Lieu de naissance des parents *réf. = Parents nés en France*	Un des deux parents né à l'étranger
	Parents nés à l'étranger
Langue parlée du responsable de l'élève *réf. = Français uniquement*	Autre langue uniquement
	Souvent le français, parfois une autre langue
	Souvent une autre langue, parfois le français
Activités extrascolaires *réf. = Trois activités ou plus*	Moins de deux activités
	Deux activités
Temps consacré à la télévision *réf. = Jamais ou presque jamais*	Régulièrement
	De temps en temps
Environnement matériel (ordinateur de famille, ordinateur personnel, internet) *réf. = Plus de deux matériels*	Aucun
	Un matériel
	Deux matériels
Possession de livres au domicile *réf. = 200 et plus*	Aucun
	de 1 à 29
	de 30 à 99
	de 100 à 199
Chambre seul *réf. = Oui*	Non

*** significatif au seuil de 1 %, ** significatif au seuil de 5 %, *significatif au seuil de 10 %,
n.s. : non significatif au seuil de 10 %.

Traitement de phrases lacunaires (TPL)	Raisonnement sur cartes à jouer (RCC)	Lecture silencieuse (LS)	Mathématiques	Mémoire encyclopédique (Lexis)
0,82***	0,58***	0,32***	0,92***	1,22***
0,03*	0,03*	0,2***	- 0,11***	- 0,1***
- 0,07***	0,09***	0,11***	0,13***	n.s.
- 0,03*	n.s.	n.s.	- 0,06***	- 0,04*
n.s.	0,07***	0,05*	0,13***	0,1***
n.s.	n.s.	0,03*	n.s.	0,04*
n.s.	n.s.	0,05*	- 0,09***	- 0,16***
n.s.	n.s.	n.s.	- 0,09*	- 0,13***
n.s.	n.s.	0,06*	- 0,08***	- 0,11***
n.s.	n.s.	n.s.	- 0,04*	- 0,05*
n.s.	n.s.	0,08*	n.s.	- 0,06*
n.s.	n.s.	n.s.	n.s.	n.s.
n.s.	n.s.	n.s.	- 0,1***	n.s.
n.s.	n.s.	0,05*	- 0,05*	n.s.
- 0,09*	n.s.	- 0,18*	- 0,19*	- 0,16*
n.s.	n.s.	n.s.	n.s.	- 0,07*
n.s.	n.s.	n.s.	n.s.	- 0,04*
n.s.	n.s.	n.s.	n.s.	- 0,09*
n.s.	n.s.	n.s.	n.s.	- 0,06*
n.s.	n.s.	n.s.	n.s.	n.s.
- 0,04*	- 0,04*	n.s.	n.s.	n.s.
n.s.	n.s.	n.s.	n.s.	n.s.
n.s.	n.s.	n.s.	n.s.	n.s.
n.s.	n.s.	n.s.	n.s.	n.s.
n.s.	n.s.	n.s.	n.s.	- 0,04*
n.s.	n.s.	n.s.	n.s.	n.s.
- 0,04*	n.s.	0,05*	n.s.	n.s.
n.s.	n.s.	n.s.	n.s.	n.s.
- 0,06*	n.s.	- 0,05*	- 0,04*	- 0,05*
n.s.	n.s.	n.s.	0,04*	n.s.
n.s.	n.s.	n.s.	n.s	n.s.
0,09*	n.s.	0,1*	n.s.	- 0,31***
n.s.	n.s.	0,05*	- 0,08***	- 0,2***
n.s.	n.s.	0,04*	- 0,04*	- 0,11***
n.s.	n.s.	n.s.	n.s.	- 0,05*
n.s.	n.s.	n.s.	n.s.	n.s.

Lecture : le tableau présente les résultats de régressions linéaires de différentes caractéristiques sur les progressions de scores observés entre la sixième et la troisième. À caractéristiques comparables, les garçons obtiennent de meilleurs résultats que les filles au test de mathématiques de fin de troisième puisque le coefficient estimé est positif (0,13) et significatif.
Source : MENESR-DEPP.

Des progressions liées à l'environnement social en mathématiques et en mémoire encyclopédique

Concernant les mathématiques et la mémoire encyclopédique, le modèle confirme l'influence des variables retenues sur la progression des collégiens (tableau 6). Néanmoins, ces modèles sont, là encore, moins significatifs que ceux représentant les scores en troisième. Pour ces deux compétences, moins de 10 % de la variance des évolutions de scores peut être expliquée par les variables comprises dans la modélisation.

À milieu social comparable, le contexte scolaire est un facteur de différenciation de la performance des collégiens. Les écarts d'évolutions de scores entre les enfants scolarisés dans le secteur privé et les autres sont significatifs. Les coefficients associés au secteur sont élevés et ont un poids similaire à ceux relatifs au retard scolaire. En effet, contrairement à ce que nous avons pu observer pour les trois compétences pour lesquelles les progressions de scores sont comparables selon l'environnement social, le retard scolaire reste ici un élément explicatif prépondérant des progressions entre la sixième et la troisième : les coefficients associés sont significatifs et négatifs. L'effet d'apprentissage identifié précédemment ne permettrait pas de réduire, voire d'inverser, les différences de progression pour ces compétences où les acquis cognitifs mesurés sont davantage liés à des connaissances strictement scolaires.

La modélisation met aussi en évidence une relation positive entre le niveau de diplôme du responsable de l'élève et la progression des scores : plus le niveau de diplôme est élevé, plus les chances de réussite de l'élève sont fortes. L'évolution des scores des collégiens s'explique dans une moindre mesure par l'indice de position sociale du responsable de l'élève. Celui-ci joue positivement uniquement pour l'épreuve d'évaluation de la mémoire encyclopédique. Quant aux variables décrivant la composition de la famille, la langue parlée à la maison et l'origine des parents, elles n'ont que peu d'influence sur les scores concernés.

Parmi les variables illustrant le contexte culturel dans lequel vit le collégien, c'est celle qui caractérise le rapport au livre qui se distingue. Ceci est particulièrement vrai pour l'épreuve de mémoire encyclopédique où les progressions des scores au collège sont plus marquées à mesure que le nombre de livres au domicile augmente. Le fait de ne pas disposer d'ordinateur ou d'accès à Internet semble avoir un effet négatif relativement faible sur la progression des performances dans ces deux compétences. Enfin, la fréquence d'écoute de la télévision n'a pas d'effet significatif sur la progression des scores.

Les influences déjà identifiées lors de notre analyse sur les variations des scores en troisième sont donc moins marquées lorsque l'on observe la progression au collège. Cependant, on peut constater que la disparité des performances scolaires reste conditionnée par certains marqueurs sociaux, tels que le diplôme du responsable de l'élève et l'environnement culturel direct de l'élève illustré par le nombre de livres au domicile.

CONCLUSION

Pour la première fois, le suivi d'un panel d'élèves du second degré a été enrichi par des données issues d'évaluations standardisées recueillies lors de deux temps de mesure, permettant ainsi d'analyser l'évolution des performances dans différentes compétences et de mettre en relation les résultats obtenus avec le contexte scolaire, social et culturel.

Les premières analyses descriptives portant sur la mesure des performances des collégiens ont ainsi conduit à identifier certains facteurs relatifs à l'environnement social et culturel déterminants dans l'explication des inégalités d'acquisition des compétences en fin de troisième. Néanmoins, le niveau de l'élève à l'entrée au collège reste un élément décisif, ce qui signifie que l'avenir scolaire de l'enfant dans le second degré serait en partie déterminé dès la sixième.

Ce bilan des compétences acquises en fin de troisième a pu être complété par une approche longitudinale. Ainsi, la modélisation des progressions de scores entre la sixième et la troisième en fonction de caractéristiques sociales, culturelles et scolaires, révèle des résultats différents selon les compétences. En effet, l'évolution des scores au test de compréhension (traitement de phrases lacunaires et lecture silencieuse) est peu sensible à l'environnement familial et scolaire de l'élève. Il en va de même pour l'épreuve de raisonnement logique. Pour ces compétences, les différenciations sociales apparaissent donc comme figées dès l'entrée au collège.

En revanche, pour les épreuves de mémoire encyclopédique et de mathématiques, le collège aurait tendance à accroître les inégalités sociales et les progressions observées seraient plus sensibles au parcours scolaire (retard, secteur de scolarisation). Pour la mémoire encyclopédique, les résultats suggèrent que les écarts sociaux auraient tendance à se creuser davantage pour des épreuves construites à partir d'un contenu strictement scolaire. Rappelons en effet que les items visant à mesurer la mémoire encyclopédique sont construits au regard du contenu des manuels scolaires de sixième et de troisième. Pour les mathématiques, la progression des écarts sociaux dans les performances fait écho aux résultats observés lors de la dernière enquête Pisa en culture mathématique selon lesquels la France apparaît comme l'un des pays de l'OCDE où la relation entre le niveau socio-économique des élèves et leurs performances est la plus grande.

BIBLIOGRAPHIE

Bourdieu P., Passeron J.-C., 1964, *Les Héritiers – Les étudiants et la culture*, Paris, éditions de Minuit, 183 p.

Bressoux P., Pansu P., 2003, *Quand les enseignants jugent leurs élèves*, Paris, Presses Universitaires de France, 190 p.

Caille J.-P., 1997, « Niveau d'acquisition à l'entrée en sixième et réussite au collège », *Note d'information*, n° 97.01, MENESR-DEPP.

Caille J.-P., Lemaire S., 2002, « Que sont devenus les élèves entrés en sixième en 1989 ? » *Données sociales*, Insee, p. 81-92.

Caille J.-P., Rosenwald F., 2006, « Les inégalités de la réussite à l'école élémentaire : construction et évolution », *France Portrait social,* Insee, p. 115-137.

Cayouette-Remblière J., 2013, *Le marquage scolaire – Une analyse « statistique ethnographique » des trajectoires des enfants de classes populaires à l'école,* Thèse de doctorat en sociologie de l'EHESS, 585 p.

Chartier P., 2012, *Évaluer les capacités de raisonnement avec les tests RCC,* Paris, Eurotests éditions, 112 p.

Duru-Bellat M., Jarousse J.-P., Mingat A., 1993, « Les scolarités de la maternelle au lycée – Étapes et processus dans la production des inégalités sociales », *Revue Française de Sociologie*, n° 34, vol. 1, p. 43-60.

Grisay A., 1997, « Évolution des acquis cognitifs et socio-affectifs des élèves au cours des années de collège », *Note d'information,* n° 97.26, MENESR-DEPP.

Le Donné N., Rocher T., 2010, « Une meilleure mesure du contexte socio-éducatif des élèves et des écoles – Construction d'un indice de position sociale à partir des professions des parents », *Éducation & formations,* n° 79, MENJVA-DEPP, p. 103-115.

Lemaire S., 2006, « Le devenir des bacheliers : parcours après le baccalauréat des élèves entrés en sixième en 1989 », *Note d'information,* n° 06.01, MENESR-DEPP.

Lemaire S., Guyon V., Murat F., 2007, « Un élève sur deux entrés en sixième en 1995 fait des études 10 ans plus tard », *Insee Première*, n° 1158, Insee.

Lieury A., 2012, *Mémoire et réussite scolaire*, Paris, Dunod (4e édition), 193 p.

MENESR-DEPP, *Repères et références statistiques*, 2014, Paris, 431 p.

Octobre S., Détrez C., Berthomier N. et Merklé P., 2010, *L'enfance des loisirs. Trajectoires communes et parcours individuels de la fin de l'enfance à la grande adolescence,* Paris, ministère de la Culture et de la Communication, La Documentation française, coll. « Questions de culture », 432 p.

O'Prey S., 2004, « Les activités extrascolaires des écoliers : usages et effets sur la réussite », *Éducation & formations,* n° 69, MENESR-DEPP, p. 89-105.

Vanhoffelen A., 2010, « Les bacheliers du panel 1995 : évolution et analyse des parcours », *Note d'information*, n° 10.13, MEN-DEPP.

NOUVELLES ANALYSES DE L'ENQUÊTE PISA 2012 EN MATHÉMATIQUES

Un autre regard sur les résultats

Éric Roditi
Sorbonne Paris Cité, université Paris Descartes,
laboratoire EDA (Éducation Discours Apprentissages)

Franck Salles
MENESR-DEPP, bureau de l'évaluation des élèves

Les enquêtes PISA visent un suivi des acquis scolaires des élèves de 15 ans. En ce qui concerne ceux de la culture mathématique, le choix de l'OCDE est d'évaluer des compétences, c'est-à-dire des capacités à mobiliser ses connaissances pour résoudre un problème en lien avec une situation de la vie réelle. Un regard didactique porté sur l'évaluation de 2012 montre que les classifications utilisées par l'OCDE ne permettent ni de recenser précisément les connaissances acquises des élèves ni d'estimer le niveau d'acquisition de ces connaissances.
Les auteurs proposent ici une nouvelle classification des items permettant de distinguer différents niveaux d'utilisation des connaissances mathématiques pour résoudre les problèmes proposés. Ils cherchent ainsi à mieux connaître les acquis des élèves. La présentation de cette classification et de son intérêt s'appuie sur l'analyse de quelques exemples extraits de PISA 2012.
Une étude complète de l'ensemble des items PISA 2012 à l'aune de cette nouvelle classification est ensuite proposée. Elle confirme la pertinence de la classification, notamment par une mise en lien du niveau d'exigence des items et de la réussite des élèves à ces items.
Puis les auteurs procèdent à un examen particulier du cas de la France. En s'appuyant sur cette même classification, ils enrichissent et nuancent les résultats de l'OCDE concernant les inégalités de performances des élèves selon le sexe, l'origine sociale ou le retard scolaire. Leurs analyses montrent notamment que les filles sont d'autant plus pénalisées que les tâches leur demandent de l'initiative, et que les difficultés des élèves en retard scolaire ou de milieu populaire ne sont pas accrues lorsque les activités attendues d'eux sont plus exigeantes.

Les enquêtes PISA (programme international pour le suivi des acquis des élèves) visent à mettre en lumière ce que les élèves de 15 ans ont appris à l'école, avec quelles différences suivant les pays participants comme au sein de chacun d'entre eux. Ces enquêtes conduisent à un classement international, mais pas seulement. Elles visent aussi à rendre compte du niveau atteint par les élèves les plus performants comme par les plus faibles, à pointer les inégalités d'acquis entre les filles et les garçons ou selon les catégories sociales. Les connaissances issues de PISA dépendent de ce qui est précisément mesuré. En mathématiques, par exemple, les élèves ne sont pas testés sur leur capacité à restituer leur connaissance des définitions, des notions ou des énoncés des règles. Ce sont leurs compétences qui sont évaluées, c'est-à-dire ce qu'ils mobilisent pour comprendre et résoudre un problème. Les organisateurs de PISA cherchent également à évaluer des compétences variées : les questions n'ont pas le même niveau de difficulté ; elles ne font pas appel aux mêmes processus psycho-cognitifs ; elles correspondent à des domaines mathématiques différents ; et elles sont posées dans des contextes diversifiés de la vie réelle.

Néanmoins, les analyses ne différencient pas, par exemple, les questions qui nécessitent l'application directe d'une règle mathématique de celles qui exigent une prise d'initiative. Une nouvelle catégorisation des items est proposée dans cet article ; elle distingue les compétences en fonction de différents niveaux concernant les activités mathématiques requises. Sur quelques items, des informations présentées dans le rapport de l'OCDE et des analyses auxquelles conduit cette nouvelle catégorisation sont mises en regard. Puis une étude des données concernant la France conduit à une réinterprétation des connaissances apportées par PISA quant aux difficultés en mathématiques, aux inégalités filles-garçons, à l'effet des différences sociales ou du retard scolaire sur les acquis en mathématiques.

COMMENT DÉCRIRE LES COMPÉTENCES MATHÉMATIQUES DES ÉLÈVES DE 15 ANS ?

Piloté par l'OCDE (Organisation pour la coopération et le développement et économique), PISA fait référence, quant à l'évaluation des acquis des élèves, à la fin de la scolarité obligatoire. À travers l'opinion publique ou les décideurs, il influence les politiques éducatives de nombreux pays et notamment celle de la France depuis le début des années 2000. Ainsi peut en effet s'interpréter la conception et la mise en œuvre du socle commun des connaissances et des compétences de 2005. En 2012, comme pour la dernière fois en 2003, le domaine majeur de l'évaluation PISA fut la culture mathématique. Mais qu'évalue vraiment PISA en mathématiques ? Qu'est-ce que la culture mathématique et comment est-elle représentée dans les items du test ?

Dans cette première partie, nous nous attachons à répondre à ces questions. Nous décrivons les objectifs de l'OCDE et des concepteurs internationaux du test, puis nous proposons une étude nouvelle des items et des activités qu'ils

conduisent à réaliser. Cette étude a été effectuée au sein de la DEPP (direction de l'évaluation, de la prospective et de la performance) par un groupe d'experts de l'enseignement des mathématiques.

Cadre défini par l'OCDE pour l'évaluation de la culture mathématique

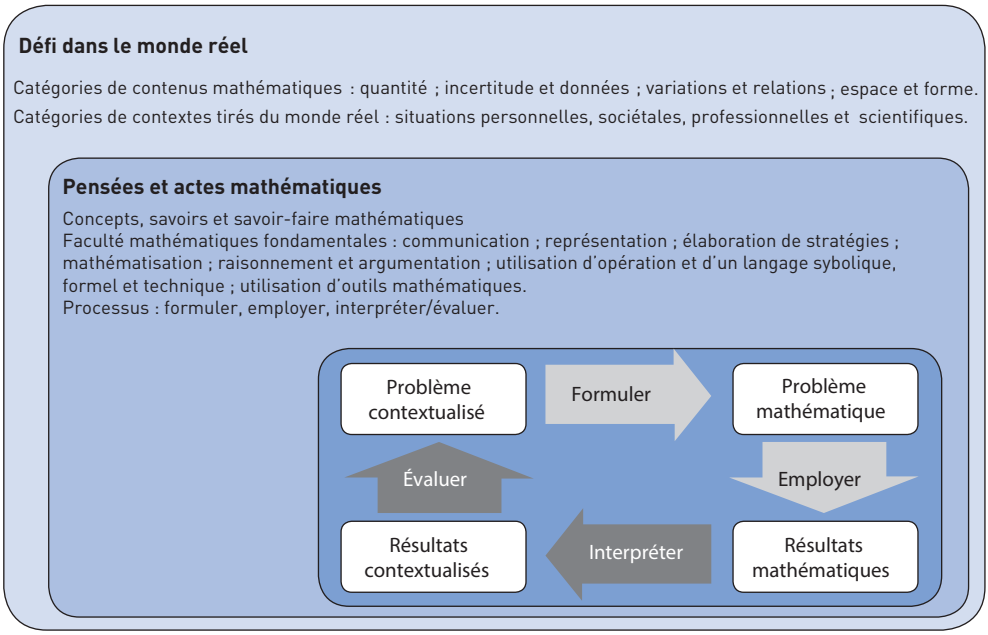
L'OCDE, à travers la documentation qu'elle publie sur PISA, énonce les principes directeurs du programme d'évaluation. En particulier, elle définit les connaissances que les élèves doivent mobiliser, les processus qu'ils doivent mettre en œuvre pour répondre aux questionnaires ainsi que les contextes dans lesquels leurs savoirs et savoir-faire sont évalués. Le cadre d'évaluation de la culture mathématique est fondé sur une approche psychologique de l'activité mathématique des élèves répondant aux items du test. Il a été élaboré pour l'OCDE conjointement par l'*Australian Council for Educational Research* (ACER) et par une organisation de recherche pédagogique basée aux États-Unis, *Achieve Inc*. [OCDE, 2013]. Indiquons-en les éléments essentiels à partir de quelques extraits de cette documentation.

> « L'enquête PISA se fonde sur une conception de l'évaluation des connaissances, des compétences et des attitudes qui reflète l'évolution des programmes d'enseignement : elle va au-delà des acquis purement scolaires et se concentre sur la mise en œuvre des savoirs et savoir-faire dans des tâches et des défis quotidiens, que ce soit en famille ou dans le monde du travail. [...] L'enquête PISA cible des activités que les élèves âgés de 15 ans auront à accomplir dans l'avenir et cherche à identifier ce qu'ils sont capables de faire avec ce qu'ils ont appris [...]. Les épreuves sont conçues à la lumière du dénominateur commun des programmes scolaires des pays participants, sans toutefois s'y cantonner. Elles servent à évaluer les connaissances des élèves, certes, mais aussi leur faculté de réflexion et leur capacité à appliquer leurs connaissances et leurs expériences dans des situations qui s'inspirent du monde réel. » [OCDE, 2013, p. 13]
>
> « La culture mathématique est l'aptitude d'un individu à formuler, employer et interpréter des mathématiques dans un éventail de contextes, c'est-à-dire à raisonner en termes mathématiques et à utiliser des concepts, procédures, faits et outils mathématiques pour décrire, expliquer et prévoir des phénomènes. Elle aide les individus à comprendre le rôle que les mathématiques jouent dans le monde et à se comporter en citoyens constructifs, engagés et réfléchis, c'est-à-dire à poser des jugements et à prendre des décisions en toute connaissance de cause. » [OCDE, 2013, p. 27]

Afin de mesurer les acquis des élèves en mathématiques, retenons que l'OCDE catégorise les items suivant quatre grands domaines mathématiques (*quantité*, *incertitude et données*, *variations et relations*, *espace et forme*), suivant quatre types de contextes (*personnel*, *sociétal*, *professionnel, scientifique*) et suivant trois processus psycho-cognitifs en jeu dans la résolution d'un problème. Ces processus peuvent se décrire comme suit : *formuler* (mathématiser les situations de vie réelle), *employer* (travailler au sein du modèle mathématique) et *interpréter/évaluer* (mettre un résultat mathématique à l'épreuve d'une situation réelle). Ils sont considérés comme constitutifs d'un cycle [OCDE, 2013, p. 29] ▸ **Figure 1 p. 238**.

Les experts s'accordent pour penser qu'il ne serait toutefois pas pertinent de construire un dispositif d'évaluation portant, pour toutes les questions de ce dispositif, sur chacun des processus du cycle. Il est ainsi fréquent, dans les items de l'enquête PISA, que plusieurs d'entre eux soient déjà pris en charge dans l'énoncé

Figure 1 **Les catégories retenues par PISA pour classer des items**



Source : OCDE, 2013.

et que l'élève qui cherche à résoudre le problème n'en ait qu'un ou deux à mettre en œuvre [OCDE, 2013, p. 28].
Les références psychologiques qui ont conduit à définir les facultés mathématiques fondamentales comme les processus cognitifs confèrent à ces catégories une relative indépendance de la discipline mathématique. Cela correspond sans doute à une volonté de l'OCDE puisque PISA n'évalue jamais la capacité des élèves à appliquer une connaissance ou une technique mathématique isolément. Il convient à ce propos de remarquer une certaine évolution par rapport aux précédents cycles de l'évaluation. Le processus *« employer »* qui a été introduit dans PISA 2012 renouvelle le regard porté par l'OCDE sur les acquis des élèves dans leur utilisation des outils mathématiques : application de théorèmes, techniques de calcul arithmétique, algébrique, etc.

Ainsi, bien que tous les problèmes soient posés dans un contexte de vie réelle, ce dernier n'a pas toujours d'influence effective sur l'activité de l'élève. En outre, même lorsque l'élève doit *« formuler »* en langage mathématique la situation issue de ce contexte de la vie réelle et/ou *« interpréter/évaluer »* les résultats obtenus par rapport à ce contexte, il reste toujours une partie de l'activité de résolution qui consiste à *« employer »* des connaissances mathématiques. C'est donc la subdivision des étapes de la résolution du problème en différents items qui permet aux experts de PISA de classer chaque item dans une et une seule de ces trois catégories ; et ce classement témoigne, non pas d'une seule activité, mais plutôt de l'activité dominante. Il reste que tous les items ne sont pas équivalents quant

au niveau de mise en fonctionnement des connaissances mathématiques, et qu'ils ne reflètent donc pas le même niveau d'acquisition. C'est justement pour se donner les moyens de distinguer ces niveaux qu'une étude a été menée en 2013. D'autres études complémentaires avaient d'ailleurs déjà été menées, en France, après PISA 2003, pour étudier la correspondance entre les items du questionnaire et les pratiques usuelles d'enseignement en fonction des programmes scolaires en vigueur en France [Bodin, 2009].

Différents niveaux de mise en fonctionnement des connaissances mathématiques dans les items de mathématiques de PISA 2012

L'étude complémentaire dont il est question a été réalisée par un groupe d'experts de la DEPP ; c'est elle en effet qui, en France, administre le test PISA. Le groupe était composé d'enseignants, de formateurs, d'inspecteurs académiques et généraux de l'Éducation nationale et d'un professeur des universités, didacticien des mathématiques[1]. Ce dernier et l'enseignant responsable du groupe sont les deux auteurs de cet article.

Les analyses produites se fondent sur des apports de la recherche en didactique des mathématiques ; elles s'appuient sur une classification différente des items du test et conduisent à des interprétations nouvelles des résultats de l'enquête PISA. Cette classification s'applique à tous les items du questionnaire. Elle repose sur une analyse de l'énoncé visant à déterminer la nature de la mise en fonctionnement des connaissances mathématiques nécessaires pour répondre à la question posée dans l'item. Elle est par conséquent indépendante des passations préalables qui permettent de déterminer la difficulté relative des items et leur pouvoir discriminant, en référence à la théorie de la réponse à l'item utilisée par les experts de PISA. Précisons enfin que depuis les années 1970, la recherche en didactique des mathématiques s'est développée en se dotant d'un corpus théorique et méthodologique spécifique afin d'étudier les phénomènes d'enseignement et d'apprentissage de contenus mathématiques précis dans des institutions données [Brousseau, 1998 ; Chevallard, 1992 ; Vergnaud, 1990]. Les didacticiens se sont encore peu consacrés aux questions d'évaluation hormis pour l'analyse des productions d'élèves en situations scolaires afin de comprendre les conceptions que produit l'enseignement quant aux notions dont les recherches sont l'objet [Roditi, 2012]. De tels travaux ont par exemple mis en évidence différentes conceptions d'élèves, difficiles à faire évoluer, à propos des nombres décimaux, d'éléments d'algèbre élémentaire (la lettre, le signe d'égalité, etc.), de la symétrie orthogonale, etc.

Il est intéressant toutefois de tirer profit des travaux produits en didactique pour l'analyse de situations d'enseignement afin d'étudier des questions d'évaluation. Les items peuvent ainsi être différenciés selon deux premières catégories de mise en fonctionnement des connaissances mathématiques : ceux pour lesquels la réponse repose uniquement sur la compréhension qualitative de contenus – concepts, théorèmes, etc. – sans réalisation de la part de l'élève, et ceux qui

1. En plus des auteurs, ont participé à ce groupe : A.-M. Camper (enseignante), A. Diger (IA-IPR), N. Grapin (formatrice), S. Herrero (enseignant), M.-C. Obert (IA-IPR), J. Yebbou (IGEN).

nécessitent effectivement la mise en œuvre d'une procédure reposant sur des contenus mathématiques. Les questions posées dans PISA émergent toujours de situations liées à la vie réelle. Les items de la première catégorie évaluent ainsi la compréhension d'un savoir mathématique en contexte, mais seulement en tant qu'objet, les élèves n'ayant pas à l'utiliser. Les items de la seconde catégorie évaluent, en revanche, l'acquisition de ces savoirs en tant qu'outils, c'est-à-dire la capacité à les mettre en œuvre pour résoudre un problème. Cette distinction entre les caractères *objet* et *outil* des savoirs mathématiques avait été effectuée par Douady [1986], didacticienne, pour rendre compte de la dynamique à l'œuvre lors de la construction de nouvelles connaissances mathématiques, ces deux caractères entretenant une relation dialectique au cours de l'activité.

Ainsi, certaines questions d'évaluation portent sur des contenus mathématiques pour attester de leur compréhension pour eux-mêmes ; elles visent le caractère *objet* de ces contenus. Il en va ainsi de nombreux exercices classiques d'entraînement de calcul numérique ou algébrique où les élèves attestent de leur capacité à effectuer une opération sans même que soit interrogée l'opportunité de poser cette opération dans un problème. Comme cela a déjà été expliqué, il n'y a pas d'items de la sorte dans PISA. Il y a, en revanche, des items qui portent sur le caractère *objet* d'un concept, et où les élèves doivent témoigner d'une compréhension de ce concept sans avoir à le mettre en œuvre, ce que certains auteurs appellent une *compréhension conceptuelle* [Kilpatrick, Swafford, Findel, 2001]. Ce serait le cas, par exemple, d'un item demandant si un enfant qui jette un dé qui est tombé sur « 6 » la première fois possède plus ou moins de chance d'obtenir « 6 » la deuxième fois. Sans demande de justification, aucune technique ou méthode n'est requise, il s'agit seulement d'exprimer par une réponse sa compréhension de l'indépendance des événements aléatoires. Les savoirs ainsi évalués dans PISA concernent souvent la probabilité, la notion de moyenne, les fonctions et les grandeurs. Nous avons réuni ces items dans une même catégorie appelée « compréhension qualitative de concepts » ou plus simplement « concept ».

D'autres questions évaluent le caractère *outil* des savoirs, l'élève doit alors mettre une connaissance mathématique en fonctionnement après s'être assuré de la pertinence de cette connaissance pour traiter la question posée dans le contexte indiqué. Nous distinguons ces mises en fonctionnement suivant qu'elles sont plus ou moins suggérées par l'énoncé, suivant aussi le degré d'initiative demandé à l'élève. Cela correspond en effet, selon nous, à différents niveaux d'acquisition des connaissances. En nous inspirant de travaux déjà effectués sur ce sujet en didactique [Robert, 1998], nous considérons trois niveaux de mise en fonctionnement des contenus mathématiques.

Le premier niveau est celui où l'élève effectue une tâche courante et obtient directement le résultat attendu par la mise en œuvre d'une procédure, souvent unique, qui est indiquée ou suggérée par l'énoncé, et dont les programmes scolaires permettent de penser qu'elle est automatisée pour les élèves. Dans les items PISA, de tels items conduisent généralement à l'application d'une propriété géométrique, d'une règle de calcul, d'une lecture graphique directe, etc. Il peut s'agir également d'une simple mise en lien de connaissances mathématiques avec le contexte de la situation. Les items correspondant à ce premier niveau de mise en fonctionnement sont regroupés dans une catégorie appelée « mise en fonctionnement directe d'une procédure » ou plus simplement « directe ».

Les items qui relèvent du second niveau nécessitent que l'élève adapte ou transforme l'énoncé – les données ou la question posée – avant d'appliquer ses connaissances. La transformation peut prendre la forme d'une transformation d'information : convertir, par exemple, une donnée dans une autre unité de mesure. Il peut s'agir d'un changement de point de vue sur des objets mathématiques ou sur une relation entre des objets : isoler, par exemple, une figure plane d'une figure de l'espace ; ou, ayant à établir que trois points sont alignés, considérer la droite qui passe par les deux premiers et montrer que le troisième appartient à cette droite. L'élève peut aussi avoir à changer de cadre [Douady, 1986] ou de registre de représentation [Duval, 1995] : passer, par exemple, dans le cadre graphique pour résoudre un problème numérique ; convertir une procédure indiquée dans le registre langagier en un calcul appartenant au registre numérique ou algébrique. Tous ces items ont été regroupés dans une catégorie appelée « mise en fonctionnement d'une procédure avec adaptation de l'énoncé » ou plus simplement « adaptation ».

Dans les items du troisième niveau, la mise en fonctionnement des contenus nécessite que l'élève, de manière autonome, introduise un ou plusieurs intermédiaires. Ils peuvent concerner le processus de résolution lui-même : décomposer une question en plusieurs étapes ; introduire une notation pour traiter le problème (par exemple en attribuant une lettre à différentes variables), etc. Il peut s'agir également d'intermédiaires ajoutés aux données : considérer une nouvelle variable combinant, par exemple, deux variables déjà explicitées dans un problème numérique ; utiliser un nouvel objet géométrique pour résoudre un problème, par exemple en considérant une droite ou un cercle qui n'apparaît pas dans l'énoncé ; introduire une fonction là où deux variables étaient indiquées avec une relation numérique les reliant, etc. Parce qu'il s'agit d'un intermédiaire qui n'est pas suggéré par l'énoncé, l'introduction correspond à une initiative de la part de l'élève ; elle est totalement à sa charge. Les items de ce type sont regroupés dans une catégorie appelée « mise en fonctionnement d'une procédure avec introduction d'intermédiaires » ou plus simplement « intermédiaires ».

La distinction de ces quatre catégories de mise en fonctionnement des connaissances mathématiques, qui portent sur leur dimension *objet* comme sur leur dimension *outil*, conduit à poser un nouveau regard sur les items de PISA, ainsi que sur les résultats produits par ce programme.

APPORTS DE LA NOUVELLE CLASSIFICATION À L'ANALYSE D'UN ITEM DE MATHÉMATIQUES DE PISA 2012

Illustrons l'intérêt de cette nouvelle classification pour l'étude des items du questionnaire de culture mathématique de PISA 2012. Dans cette partie, nous proposons une analyse détaillée de quelques items rendus publics (les experts parlent d'items « libérés »). Dans la partie suivante, nous présentons une étude synthétique de l'ensemble des items du questionnaire. Précisons que l'analyse complète a été possible parce que le groupe d'experts de la DEPP a accès à l'ensemble des items du test, qu'ils soient libérés ou non.

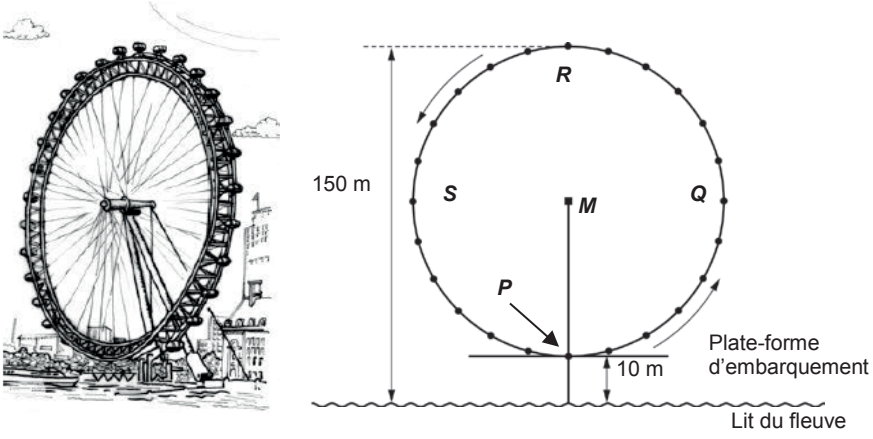
Intérêt des niveaux de mise en fonctionnement des connaissances pour l'analyse des items PISA 2012

Les items de PISA sont tous liés à des contextes de la vie réelle. Néanmoins, comme nous l'avons déjà signalé, ce contexte n'a pas toujours d'influence sur l'activité de l'élève. Ces items sont regroupés dans la catégorie *« employer »* selon le cadre de PISA. C'est le cas dans l'exemple de la **figure 2**, où l'étude d'une roue de manège est proposée, mais où l'activité de l'élève porte essentiellement sur la figure géométrique représentée dans l'illustration figurant dans l'énoncé.

Dans cet item, l'élève peut appliquer les propriétés du diamètre et du rayon d'un cercle à une figure où les mesures sont de simples nombres entiers puis tenir compte de la longueur séparant le bas de la roue avec le lit du fleuve. La figure proposée rend possible une utilisation implicite de ces connaissances puisque le point *M*, défini comme le centre du cercle, est placé au milieu du segment [*PR*] qui en est un diamètre. Différentes procédures directes et équivalentes sont envisageables : calculer la moitié de 140 et ajouter 10, enlever la moitié de 140 à 150, etc. Ce qui est important ici, c'est de

Figure 2 **Item libéré, PISA 2012**

Une grande roue est installée sur les rives d'un fleuve.
En voici une image et un schéma :



Le diamètre externe de la grande roue est de 140 mètres et son point le plus élevé se situe à 150 mètres au-dessus du lit de la Tamise, sur la rive du fleuve. Elle tourne dans le sens indiqué par les flèches.

Question : LA GRANDE ROUE

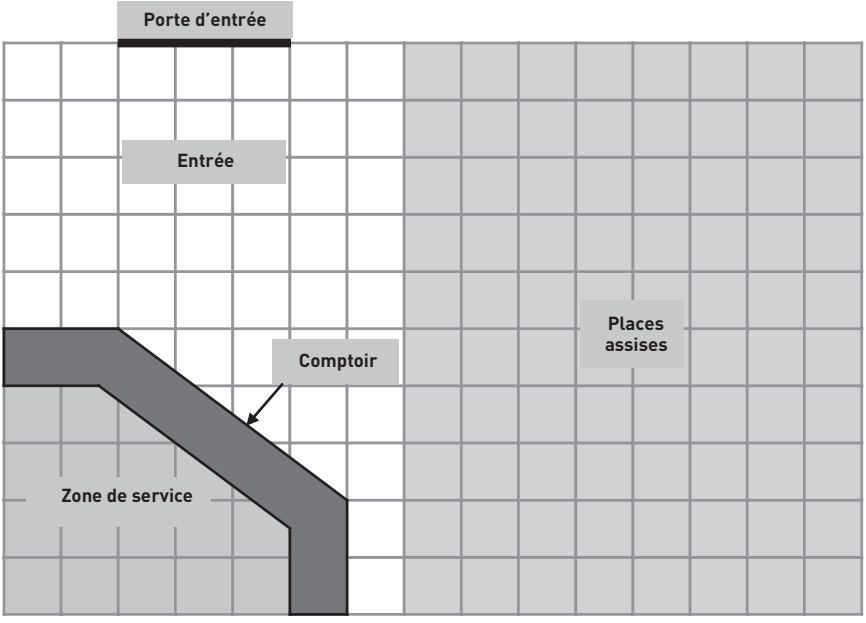
La lettre *M* dans le diagramme indique le centre de la roue.
À combien de mètres (m) au-dessus du lit du fleuve se trouve le point *M* ?
Réponse : ..... m

noter que l'activité de l'élève n'est pas influencée par le contexte de la situation (fleuve, plate-forme d'embarquement, sens de rotation de la roue, etc.), elle porte seulement sur la figure géométrique donnée dans l'énoncé. Ces faits expliquent pourquoi les concepteurs de PISA ont associé à cet item le processus psycho-cognitif *« employer »* et le domaine mathématique *« espace et formes »*.
Examinons maintenant un autre item portant lui aussi sur une situation du champ géométrique et ne demandant pas non plus à l'élève d'utiliser le contexte dans son activité de résolution du problème posé ▸ **Figure 3**.

Détaillons l'analyse *a priori* de cette activité. Après lecture de l'énoncé et identification de la paroi extérieure du comptoir sur le plan, plusieurs méthodes de résolution sont possibles pour un élève en fin de scolarité obligatoire en France, qui reposent sur des connaissances mathématiques différentes. Par exemple une méthode par mesure et application d'échelle est possible : déterminer l'échelle en mesurant à la règle graduée la longueur de deux carreaux sur le dessin qui représentent un mètre dans la réalité ; mesurer la longueur sur le plan de la paroi extérieure ; appliquer

▸ **Figure 3** **Item libéré, PISA 2012**

Voici le plan du magasin de glaces de Marie, qu'elle est en train de rénover.
La zone de service est entourée d'un comptoir.



Remarque : chaque carré de la grille représente 0,5 mètre sur 0,5 mètre.

Question 1 : CHEZ LE GLACIER

Marie veut installer une nouvelle bordure le long de la paroi extérieure du comptoir.
Quelle est la longueur totale de bordure dont elle a besoin ? Montrez votre travail.

enfin l'échelle précédemment déterminée pour calculer la longueur réelle recherchée. Il est à noter qu'une valeur approchée de la réponse sera acceptée par le correcteur, les consignes de correction internationales étant appliquées. L'élève peut également travailler tout autrement et mener un raisonnement géométrique fondé sur le théorème de Pythagore après avoir introduit sur le plan un triangle rectangle dont l'hypoténuse est la partie oblique du comptoir. Il devra tenir compte de l'échelle de 1 carreau pour 0,5 mètre, ce qui peut être fait avant ou après application du théorème de Pythagore. Ce qui doit être remarqué ici, encore une fois, c'est que le contexte du magasin de glace n'intervient en rien sur l'activité permettant de trouver la bonne réponse. C'est pourquoi les concepteurs de PISA ont associé à cet item le processus psycho-cognitif *« employer »* et le domaine de connaissances mathématiques *« espace et formes »*.

Poursuivons l'analyse. Dans les deux items que nous venons d'examiner, on ne peut être certain de la connaissance mathématique mobilisée pour répondre aux questions posées. L'enquête PISA ne cherche donc pas à connaître précisément les connaissances mathématiques acquises par les élèves, mais seulement leur domaine parmi les quatre qui sont distingués pour cette discipline. Elle ne vise pas non plus à rendre compte des différentes modalités d'expression des connaissances mathématiques acquises par les élèves. Les deux items, qui se retrouvent en effet classés dans les mêmes catégories *« employer »* et *« espace et formes »* par les experts PISA, requièrent pourtant des modalités d'expression des connaissances mathématiques très différentes. Dans le premier, l'élève effectue un calcul explicitement demandé dans la consigne, ce calcul portant sur deux longueurs clairement indiquées sur une figure elle-même fournie dans l'énoncé. Dans le second, l'emploi d'un calcul d'échelle ou d'un raisonnement basé sur le théorème de Pythagore nécessite des étapes qui, n'étant absolument pas induites par l'énoncé, sont entièrement à la charge de l'élève. Ces deux items évaluent donc bien la capacité à *« employer »* des connaissances dans des situations géométriques déjà mathématisées, mais ils ne sont absolument pas équivalents quant au niveau de mise en fonctionnement de ces connaissances. L'échec ou la réussite à ces deux items ne témoigne donc pas du même niveau d'acquisition. C'est ce dont la classification que nous proposons permet de justement rendre compte.

Analyse qualitative de quelques items à l'aide de la nouvelle classification

L'analyse de quelques items libérés permet de rendre compte des informations que la classification élaborée par le groupe d'experts de la DEPP peut apporter en complément de celles déjà produites par PISA. Une analyse systématique de tous les items a également été effectuée. Elle conduit à un nouveau regard sur l'évaluation de la culture mathématique en 2012.

Commençons par un item posé pour, selon notre classification, évaluer la compréhension qualitative d'un concept mathématique. Cet item porte sur le caractère *objet* du concept, il ne conduit à la mise en œuvre d'aucune procédure ▸ **Figure 4**. Pour répondre à cette question, l'élève doit manifester sa compréhension de la relation entre le temps et le débit lors d'une perfusion, en utilisant éventuellement la formule donnée. Si une justification précise devait être apportée à la réponse, l'élève pourrait convoquer la notion de proportionnalité inverse et en déduire, soit à partir du contexte

Figure 4 Item libéré, classé dans la catégorie « concept », PISA 2012

Les perfusions servent à administrer des liquides et des médicaments aux patients.
Les infirmières doivent calculer le débit D d'une perfusion en gouttes par minute.

Elles utilisent la formule $D = \frac{f \times V}{60 \times n}$ où

f est le facteur d'écoulement en gouttes par millilitre (mL)
V est le volume (en mL) de la perfusion
n est le nombre d'heures que doit durer la perfusion.

Question 1 : DÉBIT D'UNE PERFUSION

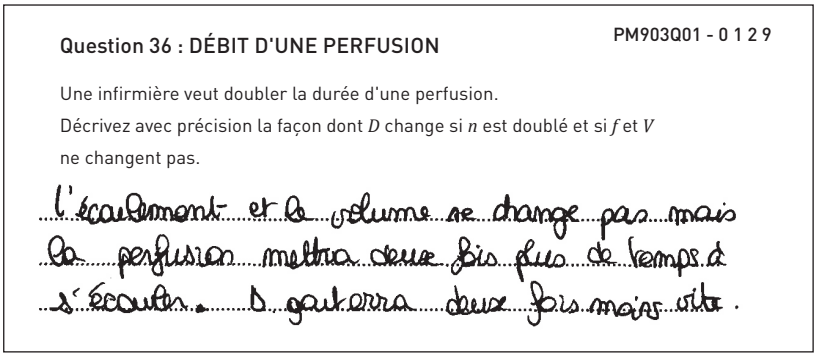
Une infirmière veut doubler la durée d'une perfusion.
Décrivez avec précision la façon dont D change si n est doublé et si f et V ne changent pas.

de vie réelle (toutes choses égales par ailleurs, si le temps d'écoulement est deux fois plus long, c'est que le débit est deux fois moins important) soit algébriquement (si le dénominateur est doublé, les autres variables restant constantes, le quotient est divisé par 2). Dans le cas de l'item étudié, l'élève de 15 ans n'a pas à justifier sa réponse, il n'applique donc vraisemblablement aucune procédure ou technique : la procédure qui conduit à penser D en fonction de n, à remplacer n par $2n$ et à établir par calcul littéral que $D(2n) = D(n)/2$ est très rare chez les élèves de 15 ans, comme le confirment leurs productions **Figures 5a et 5b**.

Selon le cadre de PISA, cet item correspond à des connaissances du domaine *« variations et relations »*, il évalue le processus *« employer »*. Les résultats obtenus après passation nous apprennent qu'il n'est réussi que par 22,2 % des élèves scolarisés dans les pays de l'OCDE (17,7 % en France), et que 27,3 % d'entre eux ne répondent pas à la question posée (30,8 % en France). Ni le domaine mathématique évalué, ni le processus en jeu ne suffisent à expliquer ces résultats qui témoignent de la difficulté de cet item.
La classification proposée par la DEPP et issue de la recherche en didactique des mathématiques complète l'analyse. Elle met en relief le fait que l'activité requise repose sur la compréhension d'un concept – le débit comme grandeur, quotient du volume et de la durée – sans procédure type associée pour le mettre en œuvre comme un outil. Cela n'est pas encore suffisant pour expliquer précisément les performances des élèves. Leur origine est multifactorielle : il faudrait également prendre en compte la familiarité avec la situation, la notion mathématique précisément en jeu, la difficulté du texte de l'énoncé, etc. Cela permet néanmoins de mieux comprendre à la fois le faible score de réussite et le pourcentage élevé de non-réponses.

Analysons maintenant des items qui visent l'évaluation du caractère *outil* des savoirs mathématiques, c'est-à-dire où l'élève doit mettre une connaissance en fonctionnement après s'être assuré de la pertinence de cette connaissance pour traiter la question posée dans le contexte indiqué. Nous avons distingué trois niveaux différents de mise en fonctionnement. Commençons par illustrer le premier où les élèves ont à mobiliser une procédure directe ou bien simplement, comme c'est le cas présenté **figure 6**, à mettre leurs connaissances en relation avec le contexte de la situation.

Figure 5a **Réponse d'un élève à l'item portant sur le débit d'une perfusion, PISA 2012**



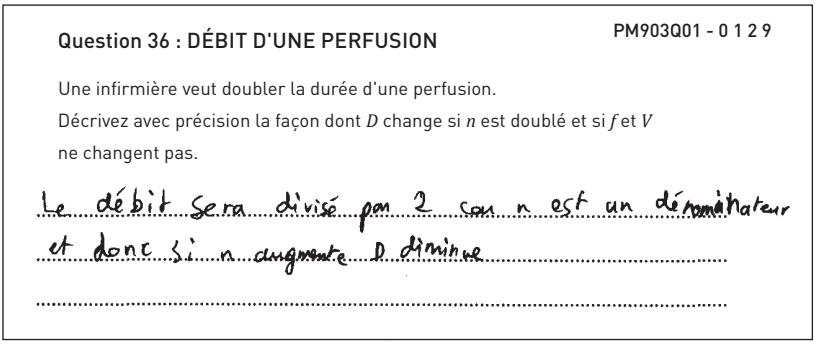
Question 36 : DÉBIT D'UNE PERFUSION

PM903Q01 - 0 1 2 9

Une infirmière veut doubler la durée d'une perfusion.
Décrivez avec précision la façon dont D change si n est doublé et si f et V ne changent pas.

L'écoulement et le volume ne change pas mais la perfusion mettra deux fois plus de temps à s'écouler. D gaterra deux fois moins vite.

Figure 5b **Réponse d'un autre élève à l'item portant sur le débit d'une perfusion, PISA 2012**



Question 36 : DÉBIT D'UNE PERFUSION

PM903Q01 - 0 1 2 9

Une infirmière veut doubler la durée d'une perfusion.
Décrivez avec précision la façon dont D change si n est doublé et si f et V ne changent pas.

Le débit sera divisé par 2 car n est un dénominateur et donc si n augmente D diminue

Après avoir lu l'énoncé et reconnu une situation de proportionnalité, l'élève doit appliquer ses connaissances sur cette notion dans un cas numériquement simple. La reconnaissance du savoir en jeu ici est fortement suggérée par la situation de la recette : elle est très familière pour les élèves et toujours reliée – souvent implicitement – à la notion de proportionnalité. Plusieurs méthodes sont possibles, mais toutes relèvent de la même procédure et du même registre numérique : passage à l'unité, coefficient de proportionnalité, produit en croix, etc. Il n'y a pas de conversion à effectuer. Cet item relève donc de la catégorie des questions nécessitant la mise en fonctionnement directe d'une procédure connue.
La documentation produite par PISA [OCDE, 2014] nous apprend que cet item du domaine « *quantité* » évalue le processus « *formuler* » et que les élèves de l'OCDE l'ont réussi à 63 % d'entre eux (56,2 % en France) avec 3 % de non-réponses (5,6 % en France). Bien que les problèmes de proportionnalité soient toujours difficiles pour beaucoup d'élèves, le fait que la réponse à la question ne nécessite que la mise en œuvre d'une procédure classique, de manière directe et sans doute déjà appliquée de nombreuses fois au cours de la scolarité, explique, par une analyse indépendante des passations du questionnaire, les résultats globaux que les élèves obtiennent sur cet item, qui sont à la fois bien meilleurs que ceux évoqués pour l'item précédent et avec une absence de réponse beaucoup plus faible.

L'exercice suivant demande davantage aux élèves quant à la mise en fonctionnement des connaissances : sa résolution repose sur une adaptation de l'énoncé ▸ Figure 7. Cet item est ancien, mais les items libérés de l'enquête de 2012 ne permettent pas de couvrir l'ensemble de la classification que nous proposons.

▸ Figure 6 Item libéré, classé dans la catégorie « directe », PISA 2012

Question 1 : SAUCE

Vous préparez votre propre vinaigrette pour une salade.
Voici une recette pour préparer 100 millilitres (mL) de vinaigrette:

Huile pour salade	60 mL
Vinaigre	30 mL
Sauce soja	10 mL

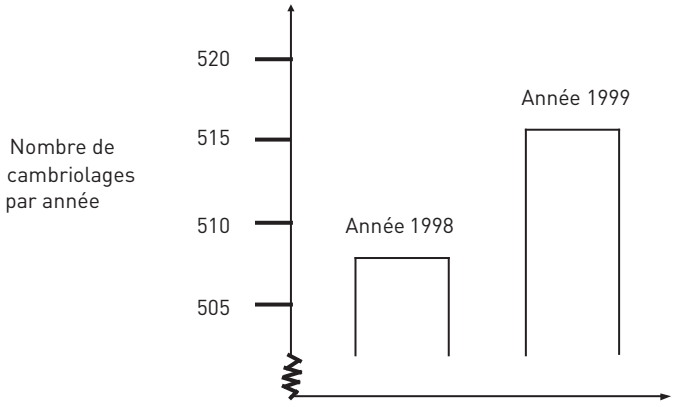
De combien de millilitres (ml) d'huile pour salade avez-vous besoin pour préparer 150mL de cette vinaigrette ?

Réponse : ..... mL

▸ Figure 7 Item libéré, classé dans la catégorie « adaptation », PISA 2000

Question 1 : CAMBRIOLAGES

Lors d'une émission télévisée, un journaliste montre ce graphique et dit :
« Ce graphique montre qu'il y a eu une très forte augmentation du nombre de vols entre 1998 et 1999. »



Considérez-vous que l'affirmation du journaliste est une interprétation correcte de ce graphique ?
Justifiez votre réponse par une explication.

La tâche proposée consiste à croiser deux informations : celle portée par un graphique et celle de l'affirmation d'un journaliste fictif à propos de ce graphique. L'analyse *a priori* de la tâche montre que le graphique laisse apparaître une différence importante de hauteur entre les deux barres représentant les cambriolages en 1998 et en 1999, et que l'élève doit relativiser cette information visuelle en se référant à l'axe des ordonnées dont l'origine n'est pas sur le graphique : le nombre de cambriolages passe de 507 à 516, il augmente de moins de 2 %. L'élève doit donc adapter l'information visuelle du graphique qui devrait être celle à percevoir (c'est en effet le rôle d'un graphique) pour lui associer une variation quantitative précise. L'adaptation est légèrement induite par l'énoncé qui ne demande pas une lecture directe du graphique, mais de juger de la qualité de l'interprétation de ce graphique par un journaliste.

L'item appartient donc à la catégorie de ceux nécessitant la mise en fonctionnement d'une connaissance (la lecture d'un graphique en barres) avec adaptation. Il est associé au domaine *« incertitude et données »* et le processus psycho-cognitif requis est *« interpréter »*. Comme l'analyse précédente le laissait prévoir, bien que la lecture directe des effectifs d'un diagramme en barres soit une connaissance acquise par de nombreux élèves de 15 ans scolarisés en France, l'item n'est pas bien réussi : moins d'un quart des élèves (24 %) trouve le commentaire erroné, et moins de 10 % peuvent expliquer la raison de cette erreur. Ici aussi, les analyses produites en référence à la didactique des mathématiques enrichissent celles de PISA.

Pour terminer l'illustration des catégories de notre classification, analysons un item nécessitant que l'élève, à son initiative, introduise des intermédiaires au cours de la résolution du problème. Cet item est présenté **figure 8**, il s'agit d'une question relative à une situation de vente de CD.

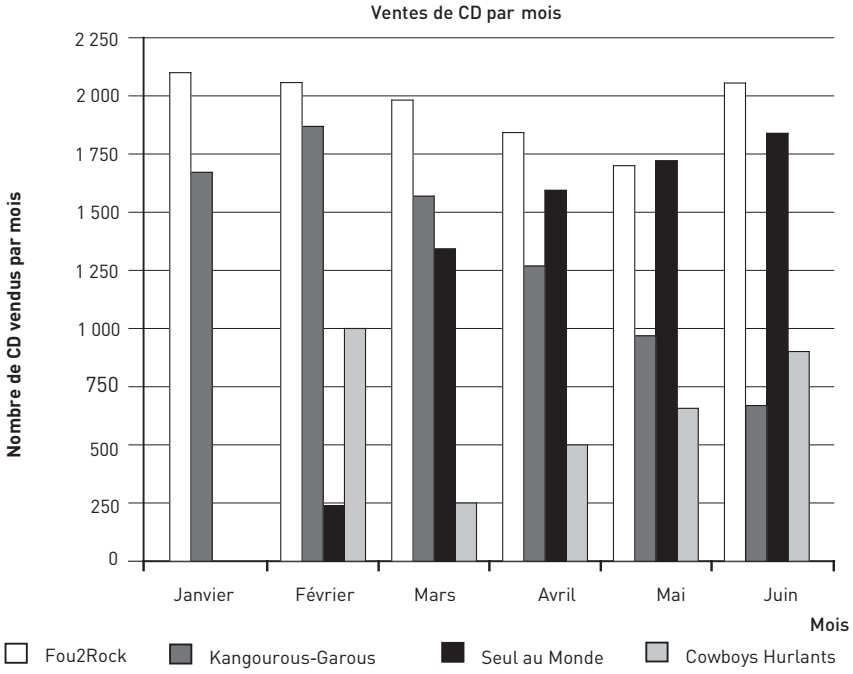
Pour répondre à cette question, l'élève doit obtenir la valeur des ventes à partir de l'extrapolation du graphique, ce qui nécessite l'introduction d'un intermédiaire : tracer une droite de régression si une méthode graphique est mise en œuvre ou, numériquement, effectuer un calcul des différences observées chaque mois, soit par les différences de vente entre chaque mois, soit en divisant la différence globale (environ 1 000) par 4 et l'appliquer au mois de juin. Il est à remarquer que le format QCM peut aussi conduire les élèves à trouver leur réponse par élimination des autres propositions.

En procédant ainsi, après avoir éliminé la réponse D qui ne correspond pas à une baisse, les élèves ont encore à choisir entre les propositions A, B et C. Le fait que les chiffres de ventes proposés en A et C correspondent à une baisse soit trop importante (A) soit pas assez importante (C) invitera sans doute les élèves procédant par élimination à choisir la réponse B qui est la réponse correcte. Il est donc très vraisemblable que les élèves auront choisi cette réponse après calcul de la baisse, à leur initiative, et application de cette baisse. L'item appartient donc bien à la catégorie de ceux nécessitant la mise en fonctionnement d'une connaissance (la lecture d'un graphique en barres) avec introduction d'intermédiaire (la valeur de la baisse). Cet item du domaine *« incertitude et données »* dont le processus psycho-cognitif associé est *« employer »* est réussi par 76,7 % des élèves de l'OCDE (80,6 % en France). Les analyses didactiques permettent de comprendre que l'item ne soit réussi que par les trois quarts des élèves, bien que le calcul à mettre en œuvre soit très élémentaire : l'initiative à prendre rend

la tâche d'application de la baisse plus difficile que si la baisse avait été donnée ou explicitement demandée.
Notre classification des items, à partir d'une analyse didactique de la mise en fonctionnement des savoirs mathématiques requise (*concept*, *directe*, *adaptation* ou *intermédiaires*) a permis d'éclairer les résultats des élèves à quelques items analysés. Une étude systématique de l'ensemble du questionnaire a été réalisée afin de mieux connaître les questions posées aux élèves et de mieux comprendre les résultats de l'enquête PISA 2012.

Figure 8 **Item libéré, classé dans la catégorie « intermédiaires », PISA 2012**

En janvier, les groupes *Fou2Rock* et *Kangourous-Garous* ont chacun sorti un nouveau CD. En février, c'était au tour des groupes *Seul au Monde* et *Cowboys Hurlants* de sortir chacun leur CD. Le diagramme suivant montre les ventes de ces CD de janvier à juin.



Question : HIT-PARADE
Le producteur des *Kangourous-Garous* s'inquiète car le nombre de CD qu'ils ont vendus a diminué de fevrier à juin.
À combien peut-on estimer leurs ventes du mois de juillet si cette tendance à la baisse continue ?

A : 70 CD B : 370 CD C : 670 CD D : 1340 CD

ANALYSES COMPLÉMENTAIRES DU QUESTIONNAIRE DE CULTURE MATHÉMATIQUE ET DES RÉSULTATS DE PISA 2012

Nous présentons une étude de l'ensemble des items du questionnaire puis nous examinons les résultats des élèves selon les niveaux requis de mise en fonctionnement des connaissances mathématiques.

Analyse complémentaire du questionnaire PISA 2012 à l'aide de la classification des items selon les niveaux requis de mise en fonctionnement

La répartition des 85 items de mathématiques de PISA 2012, selon le niveau requis de mise en fonctionnement des connaissances, confirme que l'OCDE vise essentiellement l'évaluation de la capacité à utiliser ses acquis en tant qu'outil dans des situations issues de la vie réelle plutôt que l'acquisition de notions pour elles-mêmes en tant qu'objet d'étude. Seulement 7 items concernent en effet la compréhension qualitative d'un concept. Les 78 autres se répartissent assez équitablement selon les trois niveaux de mise en fonctionnement : on en dénombre 29 de la catégorie regroupant les items nécessitant la mise en œuvre directe d'une procédure connue, 27 exigeant une adaptation de l'énoncé et 22 nécessitant de prendre l'initiative d'introduire des intermédiaires.

Nous avons ensuite mené une analyse croisée de la répartition des items suivant, d'une part, la nouvelle catégorisation proposée par la DEPP et, d'autre part, une catégorie de l'OCDE : le domaine mathématique d'abord, et le processus psycho-cognitif ensuite. Le croisement avec le domaine mathématique conduit au **tableau 1** où figurent, dans chaque case, l'effectif des items, à gauche, et le pourcentage-ligne, entre parenthèses à droite. En cas d'indépendance entre le niveau de mise en fonctionnement des connaissances et le domaine mathématique évalué, nous devrions observer des pourcentages globalement identiques au sein de chaque colonne.

La dernière colonne du tableau montre la volonté des experts de PISA 2012 de répartir les questions mathématiques de manière équivalente suivant chacun des quatre domaines (21 ou 22 items par domaine, soit un quart des 85 items du test complet). Alors que le niveau de mise en fonctionnement des connaissances se détermine indépendamment des contenus mathématiques, les écarts qui apparaissent dans le tableau 1 révèlent que ces différents niveaux ne sont pas évalués indépendamment des champs mathématiques. Inversement, notre nouvelle classification révèle

Tableau 1 **Domaines mathématiques et niveaux de mise en fonctionnement**

Classifications		DEPP				
		Concept	Directe	Adaptation	Intermédiaires	Total
PISA	Espace et forme	0 (0 %)	2 (10 %)	7 (33 %)	12 (57 %)	21 (100 %)
	Incertitude et données	5 (24 %)	7 (33 %)	7 (33 %)	2 (10 %)	21 (100 %)
	Quantité	0 (0 %)	15 (68 %)	4 (18 %)	3 (14 %)	22 (100 %)
	Variations et relations	2 (9 %)	5 (24 %)	9 (43 %)	5 (24 %)	21 (100 %)
	Total	7 (8 %)	29 (34 %)	27 (32 %)	22 (26 %)	85 (100 %)

que les savoirs en jeu dans les items PISA ne sont pas évalués de manière équivalente puisque les items ne conduisent pas à les mettre tous en fonctionnement aux mêmes niveaux.
Par exemple, ceux du champ *« quantité »* sont essentiellement évaluées par des tâches nécessitant la mise en œuvre directe d'une procédure connue (68 % des items) alors que ceux du champ *« espace et formes »* le sont plus souvent par des problèmes nécessitant l'introduction d'intermédiaires (57 % des items). Une étude analogue a été menée concernant l'évaluation des processus psycho-cognitifs et des niveaux de mise en fonctionnement des connaissances mathématiques ▸ **Tableau 2**.
Quelques écarts apparaissent, qui témoignent du fait que le niveau de mise en fonctionnement des connaissances n'est pas évalué indépendamment des processus psycho-cognitifs et inversement. Ainsi, par exemple, la capacité à prendre l'initiative d'introduire des intermédiaires est davantage testée dans les items où le processus attendu est *« formuler »* et pratiquement jamais dans ceux où l'élève doit *« interpréter »*. De même, la capacité à *« employer »* des connaissances mathématiques est surtout évaluée par des items où c'est une mise en fonctionnement directe de procédure qui est requise.

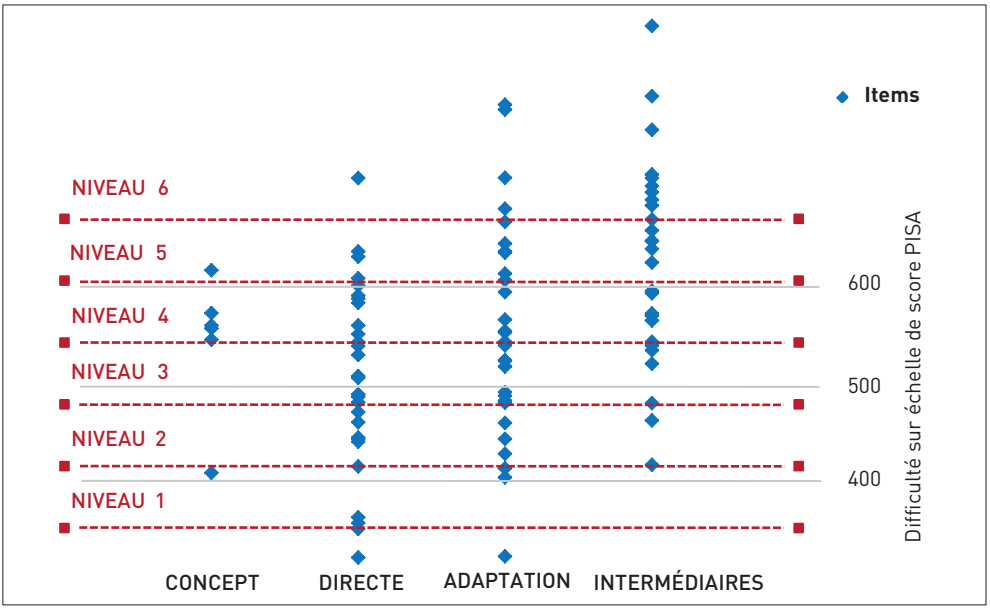
▸ **Tableau 2** **Processus psycho-cognitifs et niveaux de mise en fonctionnement**

Classifications		DEPP				
		Concept	Directe	Adaptation	Intermédiaires	Total
PISA	Employer	1 (3 %)	16 (43 %)	10 (27 %)	10 (27 %)	37 (100 %)
	Formuler	2 (7 %)	6 (22 %)	8 (30 %)	11 (41 %)	27 (100 %)
	Interpréter	4 (19 %)	7 (33 %)	9 (43 %)	1 (5 %)	21 (100 %)
	Total	7 (8 %)	29 (34 %)	27 (32 %)	22 (26 %)	85 (100 %)

Nouveau regard sur les résultats de PISA 2012 apporté par la prise en compte des niveaux de mise en fonctionnement des connaissances

La classification des items selon le niveau requis de mise en fonctionnement des connaissances s'effectue indépendamment de toute mesure de difficulté. Les trois niveaux concernant les items qui portent sur le caractère *outil* des savoirs différencient ces items selon une activité mathématique de plus en plus riche et autonome. Néanmoins, nous avons observé que le niveau de mise en fonctionnement à lui seul ne pouvait expliquer la difficulté d'un item : de nombreux autres facteurs interviennent comme la connaissance en jeu, la familiarité avec le contexte du problème, la lisibilité de l'énoncé, etc. Ainsi, les items de chaque niveau de mise en fonctionnement se répartissent dans presque tous les niveaux de difficulté, tels qu'ils sont définis par PISA. L'étude complète du questionnaire permet de croiser le niveau croissant de mise en fonctionnement des connaissances avec le niveau croissant de difficulté des items ▸ **Figure 9**.
Nous constatons, d'une part, une dispersion relativement importante de la difficulté des items de chaque niveau de mise en fonctionnement, ce qui confirme que ce critère n'est pas suffisant pour prévoir la difficulté d'un item. Le graphique montre aussi, d'autre part, qu'en moyenne les niveaux « directe », « adaptation »

Figure 9 Difficulté des items selon le niveau de mise en fonctionnement des connaissances



Lecture : parmi les 9 items de culture mathématique d'un niveau de difficulté inférieur à 2 sur une échelle de 6 niveaux, on compte 1 item de la catégorie « concept », 5 items « procédures directes », 3 items « adaptation », aucun item « intermédiaire ».
Sources : OCDE ; MENESR-DEPP.

et « intermédiaires », qui correspondent à une exigence croissante de l'activité mathématique, correspondent également à une difficulté croissante pour les élèves. Les items de ces trois niveaux sont en effet réussis en moyenne par respectivement 59,3 %, 46,8 % et 33,9 % des élèves scolarisés en France et, de manière comparable, par 59,8 %, 45,1 % et 34,8 % des élèves scolarisés dans les pays de l'OCDE. Signalons enfin que le cas des items de la catégorie « concept » n'est pas examiné, car leur effectif dans le questionnaire PISA est trop faible pour permettre des interprétations. Complétons cette étude globale par une étude des sous-groupes respectivement définis selon le sexe, la catégorie socioprofessionnelle et le retard scolaire des élèves de 15 ans scolarisés en France.

La publication de PISA sur les réussites aux items de culture mathématique révèle notamment que les filles scolarisées en France, en moyenne, réussissent moins bien que les garçons : la différence de réussite est de 2,5 pp (points de pourcentage) à l'avantage des garçons. L'étude des niveaux de mise en fonctionnement des connaissances apporte quelques informations supplémentaires. L'écart de performance à la faveur des garçons est de 1,5 pp pour les items qui requièrent la mise en œuvre directe d'une procédure connue et de 3,3 pp pour ceux qui nécessitent l'introduction d'un intermédiaire. Autrement dit, les filles sont d'autant plus en difficulté par rapport aux garçons que le niveau requis de mise en fonctionnement des connaissances est un niveau exigeant. Les travaux de Baudelot et Establet [1992] réalisés il y a plus de vingt ans, notamment à partir des résultats d'évaluations à grande échelle, montraient l'existence d'une double culture (culture du respect des

règles chez les filles et culture de la compétition chez les garçons) défavorable aux filles dans l'enseignement secondaire. Les constats indiqués précédemment à partir des résultats de PISA 2012 ne peuvent manquer de conduire à s'interroger sur la capacité du système éducatif français à former de manière équitable les filles et les garçons sur tous les types de tâches mathématiques, et cela d'autant plus que des différences d'attitudes ont été notées vis-à-vis des mathématiques : en France, 42 % des filles déclarent devenir très nerveuses dès qu'il faut résoudre un problème mathématique contre seulement 30 % des garçons, et 58 % des filles pensent que les mathématiques sont importantes pour leurs futures études contre 69 % des garçons [OCDE, 2014].

Une étude analogue a été menée concernant le lien entre les professions et catégories socioprofessionnelles (PCS) des élèves et leur réussite. Un des constats majeurs de l'étude PISA 2012 pour la France est que notre système éducatif est fortement différenciateur : les élèves issus de milieux défavorisés obtiennent une performance moyenne de 39,4 % de réussite contre 57,4 % pour ceux de milieux favorisés, soit un écart de 18 pp. En outre, un tel écart de réussite est constaté pour tous les items, sa valeur allant de 1,9 pp pour le plus faible à 31,9 pp pour le plus élevé. L'étude complémentaire menée par la DEPP montre, contrairement à ce qui a été constaté concernant les différences entre filles et garçons, que les différences de réussite selon les CSP restent stables lorsque le niveau requis de mise en fonctionnement des connaissances augmente.

Autrement dit, les élèves de milieu défavorisé n'apparaissent pas plus désavantagés que ceux de milieu favorisé par l'exigence d'autonomie mathématique requise dans les items. Un tel résultat intéresse les didacticiens des mathématiques qui s'interrogent sur les pratiques enseignantes [Robert et Rogalski, 2002 ; Roditi, 2005 ; Vandebrouck, 2008] et notamment en éducation prioritaire où les élèves de milieux modestes sont particulièrement nombreux [Peltier-Barbier, 2004 ; Charles-Pézard, Butlen, Masselot, 2012]. Les recherches montrent en effet que les enseignants proposent alors davantage de tâches qui requièrent la mise en œuvre de procédures automatisées plutôt que des activités nécessitant une prise d'initiative. Peut-être faudrait-il envisager que ces élèves, pour s'approprier les notions et les méthodes mathématiques, ont au contraire davantage besoin de tâches où ils ont à s'impliquer et des initiatives à prendre.

Le dernier aspect étudié ici est celui du retard scolaire. L'étude PISA révèle que les élèves ayant redoublé au moins une fois dans leur scolarité obtiennent une réussite moyenne de 28,9 % aux 85 items de mathématiques, alors qu'elle est de 56,0 % pour les autres, soit un écart moyen de 27,2 pp. Pour tous les items, les élèves « à l'heure » réussissent mieux que les élèves « en retard », la différence de réussite allant de 3,1 pp pour la plus faible à 46,7 pp pour la plus élevée. L'étude menée par la DEPP met en lumière le fait que la différence de réussite entre les élèves scolairement « en retard » et les élèves « à l'heure » n'est pas constante lorsque varie le niveau requis de mise en fonctionnement des connaissances. Ainsi, et peut-être contre l'idée que l'on pourrait avoir *a priori*, l'écart de performance est d'autant plus faible que le niveau de mise en fonctionnement est élevé : il est de 22,3 pp pour les tâches nécessitant l'introduction d'un intermédiaire et de 30,6 pp pour celles qui se réalisent par la mise en œuvre directe d'une procédure connue. Autrement dit, les élèves « en retard » sont plus souvent mis en difficulté par des tâches routinières

que par celles qui nécessitent davantage d'initiative. Ici encore, ces résultats invitent à s'interroger sur le système éducatif français et les pratiques des enseignants, en particulier sur les activités proposées aux élèves ayant rencontré des difficultés qui ont conduit à un redoublement au cours de leur scolarité.

CONCLUSION

Les enquêtes PISA visent donc un suivi des acquis scolaires des élèves de 15 ans. En ce qui concerne ceux de la culture mathématique, le choix de l'OCDE est d'évaluer des compétences, c'est-à-dire des capacités à mobiliser ses connaissances pour résoudre un problème en lien avec une situation de la vie réelle. Le cadre théorique qui sous-tend ces enquêtes est déterminant sur la conception des items et sur la nature des acquis effectivement évalués. Différents critères sont définis, auxquels ces items doivent satisfaire, pour que chaque enquête puisse couvrir un large spectre de compétences et qu'elle puisse être un outil efficace de positionnement et de distinction des élèves comme des systèmes éducatifs. Un regard didactique porté sur l'évaluation de 2012 ne peut manquer de pointer que l'OCDE ne se donne les moyens ni de recenser précisément les connaissances acquises des élèves (toutes les connaissances géométriques, par exemple, sont confondues au sein d'un même domaine) ni d'estimer le niveau d'acquisition de ces connaissances. Inversement, les didacticiens qui ont concentré leurs recherches sur les phénomènes d'enseignement et d'apprentissage des savoirs n'ont pas suffisamment développé d'outils théoriques et pratiques pour étudier la question de l'évaluation des connaissances des élèves.

Les auteurs de cet article proposent une nouvelle classification des items permettant de distinguer différents niveaux de mise en fonctionnement des connaissances mathématiques et donc, d'une certaine manière, d'évaluer le niveau d'acquisition de ces connaissances. Ils distinguent ainsi quatre catégories d'items suivant qu'ils nécessitent de faire preuve d'une compréhension qualitative d'une notion, de mettre en œuvre une procédure connue de façon directe, d'adapter les données ou la question de l'énoncé pour pouvoir y répondre, ou bien de faire preuve d'initiative en introduisant des intermédiaires pour résoudre le problème posé. L'analyse de quelques exemples extraits de PISA 2012 montre que cette nouvelle classification permet de différencier des items que les catégories définies par les experts de l'OCDE ne permettent pas de distinguer, qui requièrent pourtant des niveaux différents de mise en fonctionnement des connaissances évaluées et qui conduisent à des scores de réussite significativement différents.

Une étude complète de l'ensemble des items de PISA 2012 a été menée à l'aune de cette nouvelle classification. Elle montre d'une part que l'OCDE évalue peu la compréhension qualitative des concepts mathématiques. Elle montre également que les trois autres niveaux de mise en fonctionnement des connaissances, qui correspondent à une exigence croissante de richesse et d'autonomie de l'activité, correspondent également, en moyenne, à un niveau de difficulté croissant pour les élèves. Puis les auteurs ont focalisé leur attention sur le cas de la France, ils se sont interrogés sur l'enseignement des mathématiques dans leur pays en examinant

les inégalités de performances selon le sexe, l'origine sociale ou le retard scolaire. L'OCDE, dans son rapport, indique une meilleure réussite des garçons ; les auteurs, en s'appuyant sur leur classification, montrent en outre que les filles sont d'autant plus pénalisées que les tâches demandent de l'initiative. Concernant les élèves de milieux populaires comme les élèves en retard scolaire, l'étude s'appuyant sur cette même classification révèle enfin que ces élèves ne sont pas mis davantage en difficulté lorsque les activités attendues d'eux sont plus exigeantes.

Cette étude, réalisée à partir de quelques outils issus de la didactique des mathématiques, apporte des résultats qui permettent de poser un regard nouveau sur PISA et ses conclusions. Ce croisement d'approche – didactique et évaluative – sur les apprentissages scolaires s'avère fructueux. Certains chercheurs tentent depuis quelques années d'approfondir une telle démarche [Vantourout et Goasdoué, 2011 ; Sayac, 2012 ; Chesné, 2014], gageons qu'ils ouvriront de nouvelles perspectives pour notre système éducatif.

BIBLIOGRAPHIE

Baudelot C., Establet R., 1992, *Allez les filles !* Paris, Seuil.

Bodin A., 2009, « L'étude PISA pour les mathématiques. Résultats français et réactions », *La Gazette des Mathématiciens*, n° 120, *Société mathématique de France*, p. 53-67.

Brousseau G., 1998, *La théorie des situations didactiques,* Grenoble, La Pensée Sauvage.

Charles-Pézard M., Butlen D., Masselot P., 2012, *Professeurs des écoles débutants en ZEP. Quelles pratiques ? Quelle formation ?* Grenoble, La Pensée Sauvage.

Chesné J.-F., 2014, *D'une évaluation à l'autre : des acquis des élèves sur les nombres en sixième à l'élaboration et à l'analyse d'une formation d'enseignants centrée sur le calcul mental*, Thèse de doctorat, université Paris-Diderot.

Chevallard Y., 1992, « Concepts fondamentaux de la didactique : perspectives apportées par une approche anthropologique », *Recherches en didactique des mathématiques*, vol. 12, n° 1, La Pensée Sauvage, p. 73-112.

Douady R., 1986, « Jeux de cadre et dialectique outil-objet », *Recherches en didactique des mathématiques*, vol. 7, n° 2, La Pensée Sauvage, p. 5-31.

Duval R., 1995, *Semiosis et pensée humaine : registres sémiotiques et apprentissages intellectuels*, Berne, Peter Lang.

Kilpatrick J., Swafford J., Findel B., 2001, *Adding it up: Helping children learn mathematics,* Washington, National Academy Press, p. 115-135.

OCDE, 2014, *Résultats du PISA 2012 : savoirs et savoir-faire des élèves. Performance des élèves en mathématiques, en compréhension de l'écrit et en sciences,* vol. 1, Paris, OCDE.

OCDE, 2014, *PISA 2012 Results: Ready to Learn: Students' Engagement, Drive and Self-Beliefs,* vol. 3, Paris, OCDE, p. 98-106.

OCDE, 2013, *Cadre d'évaluation et d'analyse du cycle PISA 2012,* Paris, OCDE.

Peltier-Barbier M.-L. (dir.), 2004, *Dur d'enseigner en ZEP,* Grenoble, La Pensée Sauvage.

Robert A., 1998, « Outil d'analyse des contenus mathématiques à enseigner au lycée et à l'université », *Recherches en didactique des mathématiques*, vol. 18, n° 2, La Pensée Sauvage, p. 139-190.

Robert A., Rogalski J., 2002, « Le système complexe et cohérent des pratiques des enseignants de mathématiques : une double approche », *La revue canadienne des sciences, des mathématiques et des technologies*, vol. 2, n° 4, IEPO/Université de Toronto, p. 505-528.

Roditi É., 2012, « Un point de vue didactique sur les questions d'évaluation en éducation » *in* Lattuati M., Penninckx J., Robert A., *Une caméra au fond de la classe de mathématiques,* Besançon, Presses universitaires de Franche-Comté, p.275-289.

Roditi É., 2005, *Les pratiques enseignantes en mathématiques - Entre contraintes et liberté pédagogique,* Paris, L'Harmattan.

Sayac N., 2012, « Évaluations nationales ou internationales : limites et perspectives », a*ctes en ligne du colloque sociologie et didactiques*, Lausanne.

Vandebrouck F., (coord.), 2008, *La classe de mathématiques : activités des élèves et pratiques des enseignants*, Toulouse, Octarès.

Vantourout M., Goasdoué R., 2011, « Correction de dissertations en SES », *Idées économiques et sociales,* n° 63, CNDP, p. 71-78.

Vergnaud G., 1990, « La théorie des champs conceptuels », *Recherches en didactique des mathématiques*, vol. 10, n° 2-3, La Pensée Sauvage, p. 133-169.

ÉVALUATION DES COMPÉTENCES DES JEUNES EN NUMÉRATIE LORS DE LA JDC

Stéphane Herrero
MENESR-DEPP, bureau de l'évaluation des élèves

Thomas Huguet
Lycée international de Saint-Germain-en-Laye

Ronan Vourc'h
MENESR-DEPP, bureau de l'évaluation des élèves

Depuis sa création en 1998, la Journée défense et citoyenneté (JDC, ex. Journée d'appel de préparation à la défense – JAPD), permet d'évaluer chaque année les performances en lecture d'environ 700 000 jeunes. En 2013, une évaluation complémentaire s'est tenue auprès d'un échantillon de 56 000 jeunes afin de mesurer la proportion de ceux qui sont en difficulté dans l'utilisation des mathématiques de la vie quotidienne (numératie) et afin d'observer les recoupements et les différences avec les performances en lecture. Cette étude montre que 9,7 % des jeunes ont des difficultés en numératie. Pour la moitié d'entre eux, ces difficultés sont très importantes. De plus, il apparaît qu'environ 14 % des enquêtés présentent des difficultés dans au moins l'un des deux domaines et que des difficultés en lecture n'en impliquent pas nécessairement en numératie et inversement.
Dépassant la seule quantification des jeunes en difficulté en numératie, cette étude décrit qualitativement, tant sur un plan cognitif que conatif, plusieurs profils concernés. Elle distingue, en particulier, un groupe de jeunes confrontés à l'innumérisme. Elle décrit leurs acquis ainsi que leurs lacunes, sources de profondes difficultés au quotidien.
Elle montre aussi une concentration de ces difficultés dans les régions du nord de la France métropolitaine.
En apportant un éclairage inédit sur la non-maîtrise des mathématiques élémentaires à l'entrée de la vie adulte, cet article pose enfin la question d'une meilleure prise en compte de la numératie dans les dispositifs de remédiation à la difficulté, dans un contexte où la priorité est donnée à la maîtrise de la lecture.

Depuis le 1er janvier 2010, la Journée défense et citoyenneté (JDC) a pris la place de la Journée d'appel de préparation à la défense (JAPD). Passage requis dans le parcours de citoyenneté, elle est obligatoire pour tous les citoyens de nationalité française entre le recensement et l'âge de 18 ans. À l'issue de la session, les jeunes reçoivent un certificat individuel de participation qui leur permet de s'inscrire au permis de conduire, mais aussi aux examens et concours soumis au contrôle de l'autorité publique.

Cette journée est conduite par le ministère de la Défense. Après des formalités administratives, les jeunes suivent plusieurs modules d'informations. Ils passent aussi un module d'évaluation de leurs performances en lecture qui permet d'identifier les jeunes en difficulté et de les orienter vers des structures publiques d'aides.
Trois générations de tests de lecture se sont succédé depuis 1998. Les deux premières se déroulaient en utilisant un support papier. Depuis 2009, il s'agit d'un test informatisé et vidéoprojeté auquel les jeunes répondent à l'aide d'un boîtier électronique. Cette dernière innovation, parce qu'elle apporte une simplification technique significative pour l'organisation des tests, a rendu possible l'élargissement des champs évalués lors de la JDC. C'est dans cette perspective que s'inscrit l'évaluation des compétences dans l'utilisation des mathématiques de la vie quotidienne (numératie) qui s'est tenue à l'automne 2013 auprès de plus de 56 000 jeunes et dont la préparation a été initiée au début de l'année 2010.

Outre l'organisation et la préparation logistique, le travail a consisté à bien définir le cadre théorique de la numératie tout en développant un corpus d'items destinés à évaluer cette compétence.
La conception du cadre a tenu compte des programmes d'enseignement scolaires, de la variété des profils des jeunes concernés, des évaluations nationales et internationales des acquis portées par la direction de l'évaluation de la prospective et de la performance (DEPP), d'évaluations conduites dans d'autres pays ou organismes internationaux, ainsi que des travaux de recherche dans un spectre multidisciplinaire.
Un premier corpus de 130 items a ainsi été élaboré, en s'appuyant fortement sur des tests standardisés utilisés au primaire et au secondaire en Californie entre 2003 et 2009. Dans le cadre d'une pré-expérimentation, il a été testé sur un premier échantillon de plus de 1 000 jeunes ayant passé leur JDC à la fin de l'année 2010.
Dans un second temps, de septembre 2011 à décembre 2012, un groupe réunissant des représentants des inspections générales de l'enseignement primaire et secondaire, de la formation, de l'enseignement et de la recherche, a repris l'ensemble de l'outil d'évaluation afin de l'amender et d'assurer sa cohérence avec les réalités quotidiennes et professionnelles.

L'évaluation de la numératie réalisée en 2013 lors de la JDC est le fruit de l'ensemble de ce travail. Elle s'appuie sur un corpus de 66 items finalement retenus. Cet article se propose d'en présenter les principaux résultats. On cherchera tout d'abord à définir le concept de numératie et à présenter la mise en œuvre de son évaluation lors de la JDC. On décrira ensuite les performances des jeunes au test et, plus précisément, les difficultés rencontrées dans l'utilisation des mathématiques au quotidien. Ces performances seront mises en perspective avec les résultats issus d'autres évaluations. Les jeunes qui ont répondu au test de numératie ayant aussi passé le module de performance en lecture habituellement proposé lors de la JDC,

on s'attachera à décrire dans quelle mesure ces deux compétences se recouvrent. Enfin, les résultats seront analysés au regard des données de contexte disponibles concernant les jeunes ayant participé au test.

Qu'est-ce que la numératie ?

Le terme de numératie est un néologisme dérivé de l'anglais. Il recouvre les compétences numériques et mathématiques utilisées dans la vie quotidienne. Parce qu'il présente un caractère relatif de nouveauté, ce concept appelle une description détaillée, qui vise non seulement à préciser sa définition, mais aussi à éclairer les choix faits dans cette étude. L'une des priorités est d'inscrire le concept de numératie dans un contexte international, dans lequel il émerge depuis au moins deux décennies. Il apparaît alors comme le point de convergence entre des travaux de recherches et des démarches d'organismes ou d'institutions, soucieux des enjeux de développements sociaux et économiques sous-jacents.

Dans le domaine de la recherche, la numératie se situe à la confluence de disciplines variées. Elle prend en compte les questions de « support biologique » de l'apprentissage (quel fonctionnement du cerveau ? Quelles conséquences sur l'apprentissage ?) dans ce qu'il a de spécifique à l'être humain, mais aussi comparativement à d'autres espèces animales. Elle concerne le développement de l'enfant. Elle se caractérise par ses dimensions sociales (prise en compte notamment des sociétés traditionnelles), épistémologiques (histoire des sciences, didactique des mathématiques), historiques (nature et conditions d'émergence des savoirs) et étymologiques (histoire du langage mathématique).
Démarquée, en raison de ses problématiques, du vaste champ des mathématiques tout en lui étant fortement liée, la numératie s'intéresse aussi bien aux élèves qu'aux adultes de tous âges, considérés dans des dimensions cognitives, affectives, conatives et motivationnelles.
Elle concerne à la fois les enseignants, les apprenants, mais aussi l'ensemble des utilisateurs. Enfin, elle touche à une large variété de contextes tant scolaires qu'extrascolaires et de portées aussi bien locales (salle de classe ou école) que globales (population d'une région ou d'un pays).

Chez les enseignants, l'étude de la numératie vise à décrire les pratiques enseignantes et leurs impacts, en abordant les questions de formation initiale et continue. Chez les apprenants, élèves ou adultes en formation, l'étude de la numératie vise à rendre compte des acquis mathématiques (connaissances et compétences), mais s'intéresse aussi aux rapports qu'ils entretiennent avec ceux-ci. Chez les utilisateurs, l'étude de la numératie est porteuse d'enjeux sociaux majeurs. Elle propose en effet un éclairage nouveau sur l'adéquation des formations aux situations qu'ils rencontrent dans leur vie quotidienne, citoyenne et professionnelle.

La numératie couvre notamment les aptitudes à mathématiser le réel, à travailler seul et avec les autres, à mobiliser des connaissances dans une large variété de situations qui ne nécessitent pas nécessairement un passage par l'écrit. Elle s'appuie sur des compétences heuristiques et des connaissances d'ordre stratégique (maîtrise d'un répertoire de problèmes types) [Artigue, 2004]. La numératie comprend aussi la faculté à valider sa démarche en raisonnant logiquement, en employant des

raisonnements hypothético-déductifs ou en contrôlant un résultat par une seconde démarche de nature différente. Elle peut être définie par un corpus mathématique détaillé qui est résumé en annexe.
La numératie porte aussi sur le positionnement entretenu par les personnes vis-à-vis de ce corpus mathématique : confiance en soi et en ses capacités, goût et intérêt, affect, appétence, envie d'apprendre, etc. Les innombrables situations quotidiennes, citoyennes ou professionnelles dans lesquelles la numératie s'exprime ne permettent pas d'en faire une liste exhaustive. Il est cependant possible d'indiquer des exemples de tâches (remplir un formulaire, gérer un budget prévisionnel, vérifier l'addition dans un restaurant ou une facture, calculer une taxe, calculer un pourcentage d'évolution sur un prix, faire la cuisine, bricoler, etc.) ou de métiers, dont on imagine pour chacun d'eux tout ce qu'il regroupe comme variétés de compétences caractérisant la numératie (plombier, maçon, charpentier, boulanger, marin pêcheur, taxi, agriculteur, enseignant, etc.).

CONSTRUCTION DU TEST

Profil des participants au test

En septembre et en octobre 2013, 56 000 jeunes hommes et femmes de 17 ans ou plus, de nationalité française, ont pris part à l'évaluation. Elle s'est déroulée lors de toutes les sessions de la JDC qui se sont tenues en France métropolitaine à cette période. Ces jeunes ont aussi passé le test de lecture effectué dans ce cadre depuis 1998, permettant ainsi le croisement des performances obtenues dans ces deux domaines.
Leurs caractéristiques sont très proches de celles de l'ensemble des jeunes qui se sont présentés à la JDC en 2013 ▸ **Tableau 1**. Les moyennes d'âges sont comparables même si la répartition est différente de celle observée lors d'une année complète[1]. La moitié d'entre eux a un niveau de scolaire[2] qui relève des études générales ou technologiques au lycée. Ils sont un tiers à suivre ou à avoir suivi une formation professionnelle en préparant un CAP, un BEP ou un baccalauréat professionnel. Enfin, environ 8 % se sont engagés dans des études supérieures, alors qu'un peu plus de 3 % n'ont pas été au-delà de la scolarité au collège.
Leurs profils de lecteurs sont aussi très comparables à ceux de la population de référence : plus de 80 % sont des « lecteurs efficaces », mais un peu plus de 8 % ont de très faibles capacités, voire des difficultés sévères en lecture.

Contenu du test

Comme pour l'épreuve de lecture, le test de numératie est constitué d'un diaporama où chaque consigne est lue et affichée à l'écran de façon à ne pas freiner les mauvais lecteurs dans leurs calculs. Le test visant à évaluer les compétences en calcul mental,

1. Cette différence dans la répartition par âge s'explique par un effet de saisonnalité des convocations à la JDC. Quelle que soit l'année, en septembre-octobre, période pendant laquelle se sont tenus les tests de numératie, les jeunes âgés de 17 ans représentent plus de 70 % des participants.
2. Le niveau scolaire a été défini en fonction des formations que les jeunes déclarent suivre. Pour les jeunes qui ne sont plus en études lors de leur passage à la JDC (environ 9 % de l'ensemble) c'est la dernière formation suivie qui est prise en compte.

Tableau 1 **Profil des participants aux tests de la JDC en 2013 (en %)**

	Jeunes ayant passé le test de numératie (n = 56 650)	Ensemble des jeunes passés en JDC en 2013 (n = 720 391)
Sexe		
Masculin	51,0	51,2
Féminin	49,1	48,8
Structure par âge		
16 ans	0,2	0,2
17 ans	77,0	51,9
18 ans	17,1	39,2
19 ans ou plus	5,7	8,7
Âge moyen	17,4	17,6
Niveau scolaire		
Collège	3,1	3,3
CAP-BEP	12,3	12,5
Bac professionnel	25,3	25,0
Lycée général et techno.	51,3	50,6
Enseignement supérieur	7,9	8,6
Profils de lecteurs		
Difficultés sévères	3,3	3,5
Très faibles capacités	5,0	5,0
Lecteurs médiocres	7,9	8,4
Lecteurs efficaces	83,8	83,1

Lecture : parmis les jeunes ayant passé le test de numératie, on compte 51 % de garçons.
Note : par le jeu des arrondis, les totaux des colonnes de gauche peuvent être légèrement différents de 100 %.
Champ : France métropolitaine.
Sources : ministère de la Défense–DSN ; MENESR-DEPP.

l'usage de la calculatrice ou de tout autre support n'est pas autorisé. Il se compose de sept parties constituées de questions à choix multiples (QCM) auxquelles les jeunes doivent répondre grâce à un boîtier électronique. Le test comprend 66 items au total afin de couvrir le plus large champ possible (voir des exemples d'items dans l'**encadré p. 264**). Le test est passé par les jeunes comme suit :
– 16 items de **calculs dictés**, répartis sur deux épreuves, qui visent à évaluer la capacité à valider ou invalider le résultat d'un calcul. Ces épreuves sont les mieux réussies. Le format binaire des questions (validation ou non d'un résultat) explique cette réussite. La nature et l'écriture des nombres (entiers, écriture décimale ou fractionnaire) classent ces items par difficulté, plus que l'opération elle-même.
– 11 items consacrés à **l'écriture des nombres** qui vérifient les capacités des jeunes à passer d'une écriture des nombres à une autre (écriture littérale, décimale, fractionnaire, décomposée en puissances de 10). Là encore, la difficulté des items réside dans la nature des nombres.
– 29 items répartis sur deux séries de **problèmes** qui relèvent de la vie courante et traitent des champs mathématiques suivants : proportionnalité, travail sur les grandeurs usuelles, situation additive, multiplicative, traitement de données, probabilités. Les amorces sont très variées (texte brut, diagramme, tableau, figure, image) et l'énoncé est lu.
– 7 items consacrés aux **procédures** où il s'agit de compléter une suite de nombres, de compléter une opération à trou, ou bien d'utiliser des rudiments d'algèbre.

EXEMPLES D'ITEMS DU TEST DE NUMÉRATIE

1. Calcul dicté :

98

❶ Oui ❷ Non

Consigne orale : « 62 plus 35 »
Temps de réponse imparti : 10 secondes.

2a. Écriture des nombres :

Neuf cent quatre-vingt-sept peut s'écrire aussi...

❶ 900 + 87 + 7
❷ 700 + 80 + 9
❸ 980 + 70 + 0
❹ 900 + 80 + 7

Consigne orale :
« Neuf cent quatre-vingt-sept peut s'écrire aussi... »
Temps de réponse imparti : 10 secondes.

2b. Écriture des nombres :

$\frac{3}{4}$ peut s'écrire aussi...

❶ 0,34
❷ 0,75
❸ 3,4
❹ 1,33

Consigne orale :
« 3/4 peut s'écrire aussi... »
Temps de réponse imparti : 15 secondes.

3. Problème :

Vanessa a acheté des pommes pour un montant de 2,50 €.
Elle a payé avec un billet de 10 €.
Quel calcul permet de trouver le montant rendu par le commerçant ?

❶ 10 + 2,50
❷ 10 - 2,50
❸ 10 x 2,50
❹ 10 + 2,50

Consigne orale :
« Vanessa a acheté des pommes pour un montant de 2,50 €. Elle a payé avec un billet de 10 €. Quel calcul permet de trouver le montant rendu par le commerçant ? »
Temps de réponse imparti : 20 secondes.

4. Procédure :

Compléter l'égalité par le nombre qui convient
2 + 7 = ? x 3

❶ 2
❷ 3
❸ 6
❹ 9

Consigne orale :
« Compléter l'égalité par le nombre qui convient ».
Temps de réponse imparti : 13 secondes.

Les jeunes ont tous passé les 66 items du test et les temps de réponse ont été enregistrés ▸ **encadré** p. 268. Les analyses psychométriques mises en œuvre ont ensuite conduit à exclure 3 items[3]. Les résultats présentés dans cet article portent donc sur 63 items.

À ces épreuves s'ajoute un court questionnaire visant à recueillir des informations sur la situation scolaire des jeunes et leur jugement sur la difficulté du test.
Les deux épreuves de calculs dictés n'ont que deux modalités (vrai ou faux). Les quatre autres épreuves cognitives sont des QCM comprenant au moins quatre modalités de réponse. Ces deux types de formats de QCM n'imposent pas la conduite d'un raisonnement hypothético-déductif linéaire depuis les données de l'énoncé vers la bonne réponse. Différentes modalités étant affichées, il est aussi possible d'orienter une réponse en procédant par élimination des modalités jugées les moins plausibles. Les analyses conduites dans cette étude tiennent compte de la possible utilisation de cette deuxième stratégie de réponse.
Le dispositif permet de réduire très nettement les effets liés à la variabilité des conditions d'administration du test entre les différents centres de passation et garantit une grande fiabilité des données recueillies [De La Haye, Gombert *et alii*, 2010].

GROUPES DE PERFORMANCE

Constitution des groupes

À partir des réponses apportées par les jeunes aux items du test, une échelle de performance a été élaborée en utilisant un modèle de réponse à l'item. Selon la théorie relative à ce type de modèle, les scores des élèves et la difficulté des items sont mesurés sur une même échelle [Rocher, dans ce numéro, p. 37]. Cela permet de constituer des groupes de niveaux et de leur associer des ensembles d'items de difficulté croissante.

Sur la base des résultats estimés et proposés par le modèle de réponse à l'item, les items du test de numératie ont été initialement classés par ordre de difficulté croissante. Les items du début de la liste correspondaient à des items faciles, c'est-à-dire très réussis, et ceux de la fin étaient les plus difficiles. Un collège d'experts (inspecteurs généraux, inspecteurs d'académie, inspecteurs pédagogiques régionaux [IA-IPR], chercheurs et enseignants) n'ayant pas d'autre information sur ces items que leur ordre de classement, a ensuite été chargé d'identifier les items constituant de véritables seuils qualitatifs entre les différents groupes. Cette démarche, qui diffère de celle d'autres études telles que Cedre (Cycle des évaluations disciplinaires réalisées sur échantillon) ou PISA (*Programme for Internationnal Student Assessment*, Programme international pour le suivi des acquis des élèves) dans lesquelles les seuils sont définis *a priori*, est inspirée de la méthode des marque-pages décrite par Miconnet et Vourc'h [dans ce numéro, p. 141].
Pour la mettre en œuvre, les experts ont tout d'abord dû définir deux seuils en se référant aux questions suivantes : « *Parmi les items suivants, classés par ordre de difficulté croissante, jusqu'à quel item estimez-vous qu'ils doivent être réussis par tous ?* »

3. Les 3 items exclus présentaient des indices de discrimination trop faibles. Ce qui signifie que ces items étaient aussi bien réussis par les jeunes qui ont eu un score élevé au test que par ceux qui ont eu un score faible.

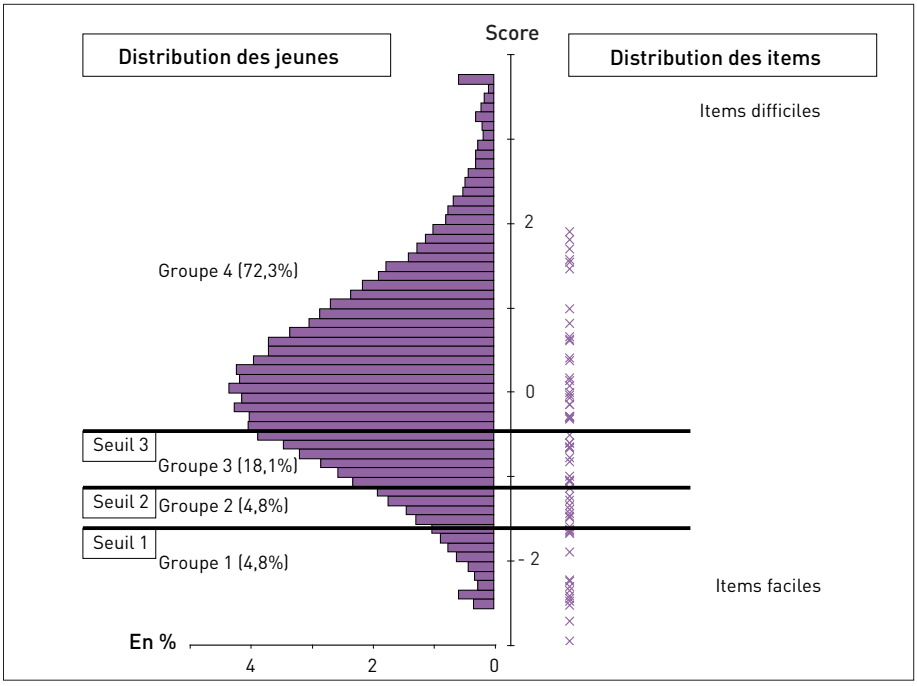
(seuil 1) ; *« Parmi les items suivants, classés par ordre de difficulté croissante, à partir duquel estimez-vous comme (pas trop a)normal qu'une partie significative de la population échoue à les réaliser ? »* (seuil 2). La définition du premier seuil vise à identifier les jeunes qui rencontrent des difficultés dès qu'ils doivent utiliser, dans une situation quotidienne, les mathématiques les plus rudimentaires (groupe 1). Le positionnement du deuxième seuil a pour objectif de discerner les jeunes en difficulté dans l'utilisation des connaissances et compétences de base requises pour conduire un calcul (groupe 2). En remontant la liste des items classés par ordre de difficulté croissante, les experts se sont aussi accordés sur un troisième seuil permettant d'identifier, parmi les jeunes qui ne sont pas en difficulté dans l'utilisation des mathématiques au quotidien, une frange de la population dont les acquis restent fragiles (groupe 3).

La **figure 1** synthétise la démarche mise en œuvre. Dans la partie gauche est représentée la distribution des élèves selon leur score issu de la modélisation, c'est-à-dire le pourcentage d'élèves (longueur de la barre en abscisse) en fonction du niveau de compétence (score en ordonnée). Dans la partie droite, chaque croix représente un item, qui est positionné en fonction de sa difficulté, du mieux réussi, en bas, au moins bien réussi, en haut. Les items les plus faciles sont placés dans le bas du graphique, en face des jeunes les moins performants et, parallèlement, les items les plus difficiles font face aux jeunes les plus performants. Chaque item est positionné à un niveau tel que les jeunes situés à ce niveau ont une chance sur deux de réussir cet item. Ainsi, les jeunes situés au-dessus du seuil ont plus d'une chance sur deux de réussir les items placés en dessous de ce seuil. Il est donc possible de déterminer la proportion de jeunes présents dans chaque groupe et de décrire les tâches qu'ils maîtrisent. Cette représentation met en évidence la gradation dans les acquis, les élèves d'un groupe donné maîtrisant les compétences acquises par ceux des groupes situés en dessous dans l'échelle.
Les deux épreuves de calculs dictés ont aussi servi de support à la détermination de seuils de maîtrise des automatismes de base en calcul. La mesure retenue est le temps moyen observé aux items réussis (97 % des jeunes réussissent plus de la moitié des 16 items proposés). Les jeunes qui présentent un temps de réponse moyen inférieur à la moyenne des temps de réponses augmentée d'un écart-type, sont considérés comme ayant acquis cette maîtrise ▸ **Tableau 2**.

Description des groupes

Une fois les groupes de performance constitués, il est donc possible de décrire les compétences des jeunes qui les composent. Il convient de préciser qu'au sein d'un même groupe, certaines caractéristiques peuvent être plus ou moins accentuées chez un jeune que chez un autre.
Les jeunes les plus en difficulté (groupe 1) forment 4,8 % de la population étudiée. Ils ne réussissent, en moyenne, qu'un tiers des items du test alors que le taux de réussite moyen à l'ensemble de l'évaluation est de 72 %. Ces items concernent en grande partie les épreuves de calculs dictés.
Ils sont en situation de réussite sur des QCM où les nombres mis en jeu sont de petits entiers ou des décimaux simples, mais il semble qu'ils ont une grande difficulté dans la compréhension des nombres ainsi que dans leur écriture. En effet, la décomposition des nombres en centaines, dizaines ou unités semble basée sur la perception orale (*je reconnais ce que j'entends*) et l'écriture décimale peut être

Figure 1 Représentation de la performance des jeunes et de la difficulté des items sur une même échelle



Lecture : les jeunes du groupe 2 (4,8 % de l'ensemble) ont tous plus de 50 % de chances de réussir les items, représentés par des croix, placés en dessous du seuil 1.
Sources : ministère de la Défense-DSN ; MENESR-DEPP.

Tableau 2 Répartition des jeunes dans les groupes et automaticité de calcul à la Journée défense et citoyenneté 2013 (en %)

	Garçons	Filles	Ensemble	Temps de réponse moyen aux épreuves de calculs dictés (en secondes)	Automaticité de calcul		
					Garçons	Filles	Ensemble
Groupe 4 Sans difficulté	75,5	68,9	72,3	4,4	91,6	91,5	91,6
Groupe 3 Acquis fragiles	15,9	20,4	18,1	5,1	68,3	73,1	70,9
Groupe 2 En difficulté	4,2	5,6	4,8	5,4	53,5	62,5	58,5
Groupe 1 En grande difficulté	4,5	5,1	4,8	5,5	46,3	54,3	50,5

Lecture : 75,5 % des garçons appartiennent au groupe 4. Les jeunes de ce groupe mettent en moyenne 4,4 secondes pour répondre aux questions de calculs dictés et pour 91,6 % d'entre eux, le calcul est automatisé.
Note : par le jeu des arrondis, les totaux des colonnes de gauche peuvent être légèrement différents de 100 %.
Champ : France métropolitaine.
Sources : ministère de la Défense–DSN ; MENESR-DEPP.

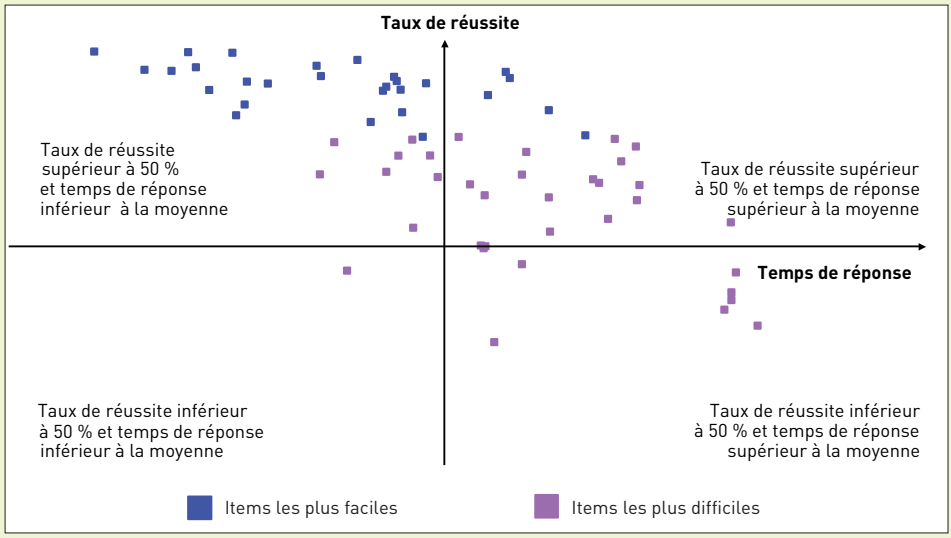
confondue à l'écriture fractionnaire, dans les deux sens : (a/b et a,b).
L'utilisation des nombres dans des contextes courants (argent, température, etc.) ne semble pas les mettre davantage en situation de réussite. Les opérations sur les nombres se limitent à l'addition de nombres entiers ou bien à des calculs simples dont il faut valider ou invalider le résultat. La méconnaissance des tables (addition et multiplication) apparaît comme un obstacle pour des opérations plus complexes. En outre, dès que les nombres décimaux sont moins familiers, le traitement des parties entières et décimales se fait séparément et la gestion de la retenue peut poser des difficultés.
La présence simultanée de plusieurs informations génère des difficultés dans le choix des opérations à effectuer ainsi que dans la gestion des différentes étapes de résolution. Les jeunes de ce groupe peuvent ainsi laisser un calcul inachevé et proposer comme réponse un résultat intermédiaire de raisonnement. Les quelques problèmes qu'ils arrivent à résoudre relèvent donc d'un modèle additif à une ou deux étapes sur des petits nombres entiers.

Par ailleurs, le calcul ou l'utilisation d'un pourcentage leur est inaccessible, même lorsque celui-ci est très simple (50 % de 60 par exemple). Ces jeunes confondent régulièrement le périmètre et l'aire d'une figure. L'utilisation de données sous forme de tableaux ou de diagrammes est restreinte au prélèvement d'informations

TEMPS DE RÉPONSE

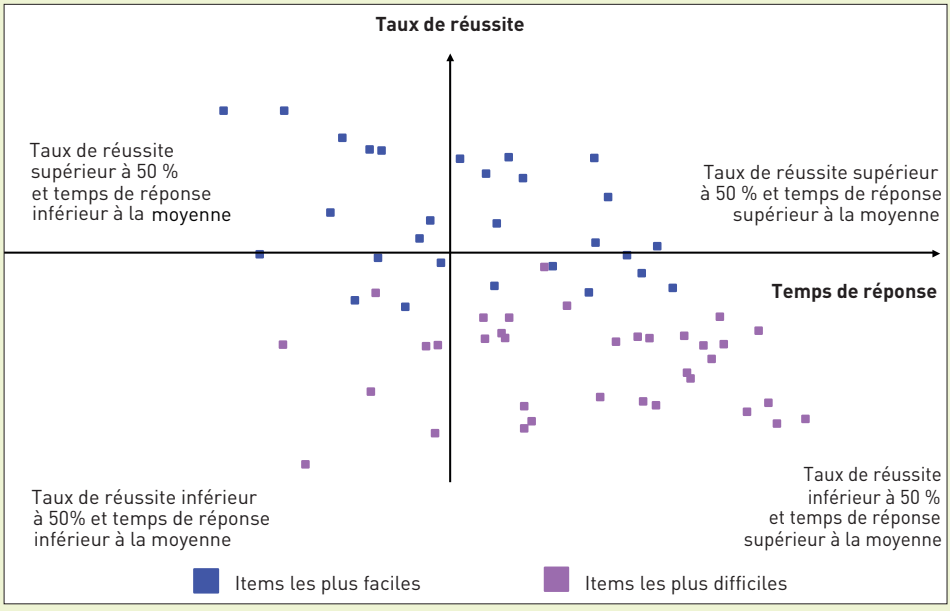
Les conditions de passation du test de la JDC en numératie ont permis de recueillir les temps de réponse des jeunes à chaque épreuve. Les temps de réponse impartis variaient de 10 à 30 secondes selon les items. Le rapport entre le temps de réponse médian et le temps total imparti a été calculé pour chaque item. On dispose ainsi d'une information sur la proportion du temps imparti qui a été mobilisée par les jeunes pour répondre à chaque item.
D'une manière générale, plus le temps de réponse est important, moins le taux de réussite aux items est élevé. Cette corrélation entre le temps de réponse et la réussite aux items apparaît clairement pour l'ensemble des jeunes sur la **figure 2**. La partie supérieure gauche comprend les items les plus faciles, au-dessous du seuil 2, avec des temps de réponse réduits. À l'opposé, la partie inférieure droite comprend exclusivement des items au-dessus du seuil 2 pour lesquels les temps de réponse sont plus élevés. La corrélation est moins visible lorsque l'analyse est limitée aux jeunes des groupes 1 et 2 ▸ **Figure 3**. Les temps de réponse mobilisés pour les items les plus simples sont comparables à ceux observés pour la majorité des items d'un niveau plus élevé. Pour ces jeunes, la grande majorité des items qui relèvent des groupes 3 et 4 se situent dans la partie inférieure droite du graphique.
Déjà plus lents que les autres sur les calculs les plus simples et en l'absence de répertoires de calculs personnels, les jeunes les plus en difficulté évitent de répondre lorsque les items deviennent plus difficiles. De fait, la part de non-réponse augmente à mesure que les compétences des jeunes baissent. Le taux de non-réponse moyen s'élève ainsi à 24,6 % pour le groupe 1 et respectivement à 14,0 %, 10,9 % et 5,1 % pour les trois autres groupes. Le taux de non-réponse moyen est de 7,5 % sur l'ensemble du test. 37 items présentent un taux de non-réponse inférieur à 5 % et seulement 13 items un taux supérieur à 10 %. À titre de comparaison, le taux moyen de non-réponse dans PISA 2012 en mathématiques pour les QCM est de 4,3 % et il y a relativement moitié moins d'items avec un taux de non-réponse supérieur à 10 %. Le temps limité et court (entre 10 et 30 secondes) pour répondre peut expliquer cet écart.

Figure 2 Taux de réussite aux items du test de numératie en fonction du temps de réponse mobilisé (ensemble des jeunes)



Sources : ministère de la Défense-DSN ; MENESR-DEPP.

Figure 3 Taux de réussite aux items du test de numératie en fonction du temps de réponse mobilisé (groupes 1 et 2)



Lecture : chaque point représente un item du test positionné en fonction de son taux de réussite (axe des ordonnées) et du temps de réponse mobilisé (axe des abscisses). Les items les plus faciles sont ceux qui se situent en dessous du deuxième seuil présenté dans la figure 1. Les items les plus difficiles sont situés au-dessus de ce seuil.
Sources : ministère de la Défense-DSN ; MENESR-DEPP.

explicites. Pour les jeunes du groupe 1, le signe « = » ne représente pas forcément une égalité, mais sert à indiquer l'exécution d'un calcul. Plus largement, cette population ne semble pas sensible aux formalismes mathématiques.
Ces jeunes mettent, en moyenne, 5,5 secondes sur les 10 secondes imparties, pour répondre aux questions de calculs dictés contre 4,6 secondes pour l'ensemble des jeunes (tableau 2 p. 267). La moitié d'entre eux n'a pas acquis les automatismes de calcul les plus fondamentaux.

Pour les jeunes de ce groupe, les compétences maîtrisées relèvent majoritairement du niveau fin de CE1. Ils ne semblent donc pas disposer des outils mathématiques requis pour répondre aux besoins de la vie courante. Au travers des difficultés qu'ils peuvent rencontrer au quotidien, ils donnent véritablement un visage à l'innumérisme en France.

Les jeunes en difficulté (groupe 2) regroupent également 4,8 % de la population étudiée. Ils ne réussissent en moyenne que la moitié des items du test et leurs compétences restent limitées. En effet, ils présentent des taux de réussite comparables à ceux du groupe 1 pour les items les plus difficiles. Cependant, relativement au groupe 1, leur connaissance des nombres est élargie, tant sur leur taille que sur celle du sens de l'écriture décimale. Ils sont aussi capables de reconnaître une proportion sur une représentation graphique ou d'utiliser une proportion élémentaire. Ainsi, pour les deux tiers des items, ils ont un taux de réussite supérieur de 10 points aux jeunes du groupe 1. Ils mettent, en moyenne, 5,4 secondes pour répondre aux questions de calculs dictés. 41,5 % d'entre eux n'ont pas acquis les automatismes de base en calcul.

Les aptitudes ainsi que les insuffisances des jeunes des groupes 1 et 2 font écho aux observations issues des évaluations Cedre et PISA. En effet, Cedre-compétences générales 2009 relève que les élèves de fin de collège des groupes 0 et 1 *« sont capables de prélever une information explicite lorsque celle-ci est facilement repérable »* [Bourny, Bessonneau *et alii,* 2010]. Les résultats issus de l'évaluation Cedre-mathématiques 2008 montrent que les élèves de ce groupe *« ne semblent pas avoir tiré bénéfice des enseignements mathématiques du collège. Ils sont en situation de réussite sur des QCM qui ne requièrent, le plus souvent, que des tâches de reconnaissance ou d'identification. Les informations à prélever sont généralement très explicites sur des supports simples. Pour résoudre des problèmes calculatoires, ces élèves tendent à privilégier une approche arithmétique. Les calculs mis en jeu portent sur des nombres entiers, tout en étant isolés »* [Brun et Huguet, 2010].
Quant aux résultats de PISA 2012, ils révèlent que les élèves de 15 ans du niveau 1, *« peuvent répondre à des questions* [...] *où toutes les informations pertinentes sont fournies et dont l'énoncé est clairement défini. Ils sont capables d'identifier les informations requises » et « peuvent exécuter des actions qui vont presque toujours de soi et qui découlent directement du stimulus donné ».* Les élèves sous le niveau 1 *« sont susceptibles de* [...] *lire une valeur dans un graphique ou un tableau.* [...] *Ils sont également capables d'effectuer des calculs arithmétiques avec des nombres entiers. »* [OCDE, 2014]. Les profils décrits sont très proches des groupes 1 et 2 de la présente étude.

Les jeunes aux acquis fragiles, mais sans difficulté majeure (groupe 3) représentent 18,1 % des jeunes et réussissent en moyenne environ 60 % des items du test. La

population de ce groupe connaît et utilise les nombres décimaux relatifs en écriture décimale même si l'écriture fractionnaire, et le passage d'une écriture à une autre posent encore des difficultés. Les réflexes calculatoires semblent plus ancrés et concernent un ensemble de nombres toujours plus élargi. Contrairement aux jeunes des groupes 1 et 2, ils peuvent résoudre des problèmes relevant de situations multiplicatives et nécessitant un traitement en plusieurs étapes ; ces problèmes pouvant être présentés sous forme textuelle, de tableaux ou de diagrammes. Ils sont susceptibles de faire la distinction entre périmètre et aire d'une figure élémentaire et de les calculer. Enfin, beaucoup d'entre eux comprennent les notations algébriques élémentaires et savent remplacer une valeur dans une expression algébrique.
Les temps de réponse moyens aux items de calculs dictés sont légèrement inférieurs à ceux des groupes 1 et 2 (5,1 secondes). Pour un tiers d'entre eux, les automatismes de base en calcul ne sont pas acquis.

Les jeunes sans difficulté (groupe 4) représentent 72,3 % de la population et réussissent en moyenne près de 80 % des items du test. Leur taux de réussite aux items est en moyenne supérieur de 21 points à celui du groupe 3. Pour deux tiers des items, cette différence dépasse 15 points de pourcentage. La population de ce groupe a acquis le sens des nombres décimaux relatifs en écriture décimale ou en lettres. Les calculs sur ces nombres semblent poser peu de problèmes, qu'ils soient proposés hors contexte ou en situation. De plus, ces jeunes peuvent traiter une situation de proportionnalité, appliquer ou calculer un pourcentage simple. Ils sont à même de comprendre les formalismes mathématiques élémentaires. Enfin, ils répondent plus rapidement aux items de calculs dictés (4,4 secondes) et pour 91,6 % d'entre eux, le calcul mental élémentaire est automatisé.

Certaines situations semblent cependant poser difficulté à tous les jeunes de l'étude. Il s'agit de celles relatives à la distinction entre le périmètre et l'aire d'une figure, à l'application et au calcul d'un pourcentage ou au dénombrement de cubes constituant une figure représentée en perspective. L'écriture fractionnaire peut causer des difficultés quand le dénominateur n'est pas simple (écrire des cinquièmes en décimal plutôt que des demis ou des centièmes). Enfin, les taux de réussite des calculs mettant en jeu des additions ou des soustractions sont tous plus élevés que ceux obtenus lorsqu'il faut utiliser des multiplications ou des divisions.

LECTURE ET NUMÉRATIE

Les jeunes qui ont participé au test de numératie ont aussi passé le module de performance en lecture mis en place depuis 1998 lors de la JDC. Les résultats issus de ce test indiquent que 8,3 % des jeunes rencontrent des difficultés en lecture. Pour une partie d'entre eux (3,3 % de l'ensemble) ces difficultés sont très importantes. 7,9 % ont une maîtrise fragile de la lecture et 83,8 % sont des lecteurs efficaces (voir tableau 1 p. 263). Il apparaît tout d'abord que la corrélation entre le test de lecture et le test de numératie est moins forte que ce que l'on peut observer dans d'autres évaluations. Ce résultat décisif renforce le constat de qualité de la mesure effectuée dans le domaine de la numératie. Ainsi, le coefficient de corrélation établi à partir

des tests de la JDC entre le score en numératie et le score en compréhension de l'écrit est de 0,54 alors qu'il s'élève à 0,86 dans PISA pour des compétences comparables. En outre, le SATO[4] moyen pour le test de la JDC est de 4,8 en numératie. Cela signifie qu'il faut environ cinq années d'études après la première année d'élémentaire pour comprendre le texte. À titre de comparaison, la lisibilité moyenne de PISA 2012 en mathématiques est à 7,8 ; chaque amorce contenant entre 17 et 324 mots, pour une moyenne de 113 mots.

Deux raisons principales peuvent donc être avancées pour expliquer la plus faible corrélation entre les deux compétences évaluées lors de la JDC. D'une part, chaque consigne est lue et affichée à l'écran, d'autres part, les consignes des items sont courtes (entre 3 et 38 mots, 14 mots en moyenne) et simples.
Les résultats obtenus au test en numératie varient tout de même de façon significative selon les compétences en lecture : près de 60 % des jeunes en sévères difficultés de lecture sont aussi en difficulté en numératie. Ils sont seulement 5,1 % dans ce cas parmi les lecteurs efficaces. En outre, le score moyen au test, mesuré par le nombre d'items réussis, augmente avec les performances en lecture et ce, même si l'on tient compte d'autres variables contextuelles au moyen d'une régression linéaire (voir tableau 5 p. 275).
Par ailleurs, les résultats croisés de ces deux tests révèlent qu'environ 14 % des jeunes sont en difficulté dans au moins un des deux domaines, quel que soit le sexe ▸ **Tableau 3**. Mais les difficultés en lecture n'impliquent pas forcément des difficultés en numératie et inversement. En effet, 5,8 % des jeunes rencontrent des difficultés uniquement en numératie tandis que 4,5 % des jeunes n'en rencontrent qu'en lecture. Ils sont 3,8 % à cumuler les difficultés dans les deux champs. Parmi les jeunes en difficulté de lecture, 54 % n'éprouvent donc pas de difficulté en numératie. Si, de manière générale, les jeunes qui ne présentent pas de difficulté en lecture réussissent mieux l'évaluation en numératie, le lien entre performances en lecture et performances en numératie est à nuancer selon les groupes. En effet, parmi les jeunes des groupes 1 et 2, ceux qui ne présentent pas de difficulté de lecture

▸ Tableau 3 Difficulté en numératie et en lecture selon le sexe (JDC 2013) (en %)

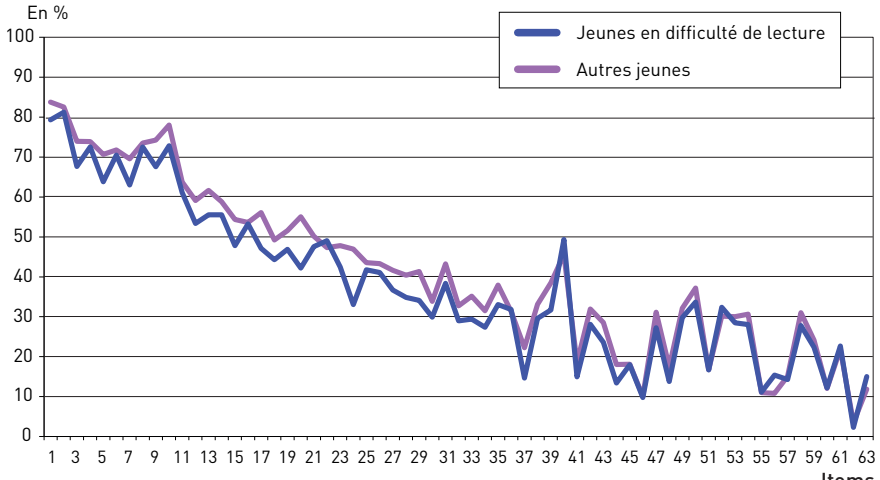
Profils	Filles	Garçons	Ensemble
Difficultés en lecture et en numératie	3,5	4,1	3,8
Difficultés en numératie seulement	7,2	4,5	5,8
Difficultés en lecture seulement	3,3	5,7	4,5
Sans difficulté	86,1	85,7	85,9
Total	100	100	100

Lecture : 4,5 % des garçons sont en difficulté en numératie.
Note : par le jeu des arrondis, les totaux des colonnes de gauche peuvent être légèrement différents de 100 %.
Champ : France métropolitaine (groupes 1 et 2) mais pas en lecture.
Sources : ministère de la Défense–DSN ; MENESR-DEPP.

4. SATO (système d'analyse de texte par ordinateur) est un logiciel d'analyse de données textuelles ou de statistique textuelle. Il permet de déterminer le nombre d'années de scolarisation depuis l'entrée dans l'enseignement primaire qui sont nécessaires à la lecture d'un texte. Ces indicateurs permettent ainsi d'attribuer un niveau à chaque énoncé d'item (indice SATO). Par exemple, un item dont l'énoncé est évalué de niveau 6 correspond à un texte qui n'est pas compréhensible avant l'âge de 13 ans (cinquième). Le logiciel a été conçu par François Daoust de la faculté des sciences humaines de l'université du Québec à Montréal.

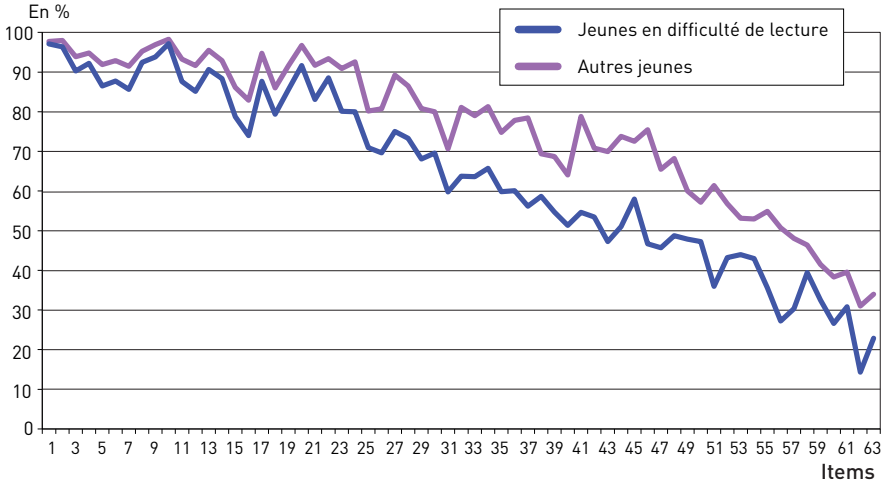
réussissent légèrement mieux les items de mathématiques les plus simples que les mauvais lecteurs, mais, dès lors que les situations se complexifient, les taux de réussite sont équivalents ▸ **Figure 4**. Être « bon lecteur » ne permet pas forcément la réussite à des items compliqués pour ces groupes. En revanche, pour les jeunes des groupes 3 et 4, les écarts de performances sont plus marqués, *a fortiori* pour les items les plus difficiles ▸ **Figure 5**.

▸ Figure 4 Pourcentage de réussite des jeunes des groupes 1 et 2 aux items du test de numératie classés par ordre de difficulté croissante



Sources : ministère de la Défense-DSN ; MENESR-DEPP.

▸ Figure 5 Pourcentage de réussite des jeunes des groupes 3 et 4 aux items du test de numératie classés par ordre de difficulté croissante



Sources : ministère de la Défense-DSN ; MENESR-DEPP.

PERFORMANCES EN NUMÉRATIE SELON LES CARACTÉRISTIQUES INDIVIDUELLES DES JEUNES

Inversement à ce que l'on observe autour des tests de lecture de la JDC [Vourc'h, Rivière *et alii*, 2014], les performances des garçons en numératie tendent à être supérieures à celles des filles. Ces résultats confirment les évolutions observées dans Cedre-mathématiques [Huguet et Everaert, 2012], dans PISA [Keskpaik et Salles, 2013], ainsi que dans d'autres enquêtes sur les compétences des adultes [Jonas, 2012 ; Jonas, 2013]. Les filles réussissent, en moyenne, 69,8 % des items du test contre 73,0 % pour les garçons. Elles sont plus souvent en difficulté que les garçons (10,7 % contre 8,7 %) (voir tableau 2 p. 267) et sont moins représentées dans le groupe 4 (68,9 % contre 75,5 %). Elles sont moins performantes que les garçons pour les épreuves de résolution de problèmes. Leurs résultats s'en rapprochent pour les épreuves de calculs dictés et d'écriture de nombres et ils sont comparables à ceux des garçons pour les questions de procédures ▸ **Tableau 4**. Enfin, les temps de réponse aux questions de calculs dictés sont assez proches quel que soit le sexe : 4,7 secondes en moyenne pour les garçons contre 4,6 secondes pour les filles. Ces dernières sont plus nombreuses à avoir acquis les automatismes de base en calcul dans les groupes 1, 2 et 3. En revanche, dans le groupe 4, les garçons sont plus performants.

Les filles sont deux fois plus nombreuses que les garçons à déclarer préférer le français aux mathématiques (44,2 % contre 21,6 %) y compris dans le groupe 4 (39,2 % contre 18,5 %)[5]. De plus, 5,7 % de filles déclarent avoir trouvé le test très facile contre 17,9 % des garçons, écart qui se maintient dans le groupe 4 (7,3 % contre 21,3 %). Cela confirme les conclusions de Cedre et de PISA qui montrent que le rapport entretenu par les jeunes aux mathématiques est nettement en défaveur des filles qui expriment une plus grande anxiété face à cette discipline, quelles que soient leurs performances.

Une analyse dite « toutes choses égales par ailleurs » au moyen de la méthode de régression linéaire confirme ces différences de performances selon le sexe ▸ **Tableau 5.** En effet, l'écart de score filles/garçons, mesuré par le nombre moyen d'items réussis à l'évaluation en numératie, reste le même, que l'on tienne compte ou pas des autres caractéristiques présentes dans le modèle (écart légèrement

▸ **Tableau 4 Scores moyens par épreuve selon les compétences en numératie et le sexe (JDC 2013)**

	Calcul (score sur 16)			Ecriture de nombres (score sur 11)			Problèmes (score sur 29)		
	Garçons	Filles	Ens.	Garçons	Filles	Ens.	Garçons	Filles	Ens.
Groupe 4 - Sans difficulté	14,6	14,3	14,4	9,2	8,9	9,1	21,5	20,0	20,8
Groupe 3 - Acquis fragiles	12,5	12,4	12,4	6,3	6,4	6,4	14,5	13,9	14,2
Groupe 2 - En difficulté	10,9	10,9	11,0	4,8	4,9	4,9	11,2	10,8	11,0
Groupe 1 - En grande difficulté	8,3	8,5	8,4	3,4	3,4	3,4	7,5	7,5	7,5
Ensemble	13,8	13,4	13,6	8,3	7,9	8,1	19,3	17,6	18,5

5. Ces résultats sont issus des réponses à la question *« Préférez-vous le français ou les mathématiques ? »*, posée lors de la JDC avant l'épreuve de numératie.

Tableau 5 Nombre moyen d'items réussis en numératie selon les caractéristiques des jeunes

	Nombre d'items réussis en moyenne	Coefficient de régression		Nombre d'items réussis en moyenne	Coefficient de régression
Sexe			**Difficulté perçue du test**		
Garçons	46,4	**réf.**	Test très facile	52,2	2,9***
Filles	44,0	- 2,3***	Test facile	46,9	**réf.**
Parcours scolaire			**Test difficile**	**40,0**	- 3,9***
Redoublement	41,1	- 2,4***	Test très difficile	39,5	- 3,9***
Aucun redoublement	47,9	**réf.**	Non-réponse	45,5	- 9,4***
Fratrie			**Âge**		
Enfant unique	45,6	- 0,3***	16 ans	44,2	- 1,6***
1 frère ou sœur	46,5	**réf.**	17 ans	45,9	**réf.**
2 frères et sœurs	45,7	- 0,1*	18 ans	43,6	- 0,2**
3 frères et sœurs	43,8	- 0,7***	19 ans	41,3	- 0,9***
4 frères et sœurs	41,0	- 1,6***	Plus de 19 ans	40,3	- 0,2
Niveau d'études			**Profil de lecteur**		
Collège	32,6	- 9,5***	Sévères difficultés de lecture	29,7	- 10,6***
CAP-BEP	37,6	- 7,1***	Très faibles capacités	34,9	- 7,3***
Bac pro	41,6	- 4,7***	Lecteur médiocre	39,9	- 3,7***
Lycée GT ou sup	49,1	**réf.**	Lecteur efficace	47,0	**réf.**
Préférence disciplinaire					
Préfèrent le français	42,4	- 3,5***			
Préfèrent les maths	48,6	**réf.**			
Aiment autant les 2	47,6	- 0,8***			
N'aiment aucun des 2	42,8	- 2,2***			
Non-réponse	43,6	- 2,4***			

n.s. : non significatif
* significatif au seuil de 0,1
** significatif au seuil de 0,05
*** significatif au seuil de 0,01

Lecture : le score moyen des jeunes en sévères difficultés de lecture est nettement inférieur à celui des lecteurs efficaces (29,7 contre 47). Cet écart de plus de 17 points ne reflète pas véritablement la différence de performance entre ces deux populations. Si, à part leur performance en lecture, les jeunes partageaient les mêmes caractéristiques (situation de référence), l'écart serait ramené à 10,6 points (coefficient de régression).
Sources : ministère de la Défense-DSN ; MENESR-DEPP.

Procédures (score sur 7)			Score total (score sur 63)		
Garçons	Filles	Ens.	Garçons	Filles	Ens.
5,7	5,9	5,8	50,9	49,1	50,1
3,3	3,9	3,6	36,7	36,5	36,6
2,3	2,7	2,5	29,3	29,4	29,3
1,4	1,6	1,5	20,6	21,1	20,9
5,0	5,1	5,0	46,4	44,0	45,2

Lecture : les garçons en situation d'innumérisme (groupe 1) ont obtenu un score moyen aux épreuves de calcul de 8,3 sur 16 items, contre 8,5 pour les filles.
Champ : France métropolitaine
Sources : ministère de la Défense-DSN ; MENESR-DEPP.

supérieur à deux points en faveur des garçons). En revanche, cette analyse montre que les effets liés à l'âge et à la taille de la fratrie s'atténuent lorsque les autres variables sont tenues constantes.

Les jeunes en difficulté en numératie sont de moins en moins nombreux à mesure que le niveau d'études s'élève : ils sont 46,3 % parmi ceux qui n'ont pas dépassé le collège et encore 26,7 % parmi ceux qui ont un niveau CAP ou BEP. Alors que ces deux groupes ne représentent que 15,5 % des jeunes de l'échantillon, ils constituent 49,2 % des jeunes en difficulté. À l'opposé, les jeunes qui suivent ou ont suivi des études secondaires générales ou technologiques, voire une formation d'enseignement supérieur, ne sont que 2,8 % à être en difficulté en numératie.
Ces différences dans les performances des jeunes selon leur niveau d'études apparaissent aussi nettement lorsque l'on s'intéresse au nombre d'items réussis en moyenne (voir tableau 5 p. 275). Il y a en effet un écart de plus de 16 points entre ceux qui n'ont pas dépassé le collège et ceux qui suivent ou ont suivi des études secondaires générales ou technologiques, voire une formation d'enseignement supérieur. Cet écart passe à 9,5 points si l'on raisonne toutes choses égales par ailleurs. D'autre part, les jeunes ayant redoublé au moins une fois pendant leur scolarité réussissent, en moyenne, 41 items du test sur 63 et sont 17,0 % à être en difficulté en numératie. Ceux qui n'ont jamais redoublé réussissent, en moyenne, 7 items de plus (2 items si l'on tient compte des autres caractéristiques) et sont seulement 4,5 % à être en difficulté en numératie.

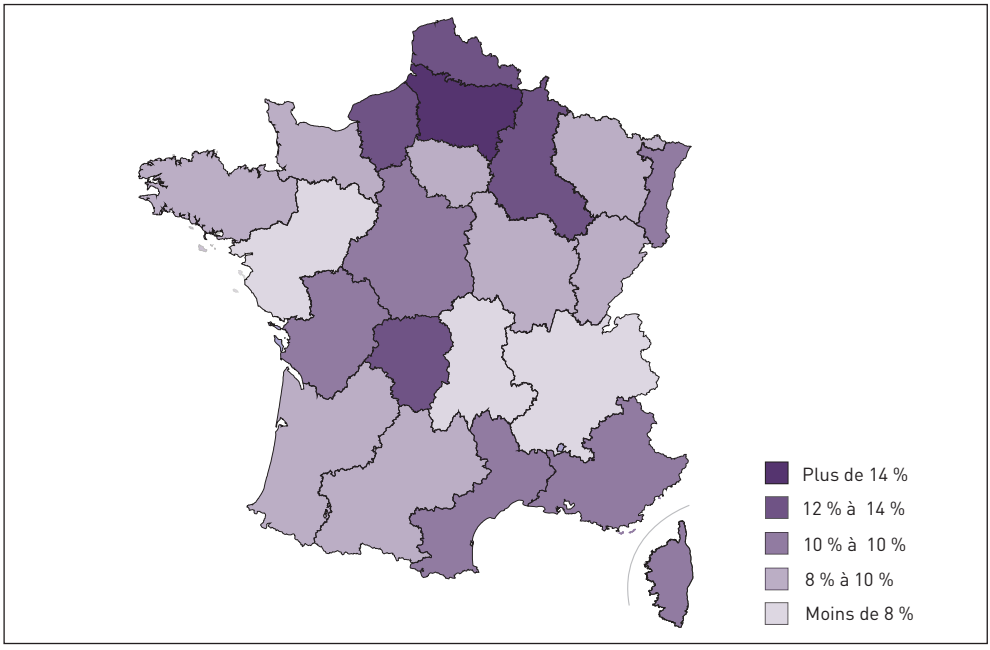
Les jeunes français sont plus nombreux à déclarer préférer les mathématiques au français : 38,1 % contre 32,7 %. Seulement 8,1 % n'ont pas de préférence et une part importante (21,1 %) déclare n'aimer « ni l'un, ni l'autre ». Le lien entre cette information et les performances au test de numératie est très fort. En effet, les jeunes qui préfèrent le français sont pratiquement trois fois plus nombreux à être en difficulté en numératie que ceux qui préfèrent les mathématiques (13,7 % contre 4,9 %). Les jeunes qui déclarent n'aimer aucune de ces deux disciplines sont 13,0 % à être en difficulté. Enfin, ceux qui n'ont pas de préférence sont 6,1 % dans ce cas. Le score moyen au test est aussi lié à ces préférences déclarées y compris lorsque l'on observe les résultats toutes choses égales par ailleurs.

Une fois le test de numératie effectué, les jeunes devaient porter une appréciation sur son niveau de difficulté. Ils l'ont jugé facile, voire très facile dans leur grande majorité (71,1 %). Ici aussi, on observe un lien entre les réponses apportées et les résultats au test. Plus les jeunes ont trouvé le test difficile, plus ils ont de risques d'être en difficulté en numératie et donc d'obtenir un score faible au test, y compris lorsque les autres caractéristiques sont tenues constantes (3,4 % des jeunes ayant trouvé le test très facile sont en difficulté en numératie contre 24,6 % parmi ceux qui l'ont trouvé très difficile).

Des disparités régionales importantes

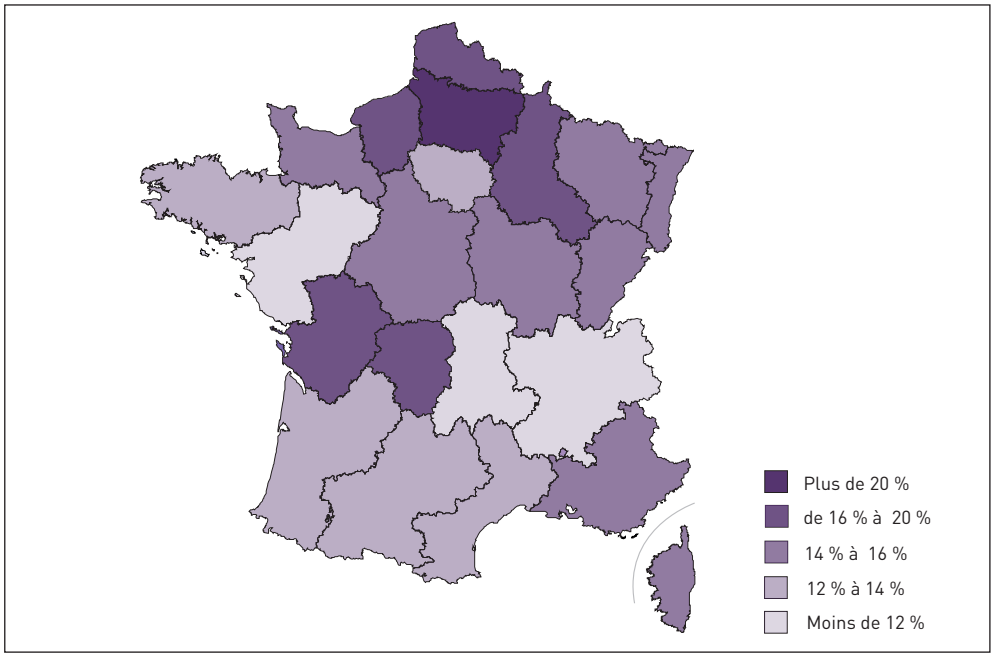
La taille de l'échantillon de jeunes ayant participé à l'évaluation permet des comparaisons territoriales. Huit régions affichent un pourcentage de jeunes en difficulté compris entre 8 % et 10 % ▸ **Figure 6**. Seules trois régions obtiennent un

Figure 6 **Pourcentage de jeunes en difficulté en numératie selon la région (JDC 2013)**



Sources : ministère de la Défense-DSN ; MENESR-DEPP.

Figure 7 **Pourcentage de jeunes en difficulté en lecture ou en numératie selon la région (JDC 2013)**



Sources : ministère de la Défense-DSN ; MENESR-DEPP.

pourcentage inférieur (Auvergne, Rhône-Alpes et Pays de la Loire). Six régions ont un taux compris entre 10 % et 12 %. Pour quatre régions, la part de jeunes en difficulté se situe entre 12 % et 14 % (Champagne-Ardenne, Limousin, Nord-Pas-de-Calais et Haute-Normandie). Pour la Picardie, ce taux atteint 19,6 %. Comme pour la lecture, les régions du nord de la France concentrent donc une partie significative des jeunes en difficulté. Si l'on prend en compte les résultats obtenus au test de lecture, la part des jeunes en difficulté dans au moins l'une des deux compétences s'élève à 27,6 % en Picardie et se situe autour de 18 % dans les quatre autres régions où la part des jeunes en difficulté en numératie est la plus élevée, ainsi qu'en Poitou-Charentes ▸ **Figure 7**. Dans trois régions seulement, cette part est inférieure à 12 %.

LA JDC, PHOTOGRAPHIE D'UNE GÉNÉRATION DE JEUNES

La JDC se révèle être un lieu unique pour observer et décrire la jeunesse française dans sa diversité. En effet, elle dispose désormais d'un outil d'évaluation informatisé novateur et adaptable. Grâce à lui, il a été possible de réaliser une prise de vue instantanée, unique et inédite d'une génération à l'entrée dans la vie adulte, simultanément dans les domaines de la lecture et de la numératie. Cette démarche vient compléter le dispositif pérenne d'observation de la maîtrise de la lecture, en apportant des résultats sans précédent permettant de décrire précisément différentes franges de la jeunesse au regard de leurs compétences en calcul.
Cette étude, réalisée auprès d'un échantillon très large et représentatif, met en lumière une proportion significative de jeunes, près de 10 %, pouvant être considérés en grande difficulté pour utiliser des mathématiques au quotidien. Parmi eux, la moitié n'est pas en mesure de mobiliser les notions parmi les plus élémentaires et peut être considérée comme étant en situation d'innumérisme. Les difficultés rencontrées par les jeunes de ce groupe renvoient directement à celles observées dans les évaluations Cedre et PISA. On peut avancer l'hypothèse que les jeunes collégiens-lycéens appartenant aux groupes de bas niveau dans ces évaluations quelques années auparavant alimentent de façon conséquente la population en situation d'innumérisme dans notre étude.

La mise en relation des compétences en lecture avec les compétences en numératie constitue aussi un apport important de cette étude. Il apparaît ainsi que presque 4 % des jeunes ont des difficultés importantes dans ces deux compétences et que près d'un jeune français sur six éprouve des difficultés dans l'un ou l'autre des domaines. Cependant, le recouvrement est loin d'être total : parmi les jeunes en difficulté de lecture, 54 % n'éprouvent pas de difficulté en numératie. En proposant des consignes simples et oralisées, cette évaluation en numératie dissocie significativement la mesure des compétences en lecture de celles mobilisées dans l'utilisation des mathématiques du quotidien. Elle permet ainsi de caractériser toutes les spécificités de la difficulté scolaire, puis quotidienne, en mathématiques.
Ces résultats amènent à questionner l'école sur les remédiations apportées aux élèves en difficulté et sur l'actuelle prévalence donnée à la maîtrise de la lecture. Certaines difficultés d'apprentissage spécifiquement mathématiques apparaissent

comme méconnues. Elles ne semblent pas bénéficier des traitements appropriés. En trouvant sa place aux côtés des autres évaluations nationales et internationales déjà existantes, cette évaluation établit un point de repère qui pourra servir lors de comparaisons ultérieures. En raison des spécificités de son échantillon, et de manière complémentaire à d'autres indicateurs, elle permet de mettre en lumière des disparités géographiques, en répondant à des besoins souvent exprimés, notamment par les collectivités territoriales. Enfin, elle confirme le rôle de prévention décisif que peut jouer la JDC dans la détection, puis l'orientation vers des centres d'information, de jeunes aux acquis scolaires fragiles.

Annexe

Corpus mathématique participant à la définition de la numératie

Conceptualisation du nombre	Nombreuses sémantiques (ordinal, cardinal, nombre concret, nombre opérateur, grandeurs) – Comptage, dénombrement, mesure.
	Tous les types de nombres (entiers, rationnels, décimaux, relatifs).
	Connaissance et coordination des registres de représentation sémiotiques. Mobilisation du registre le plus approprié pour résoudre un problème.
	Comparaison des nombres entre eux. Estimation, sens commun sur les ordres de grandeurs.
Conduite efficace des calculs	Maîtrise des quatre opérations (notions de puissances) – Sens et technique. Sur tous les nombres, dans toutes les écritures.
	Calcul mental (construction de répertoires personnels « intelligents »), calcul posé ou calcul avec des artefacts/instruments (calculatrices, tableur).
	Dialectique calcul exact/calcul approché.
	Méthode arithmétique de résolution de problème. Problèmes pouvant se ramener à des équations simples (résolues par inversement des opérations).
	Règles de priorités – Évaluer une formule.
Proportionnalité	Tableau de proportionnalité, propriétés de linéarité, produits en croix, règles de trois, proportions, représentation graphique.
	Pourcentage et pourcentages d'évolution – Calcul de vitesses, durées, temps.
Regard géométrisé	Sensibilités aux (égalités de) longueurs, aux angles, aux positions respectives des objets dans l'espace (alignement, parallélisme), etc. Connaissance des objets élémentaires du plan et de l'espace, ainsi que de leurs propriétés utilisées « en acte » dans des raisonnements déductifs.
	Compréhension et construction de représentations de figures dans le plan et dans l'espace (perspectives parallèle et centrale).
	Maîtrise des instruments de représentation papier-crayon (règle, compas, rapporteur, etc.) et numériques (logiciels de géométrie dynamique).
	Capacité à modéliser le réel par une figure géométrique.
Exploitation coordonnée de ressources	Choix des ressources appropriées – Recherche documentaire.
	Connaissance des outils et des instruments standards de représentation – Tableaux – Graphiques (diagrammes en bâtons, histogrammes, diagrammes circulaires, graphique).
Analyse et synthèse statistique	Groupement en classes pertinentes, effectifs, fréquences.
	Indicateurs de position et de dispersion – Non-confusion entre les indicateurs (moyenne/médiane) – Mérites comparés.
Maîtrise d'une large famille de grandeurs physiques	Propriétés d'additivité, de multiplication par un scalaire, de mise en rapport. Familiarité avec la grandeur séparément de toute démarche de mesure.
	Liens entre grandeurs, non-confusion entre certaines grandeurs. Grandeurs quotient.
	Mesure d'une grandeur, imprécision. Unités, conversions, unités du système international et autres. Instrument de mesure spécifique à chaque grandeur.
	Formules de calculs.
Sensibilité aux probabilités	Approche fréquentiste. Représentation à l'aide d'arbres, de tableaux.
	Familiarité avec les situations courantes (dont équiprobabilité).
Fondements algorithmiques	Algorithmes représentés par des instructions textuelles ou imagées.
	Connaissance d'algorithmes usuels et élémentaires.
	Exécution et communication à autrui de l'algorithme.

BIBLIOGRAPHIE

Artigue M., 2004, « L'enseignement du calcul aujourd'hui : problèmes, défis et perspectives », *Repères – IREM,* n° 54, p. 23-39.

Bourny G., Bessonneau P., Daussin J.-M., Keskpaik S., 2010, « L'évolution des compétences générales des élèves en fin de collège de 2003 à 2009 », *Note d'information*, n° 10.22, MENJVA-DEPP.

Brun A., Huguet T., 2010, « Les compétences en mathématiques des élèves en fin de collège », *Note d'information,* n° 10.18, MENJVA-DEPP.

De La Haye F., Gombert J.-É., Rivière J-P., Rocher T., 2010, « Les évaluations en lecture dans le cadre de la journée d'appel de préparation à la défense – Année 2009 », *Note d'information,* n° 10.11, MEN-DEPP.

Huguet T., Everaert V., 2012, *Mathématiques en fin de collège : le bilan des compétences,* Chasseneuil-du-Poitou, SCÉRÉN [CNDP-CRDP], coll. « Évaluations élèves », 64 p.

Jonas N., 2013, « Les capacités des adultes à maîtriser des informations écrites ou chiffrées – Résultats de l'enquête PIAAC 2012 », *Insee Première,* n° 1467.

Jonas N., 2012, « Pour les générations les plus récentes, les difficultés des adultes diminuent à l'écrit, mais augmentent en calcul », *Insee Première,* n° 1426.

Keskpaik S., Salles F., 2013, « Les élèves de 15 ans en France selon PISA 2012 en culture mathématique : baisse des performances et augmentation des inégalités depuis 2003 », *Note d'information,* n° 13.31, MEN-DEPP.

OCDE, 2014, *Résultats du PISA 2012 – Savoirs et savoir-faire des élèves : performance des élèves en mathématiques, en compréhension de l'*écrit *et en sciences,* vol.1, Paris, OCDE, 580 p.

Vourc'h R., Rivière J.-P., De La Haye F., Gombert J.-É., 2014, « Journée défense et citoyenneté 2013 : des difficultés en lecture pour un jeune français sur dix », *Note d'information*, n° 12, MENESR-DEPP.

ÉVALUATION DES EFFETS DU DISPOSITIF EXPÉRIMENTAL D'ENSEIGNEMENT INTÉGRÉ DE SCIENCE ET TECHNOLOGIE (EIST)

Marion Le Cam
MENESR-DEPP, bureau de l'évaluation des élèves

Olivier Cosnefroy
Université Grenoble Alpes, laboratoire des sciences de l'éducation

Remerciements à **Pascal Bessonneau** (MENESR-DEPP) pour la mise en œuvre d'une partie des analyses psychométriques.

Cet article présente les résultats de l'évaluation des effets du dispositif expérimental d'enseignement intégré de science et technologie (EIST). Ce dispositif a été évalué à partir de la rentrée 2008. Pendant quatre ans, une cohorte composée d'élèves ayant bénéficié de l'EIST en classe de sixième et d'élèves n'en ayant pas bénéficié, a été suivie jusqu'en fin de troisième et évaluée à cinq reprises. Ces évaluations permettent de mesurer l'évolution d'un score cognitif de performance en sciences des élèves, ainsi que de deux scores conatifs mesurant la motivation intrinsèque pour les sciences et l'intérêt pour les sciences en dehors de l'école. Des modélisations ont été engagées pour identifier l'existence d'un effet de l'EIST sur la progression des élèves en sciences, et sur l'évolution de leurs attitudes envers les sciences au cours du temps. Nous n'observons pas d'effet de l'EIST sur la progression des performances en sciences des élèves tout au long de leur scolarité au collège. En moyenne, l'intérêt et la motivation des élèves pour les sciences tendent à décroître au cours des années de collège. Si les élèves ayant bénéficié de l'EIST présentent un niveau de motivation légèrement plus élevé en début de sixième, cet écart moyen reste cependant stable jusqu'en fin de collège.

De nombreux rapports d'institutions internationales, parus au début des années 2000, font état d'une désaffection des étudiants pour les filières scientifiques. En 2000, le Conseil européen de Lisbonne avait émis le souhait que l'Union européenne devienne l'économie basée sur la connaissance

la plus compétitive et la plus dynamique d'ici 2010, ce qui se traduisait par la volonté d'attirer et de retenir des chercheurs de haut niveau. Or, un rapport de la Commission européenne de 2004, intitulé « L'Europe a besoin de scientifiques » [Gago, Caro *et alii*, 2004], s'alarme du manque d'engouement pour les carrières scientifiques de la part des jeunes et émet des recommandations afin que l'Union européenne se donne les moyens d'atteindre les objectifs de recrutements, dans le secteur de la recherche, nécessaires pour mettre l'Europe en tête en matière d'excellence scientifique et technologique. En 2006, un rapport d'orientation de l'OCDE fait le même constat, cherche à en étudier les causes et émet également un certain nombre de recommandations pour modifier cette tendance [OCDE, 2006]. Selon ce rapport, il convient, notamment, de réformer l'enseignement des sciences et technologies, et les programmes, qui ne favoriseraient pas l'intérêt des jeunes pour les sciences. En particulier, la pédagogie devrait *« être concentrée plutôt sur les concepts et les méthodes scientifiques que sur la seule mémorisation de l'information », ces objectifs étant « particulièrement importants dans l'enseignement secondaire »*. D'autres rapports [Osborne et Dillon, 2008 ; Rocard, Cserlely *et alii*, 2007] relient le déclin de l'intérêt des jeunes pour les études scientifiques à la façon dont les sciences sont enseignées, et préconisent de développer les pratiques pédagogiques basées sur des méthodes d'investigation, qui seraient plus efficaces pour accroître l'intérêt des élèves pour les sciences, ainsi que leur niveau de réussite dans ces disciplines.

Toutefois, si les méthodes d'enseignement privilégiant les activités expérimentales et un rôle actif des élèves permettent effectivement de susciter un engagement de court terme de la part des élèves, elles ne suffisent pas forcément pour les motiver à poursuivre des études scientifiques [Abrahams et Reiss, 2012 ; Abrahams et Millar, 2008], et leur impact sur les performances des élèves en sciences reste largement discuté [Alfieri, Brook *et alii*, 2011 ; Kirchner, Sweller, Clark, 2006]. Différentes études montrent que la mise en œuvre de la démarche d'investigation en cours de sciences ne va pas de soi, ni pour les enseignants, ni pour les élèves qui peuvent rencontrer des difficultés à faire le lien entre les expériences et les théories. Il est donc nécessaire d'apporter à ces derniers une aide appropriée pour engendrer des effets positifs [Eurydice, 2006].

En France, la définition du socle commun de connaissances et de compétences arrêtée en 2006 vise une plus grande cohérence des enseignements de sciences et de technologie. Il regroupe les principaux éléments de mathématiques et la culture scientifique et technologique dans l'un de ses sept piliers. En 2008, les programmes sont adaptés discipline par discipline, mais sont alors, dans ce cadre, soumis à des objectifs communs plus généraux, notamment de formation de l'esprit et de la personne en tant que citoyen. En ce qui concerne les sciences, une approche transdisciplinaire est clairement explicitée dans ces nouveaux programmes, et l'utilisation de la démarche d'investigation est préconisée, non seulement comme modalité pédagogique, mais également comme un objectif de formation en soi. Cependant, si le socle cherche à mettre en évidence la complémentarité des disciplines, la nature des interactions entre disciplines reste implicite, et la structure des programmes disciplinaires, en chaînes notionnelles, peut constituer un frein à la mise en place d'une approche interdisciplinaire [Delserieys-Pedregosa, Boilevin *et alii*, 2010].
Dans ce contexte international, de revalorisation de l'intérêt pour les sciences, et national, de rénovation de l'enseignement, notamment des sciences et technologies,

l'Académie des sciences et l'Académie des technologies, en collaboration avec la direction générale de l'enseignement scolaire (DGESCO) du ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche, portent depuis 2005 le projet d'une expérimentation d'un enseignement intégré de science et technologie (EIST) lors des premières années de collège, en classe de sixième et de cinquième. Dans la lignée du projet « La main à la pâte » lancé en 1996 à l'école primaire, l'EIST consiste en un enseignement intégré des matières scientifiques et technologiques, par opposition au découpage disciplinaire traditionnel de l'enseignement dans le secondaire dans lequel différents enseignants interviennent en physique-chimie, en sciences de la vie et de la Terre (SVT) et en technologie. L'EIST se veut donc particulièrement adapté à la mise en place de la démarche d'investigation, et à un rôle actif des élèves dans le cadre des cours de science, tout en s'inscrivant dans le respect des programmes.

Présenté à travers dix grands principes, l'EIST affiche comme objectif *« l'appropriation progressive ou la consolidation par les élèves de concepts scientifiques et de techniques opératoires en même temps que l'amélioration de la maîtrise du langage et des qualités d'expression écrite et orale »* (voir rubrique EIST au collège, sur le site Internet de la fondation La main à la pâte). L'EIST est supposé *« accroître l'intérêt et l'autonomie »* des élèves, ainsi que favoriser *« une bonne acquisition des connaissances »*, et *« une transition plus heureuse entre école et collège »*. Il suppose également *« un intense travail collectif et interdisciplinaire des professeurs »*.

L'EIST est expérimenté depuis la rentrée 2006, et a fait l'objet d'une évaluation, réalisée de 2008 à 2012 par la direction de l'évaluation de la prospective et de la performance (DEPP) du ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche. À l'aide d'évaluations répétées d'une cohorte d'élèves entrés en sixième en septembre 2008, et ce tout au long de leur scolarité au collège, nous avons cherché à mesurer l'impact de l'EIST sur les compétences en sciences des élèves, ainsi que sur leurs attitudes envers les sciences, par rapport à un enseignement disciplinaire traditionnel. Les premiers résultats, obtenus à partir des trois premiers temps de mesure réalisés entre 2008 et 2010, avaient fait l'objet d'un premier article [Le Cam et Rocher, 2012]. À ce stade, l'analyse des progressions des élèves ne laissait apparaître aucun effet significatif du dispositif sur les performances des élèves.

Après un rappel sur les conditions de mise en œuvre de l'EIST et sur le dispositif d'évaluation, nous présenterons les résultats obtenus à partir des cinq temps de mesure réalisés entre le début de sixième à l'automne 2008 et la fin de troisième en juin 2012. Le caractère longitudinal des données permet d'étudier conjointement les progressions en sciences des élèves et l'évolution de leurs attitudes envers les sciences, tout au long de leur scolarité au collège. Nous nous sommes attachés à étudier l'existence d'un effet de l'EIST à la fois sur les domaines cognitifs et motivationnels.

LA MISE EN ŒUVRE DE L'EIST

Les conditions de mise en œuvre de l'EIST sont largement détaillées dans le rapport réalisé sur le sujet par l'inspection générale de l'Éducation nationale (IGEN) [Perrot, Pietryk, Rojat, 2009], ainsi que dans le guide de découverte réalisé par la fondation La main à la pâte [2011]. L'expérimentation de l'EIST a commencé au niveau sixième à partir de la rentrée 2006, et au niveau cinquième à partir de la rentée 2007. À la rentrée 2008, 38 établissements volontaires expérimentaient le dispositif, au niveau sixième pour 35 d'entre eux et/ou cinquième pour 10 d'entre eux (7 établissements expérimentaient aux deux niveaux, 28 établissements uniquement au niveau sixième et 3 établissements uniquement au niveau cinquième). L'information étant passée par les recteurs, qui se sont appuyés sur les IA-IPR (inspecteurs d'académie – inspecteurs pédagogiques régionaux) des disciplines concernées pour solliciter les établissements, la communication auprès des établissements sur la mise en place du dispositif et la possibilité de se porter volontaire a été plus ou moins importante dans les différentes académies, selon l'intérêt porté par les recteurs à un tel dispositif. Les 38 établissements volontaires se situent dans 12 académies[1]. Ces établissements s'engageaient alors à expérimenter l'EIST pendant quatre ans.

Le dispositif prévoit la constitution de trois groupes d'élèves à partir de deux classes, chaque groupe étant pris en charge pour l'enseignement de sciences et technologie par un unique professeur, de l'une des trois disciplines (physique-chimie, SVT, ou technologie). En sixième, chaque groupe d'élèves reçoit 3 h 30 d'enseignement des sciences et technologie par semaine, soit une demi-heure de plus que dans l'enseignement traditionnel dans lequel il n'y a pas de physique-chimie à ce niveau. Les enseignants concernés bénéficient d'une heure supplémentaire par semaine pour assurer le travail de concertation nécessaire à la préparation des cours, à partir de thèmes transversaux qui permettent d'intégrer les différents programmes disciplinaires. Pour deux classes de sixième, l'EIST représente un surcoût de 5 h 30 hebdomadaires par rapport à un enseignement traditionnel. En cinquième, ce surcoût est de 7 h 30 hebdomadaires. D'après le rapport de l'inspection générale, pour l'année 2006-2007, l'EIST n'était mis en place en sixième que pour un trimestre, mais au cours des années suivantes, il devait être mis en place sur l'année entière. D'après le cahier des charges que l'on peut trouver dans le guide de découverte de la fondation La main à la pâte, les établissements s'engageaient à assurer la mise en place de l'EIST *« pendant au moins 30 % de l'année scolaire »*. Cette souplesse accordée aux établissements rend délicate l'appréciation du temps effectif consacré à l'EIST au cours de l'année par chacun d'entre eux. De plus, l'organisation et la répartition des heures d'enseignement des sciences et technologie sont laissées à l'initiative des établissements, et sont donc aussi très variables d'un établissement à l'autre. À titre d'illustration, des éléments sur la mise en œuvre du dispositif dans les établissements ont été rendus disponibles grâce à un questionnaire envoyé en janvier 2009 aux chefs d'établissements et aux enseignants engagés dans l'EIST [Benhaim-Grosse, 2012]. Pour l'année 2008-2009, première année de l'évaluation du dispositif, l'expérimentation était prévue en classe de sixième pour toute l'année scolaire dans 32

1. Clermont-Ferrand, Nice, Toulouse, Montpellier, Poitiers, Bordeaux, Orléans-Tours, Nancy-Metz, Lille, Strasbourg, Versailles et Créteil.

établissements, et sur un semestre dans 3 établissements. Pour 80 % des groupes d'élèves de sixième participant, les 3 h 30 hebdomadaires d'EIST étaient réparties en deux séances par semaine. Ces séances, généralement de durée plus longue que dans l'organisation traditionnelle, permettent notamment de mettre en œuvre plus facilement la démarche d'investigation et des séquences de manipulations où les élèves sont actifs.

Les réponses au questionnaire des enseignants engagés dans l'EIST montrent que ces derniers se sont impliqués dans le dispositif principalement sur proposition des chefs d'établissement et/ou des IA-IPR, et assez peu, finalement, de leur propre souhait. Ils se répartissent selon le sexe de façon comparable au national. Ils sont un peu plus jeunes en physique-chimie et en SVT (avec une ancienneté moyenne de 12 et 13 ans respectivement), et sont comparables au niveau national en technologie (avec une ancienneté moyenne de 19 ans). Enfin, toujours suivant les réponses des chefs d'établissement au questionnaire, la sélection des élèves participant au dispositif en classe de sixième semble s'être effectuée de manière strictement aléatoire dans la moitié des établissements. Par ailleurs pour 7 collèges, essentiellement de petite taille, tous les élèves de sixième ont bénéficié de l'EIST. Enfin, pour 8 collèges la sélection s'est faite sur la base d'un ou plusieurs critères, et notamment le niveau des élèves pour 5 d'entre eux. Dans ce cas, les élèves en difficultés sont sous-représentés dans les classes d'EIST.

LE DISPOSITIF D'ÉVALUATION

L'évaluation d'expérimentations sociales n'est pas nouvelle, mais a connu un développement particulier en France suite à la diffusion des travaux d'Esther Duflo en économie du développement [Banerjee et Duflo, 2009], et à la création du Fonds d'expérimentation pour la jeunesse (FEJ) en 2009. Celui-ci a largement contribué à diffuser et à promouvoir les méthodes d'évaluation d'expérimentation, et en particulier les méthodes d'évaluation randomisées (avec assignation aléatoire) qui sont particulièrement recommandées lorsqu'on souhaite mesurer, par une approche statistique, l'impact d'un dispositif expérimental [Gurgand et Valdenaire, 2012 ; L'Horty et Petit, 2011] : si l'EIST apparaît comme un dispositif susceptible de favoriser l'acquisition des connaissances des élèves en sciences et leur intérêt pour ces disciplines, il est nécessaire de mettre à l'épreuve l'existence d'un réel effet du dispositif, et de s'assurer qu'il est bien la cause des différences observées chez les élèves. En effet, le seul jugement des acteurs impliqués (enseignants, IA-IPR, etc.) ne peut suffire à conclure, l'enjeu étant d'identifier si un tel dispositif peut et doit être plus largement diffusé, ou encouragé.

Le principe de l'expérimentation avec assignation aléatoire est de comparer la situation des individus ayant bénéficié du dispositif expérimental avec la situation qu'ils auraient connue s'ils n'en avaient pas bénéficié. L'écart entre les deux situations correspond alors à l'effet du dispositif expérimental. Comme on ne peut observer directement ce que serait la situation des bénéficiaires en l'absence d'intervention, on observe la situation d'un groupe d'individus non bénéficiaires, semblables aux

bénéficiaires, le groupe témoin. La validité de la comparaison repose notamment sur l'assignation aléatoire des individus dans chacun des deux groupes : les caractéristiques initiales des individus dans les deux groupes sont ainsi équivalentes en moyenne et toutes explications alternatives d'un potentiel effet sont ainsi écartées. La mise en œuvre de l'EIST s'est faite sur la base du volontariat des établissements, il ne s'agit donc pas d'une assignation aléatoire. Dans ce cas, la différence observée entre les élèves du groupe expérimental et ceux du groupe témoin ne correspond plus exactement à l'effet de l'expérimentation, mais peut être également composée d'un biais de sélection [Givord, 2010].

L'échantillon

Afin d'identifier l'existence d'un effet de l'EIST sur les connaissances et compétences en sciences des élèves, ainsi que sur leurs attitudes envers les sciences, la DEPP a mis en place un dispositif d'évaluation des élèves, à partir de l'année scolaire 2008-2009 et sur quatre années consécutives. L'objectif est de suivre une cohorte d'élèves, composée d'une part des élèves bénéficiant de l'EIST dans les 35 établissements expérimentateurs au niveau sixième en France métropolitaine en 2008-2009, et d'autre part d'un groupe témoin d'élèves de sixième suivant un enseignement de sciences et de technologie traditionnel dans des établissements non expérimentateurs. L'échantillon témoin a été sélectionné dans les 12 académies concernées par l'expérimentation de l'EIST (hors établissements expérimentateurs). Il est stratifié en respectant les proportions d'élèves issus d'établissements publics en éducation prioritaire, d'établissements publics hors éducation prioritaire, et d'établissements privés, observées au niveau national. Les différences de structure entre les deux groupes seront contrôlées lors des analyses. Afin qu'il soit de même taille, en nombre d'élèves, que le groupe expérimentateur, 81 établissements ont été sélectionnés, dans chacun desquels une classe de sixième a ensuite été sélectionnée par tirage aléatoire simple pour participer au dispositif d'évaluation.
Dans l'optique de pouvoir contrôler un éventuel effet de sélection, lié au volontariat des établissements engagés dans le dispositif expérimental, un échantillon d'élèves recevant un enseignement traditionnel mais scolarisés dans les établissements engagés dans l'EIST a également été sélectionné pour participer au dispositif d'évaluation. En pratique, dans chaque établissement expérimentateur, une classe entière de sixième ne pratiquant pas l'EIST en 2008-2009 a été échantillonnée par tirage aléatoire simple. La cohorte d'élèves participant au dispositif d'évaluation comportait donc trois groupes d'élèves à l'origine : celui des élèves qui ont reçu un enseignement EIST en sixième, le groupe témoin et enfin un groupe d'élèves scolarisés dans les établissements expérimentateurs, mais recevant un enseignement traditionnel. Ce dernier groupe étant fortement affecté dès la première session par une non-réponse de niveau établissement, a de ce fait un effectif particulièrement faible et ne pourra plus remplir sa fonction de groupe contrôle. Il n'est donc pas utilisé dans les analyses présentées ici.

Tous les élèves de la cohorte ainsi constituée ont été évalués à l'automne 2008, en début de sixième, avant leur participation à l'expérimentation. Ils ont ensuite été suivis tout au long de leur scolarité au collège et ont été à nouveau évalués à chaque fin d'année scolaire, et ce jusqu'en troisième, ce qui représente cinq moments d'évaluation au total. Les élèves redoublants n'ont pas été suivis, de même que les

élèves ayant fait l'objet d'un passage anticipé, et ceux ayant changé d'établissement en cours de scolarité.

Les instruments d'évaluation

Chaque moment de mesure consiste en une évaluation des élèves de l'échantillon, sous la forme d'un test papier-crayon standardisé. Chaque test, d'une durée totale de deux fois une heure, est composé de trois séquences. Les séquences une et trois sont des séquences évaluant les compétences et connaissances des élèves en sciences (dimensions cognitives), tandis que la séquence deux évalue les attitudes des élèves envers les sciences (dimensions conatives). Afin de respecter le cadre des programmes officiels en vigueur pour le niveau concerné et celui du socle commun, à chaque temps de mesure, les épreuves cognitives sont différentes. Ces dernières ont été conçues par des enseignants des trois disciplines (physique-chimie, SVT et technologie), dont certains venaient d'établissements engagés dans l'expérimentation de l'EIST, et d'autres pas.

La DEPP a été contactée pendant l'été 2008 pour débuter l'évaluation dès la rentrée de septembre 2008, ce qui imposait des délais très courts pour mettre en place les deux premiers temps de mesure, en début et en fin de sixième : les deux premières évaluations sont donc composées uniquement de questions à choix multiples (QCM) [Le Cam et Rocher, 2012]. Les trois suivantes intègrent à la fois des QCM et des questions à réponses construites, qui supposent une réponse rédigée de la part des élèves.

La séquence deux est identique à chaque temps de mesure, et comporte 28 questions (aussi appelées items) réparties en six grands thèmes : « 1– motivation pour les sciences », « 2– activités en relation avec les sciences », « 3– les sciences dans l'avenir et le futur métier », « 4– sensibilisation aux phénomènes environnementaux », « 5– sentiment d'efficacité en sciences », et « 6– qu'est-ce que faire des sciences ? ».

ANALYSE DES DONNÉES

Participation et comparaison des groupes

Le **tableau 1 p. 290** présente la participation des établissements et des élèves aux cinq moments d'évaluation. Sur les 116 établissements visés au départ (35 expérimentateurs et 81 témoins), 102 ont participé à tous les moments d'évaluation (34 expérimentateurs et 68 témoins). Sur les 4 012 élèves visés au départ, 2 004 ont finalement participé aux cinq temps de mesure, soit 50 %. Ce taux de participation est un peu plus élevé dans le groupe EIST (53 %) que dans le groupe témoin (47 %).

Le groupe témoin présente un taux de non-réponse au niveau établissement plus important que le groupe EIST : à chaque session, entre 2 et 5 établissements n'ont pas participé, amenant à 68 le nombre d'établissements ayant effectivement participé aux 5 temps de mesure. Au-delà de cette non-réponse au niveau établissement, l'attrition

Tableau 1 Participation des établissements et des élèves aux cinq moments d'évaluation

		EIST			Témoins	
		Établissements EIST	Élèves EIST	Élèves (en %)	Établissements témoins	Élèves témoins
2008-2009	Nombre visé	35	2 035	100,0 %	81	1 977
Session 1 : début de 6e	Nombre testé	35	1 961		79	1 845
Session 2 : fin de 6e	Nombre testé	35	1 903		81	1 870
Participation sessions 1 et 2		35	1 853	91,1 %	79	1 755
2009-2010	Nombre visé	35	1 798		81	1 750
Session 3 : fin de 5e	Nombre testé	34	1 647		76	1 547
Participation sessions 1, 2 et 3		34	1 542	75,8 %	74	1 402
2010-2011	Nombre visé	35	1 629		81	1 584
Session 4 : fin de 4e	Nombre testé	35	1 435		78	1 361
Participation sessions 1, 2, 3 et 4		34	1 280	62,9 %	70	1 132
2011-2012	Nombre visé	35	1 460		81	1 401
Session 5 : fin de 3e	Nombre testé	35	1 289		78	1 239
Participation sessions 1, 2, 3, 4 et 5		34	1 068	52,5 %	68	936

observée à partir du troisième temps de mesure est due notamment aux élèves ayant quitté leur établissement et aux élèves redoublants ou ayant fait l'objet d'un passage anticipé, qui ne sont plus suivis et sortent de la cohorte. À cette perte s'ajoute de la non-réponse au niveau élève, la part de chaque cause dans la perte totale est difficile à évaluer, car ces informations ne sont transmises par les établissements que de façon très partielle.

La principale différence initiale entre les deux groupes concerne la proportion d'élèves issus d'établissements privés qui s'élève à 4 % dans le groupe EIST contre 23 % dans le groupe témoin **Tableau 2**. La comparaison des caractéristiques des élèves ayant participé à la première session, et des élèves ayant participé aux trois

Tableau 2 Caractéristiques des groupes de répondants

	Répondants à la première session (N=3 840)		Répondants aux cinq sessions (N=1 894)	
	EIST	Témoins	EIST	Témoins
N	1 961	1 845	1 068	936
% d'élèves du secteur privé	4,04	22,89	3,89	23,21
% d'élèves en éducation prioritaire	14,00	14,29	12,06	9,47
% de filles	48,03	48,96	50,10	53,12
% d'élèves en retard	18,96	20,55	11,19	12,93
Indice de position sociale[1] moyen du père	- 0,43	- 0,52	- 0,35	- 0,46
Indice de position sociale[1] moyen de la mère	- 0,28	- 0,39	- 0,18	- 0,31

1. L'indice de position sociale est calculé à partir des professions et catégories socioprofessionnelles [Le Donné et Rocher, 2010].

Lecture : les filles représentent 48 % des 1 961 élèves du groupe EIST ayant participé à la première session, et 50 % des 1 068 élèves du même groupe ayant participé aux cinq sessions.

Source : MENESR-DEPP.

	Ensemble		
Élèves (en %)	Établis-sements	Élèves	Élèves (en %)
100,0 %	116	4 012	100,0 %
	114	3 806	
	116	3 773	
88,8 %	114	3 608	89,9 %
	116	3 548	
	110	3 194	
70,9 %	108	2 944	73,4 %
	116	3 213	
	113	2 796	
57,3 %	104	2 412	60,1 %
	116	2 861	
	113	2 528	
47,3 %	102	2 004	50,0 %

Lecture : le suivi de cohorte visait 2 035 élèves de sixième participant au dispositif expérimental EIST à la rentrée 2008, dans 35 établissements. Parmi eux, 1 961 ont participé à l'épreuve du début de sixième, 1 903 ont participé à celle de fin de sixième, et 1 853 ont participé aux deux, ce qui représente 91 % des élèves visés au départ. Le nombre d'élèves visés dès la seconde année n'est pas celui de 2008-2009, car il ne prend pas en compte les élèves sortis de la cohorte (passage anticipé, redoublement, etc.).
Source : MENESR-DEPP.

premières sessions, avait montré que l'attrition ne présentait pas jusque-là de caractère différentiel entre le groupe EIST et le groupe témoin [Le Cam et Rocher, 2012]. Ces analyses ont été complétées en ajoutant les deux derniers moments de mesure. L'attrition concerne un peu plus les garçons que les filles, en particulier dans le groupe témoin, et surtout les élèves qui présentaient déjà un retard scolaire à l'entrée en sixième. Ces caractéristiques seront cependant contrôlées dans l'analyse de l'impact de l'EIST sur les performances et les attitudes des élèves.

Analyse des items et construction des scores

Les évaluations passées par les élèves doivent permettre de mesurer leur performance en sciences, ainsi que leurs attitudes envers les sciences, à chaque moment de mesure. Le caractère longitudinal des données permet alors d'étudier la progression des acquisitions des élèves en sciences tout au long de leur scolarité au collège, de même que l'évolution de leurs attitudes envers les sciences.

Concernant la partie cognitive des évaluations, les épreuves étaient différentes à chaque session, puisqu'elles ont été conçues dans le respect des programmes du niveau considéré chaque année. Le nombre d'items cognitifs passés par les élèves varie de 44 items lors de la session 1, à 72 items lors des sessions 4 et 5. À chaque temps de mesure, un certain nombre d'items de la session précédente sont repris (de 8 à 18 items selon les sessions). Ils sont utilisés comme ancrage afin de construire des scores comparables d'un temps de mesure à l'autre. Une analyse des caractéristiques psychométriques des items a été réalisée dans un premier temps, afin d'apprécier la qualité des items. Une première sélection des items a été réalisée à l'aide du Rbis, indice mesurant la discrimination de l'item. Les items faiblement discriminants à une ou plusieurs évaluations ont été écartés des analyses ultérieures. La fidélité de chaque épreuve a été mesurée par l'alpha de Cronbach. Ce dernier varie de 0,85 à 0,89 selon les sessions d'évaluation, indiquant

une bonne cohérence interne de l'épreuve à chaque session. La dimensionnalité des épreuves a été étudiée grâce à des analyses factorielles. Ces analyses ont laissé apparaître une large dimension dominante, correspondant à la compétence en sciences testée par chaque épreuve. Enfin, une analyse a été réalisée concernant d'éventuels fonctionnements différentiels d'items (FDI). Un FDI peut être défini comme une probabilité de réussite différente, pour deux élèves, liée à une variable autre que la compétence de l'élève. À niveau de compétence égal, la réussite différente à certains items entre groupe témoin et groupe expérimental, ou entre deux années consécutives pour les items d'ancrage, ne doit pas perturber la comparabilité des épreuves. Les items présentant un FDI ont donc été écartés des analyses. Le nombre d'items sélectionnés pour le calcul des scores est présenté dans le **tableau 3**.

Tableau 3 Nombre d'items sélectionnés pour le cacul des scores cognitifs

Session	Nombre total	Communs avec la session précédente	Mauvaise discrimination	FDI	Nombre final	Communs avec la session précédente
1	44		6	3	35	
2	62	10	15	9	38	7
3	54	8	1	12	41	5
4	72	18	5	13	54	8
5	72	17	2	4	66	13

Lecture : parmi les 72 items de l'épreuve de sciences du cinquième moment de mesure, 17 étaient repris de la session précédente. Sur ces 72 items, 2 ont été écartés en raison d'un indice de discrimination (ou corrélation item-test) trop faible, 4 ont été écartés car ils présentaient un fonctionnement différentiel (FDI). Au final, le score de la session 5 est calculé sur 66 items, dont 13 communs avec la session précédente.
Source : MENESR-DEPP.

Une fois les items sélectionnés, un score, correspondant au niveau de compétence de l'élève, a été calculé à chaque moment de mesure en utilisant des modèles de réponse à l'item à deux paramètres. Cette technique de modélisation permet, grâce aux items d'ancrage, de mettre sur une même échelle et rendre comparable les scores d'une année sur l'autre, ce qui est indispensable pour étudier les progressions des élèves en sciences tout au long de leur scolarité au collège[2]. Pour mettre sur une même échelle les scores des cinq moments de mesure, les paramètres sont estimés ensemble sur les trois premières sessions d'une part et sur les deux dernières d'autre part. Les items communs aux sessions 3 et 4, permettent de relier ces deux échelles par la méthode de Stocking et Lord [1983]. Les scores sont standardisés, de moyenne 0 et d'écart-type 1, sur les résultats des élèves du groupe témoin à la première session.

Concernant la partie conative des évaluations, le questionnaire est identique à chaque session. Les thèmes 4, sur les questions environnementales, et 6, qui concerne les représentations des élèves sur ce qu'est « faire des sciences », ont été écartés, car ces items se prêtent mal aux analyses engagées ici **Encadré**. Sur les 18 items restants, sur la base d'analyses factorielles exploratoires et confirmatoires,

2. Voir Le Cam et Rocher [2012], ainsi que l'article de Rocher, dans ce numéro, p. 37, pour plus de détails sur les modèles de réponse à l'item, et les techniques d'*equating*.

13 sont retenus, organisés en quatre facteurs. L'invariance du construit au cours du temps n'étant satisfaisante que pour les facteurs « 1– motivation intrinsèque pour les sciences », et « 2– intérêt pour les sciences en dehors de l'école », seules ces deux dimensions ont été considérées dans les analyses.

CONSTRUCTION DES SCORES CONATIFS

Le questionnaire intitulé « attitudes envers les sciences » compte 6 sections comprenant chacune entre 4 et 5 items, soit au départ 28 items. Chaque section se rapporte à un grand thème :

1 – Motivation pour les sciences
2 – Activités en relation avec les sciences
3 – Les sciences dans l'avenir et le futur métier
4 – Sensibilisation aux phénomènes environnementaux
5 – Sentiment d'efficacité en sciences
6 – Qu'est-ce que faire des sciences ?

La section 4 qui comprend 5 items portant sur les questions environnementales a été écartée. En effet, les items s'intéressent davantage aux connaissances des élèves relatives aux thématiques environnementales qu'aux aspects affectivo-motivationnels. De la même manière, la section 6, qui concerne les représentations des élèves sur ce qu'est « faire des sciences » permet difficilement de faire une analyse globale, et se prête davantage à l'analyse individuelle et descriptive de l'évolution de chaque item. Les analyses portent donc sur 4 sections, soit 18 items.

Les élèves devaient répondre selon une échelle de Likert en quatre points, allant de « tout à fait d'accord » à « pas du tout d'accord », ou de « très souvent » à « jamais » pour les items de la section 2 sur les activités en relation avec les sciences.

L'échantillon des répondants au premier temps de mesure a été subdivisé aléatoirement en deux sous-échantillons de même taille. Le premier sous-échantillon est considéré comme l'échantillon de *calibration* et le second comme l'échantillon de *réplication*.

L'échantillon de *calibration* est utilisé dans un premier temps pour réaliser une analyse factorielle confirmatoire, afin de tester la dimensionnalité des 18 items des quatre sections sélectionnées en 4 facteurs distincts. Les indicateurs d'ajustement ne sont pas complètement satisfaisants, et quatre items présentent une variance résiduelle élevée, ce qui indique qu'ils représentent mal la dimension à laquelle ils étaient associés dans le questionnaire.

Toujours en s'appuyant sur l'échantillon de *calibration*, une analyse factorielle exploratoire est alors mise en œuvre, sur les 18 items des quatre sections sélectionnées. Chaque nouvelle dimension ainsi identifiée est par ailleurs testée, en utilisant un modèle distinct (analyse confirmatoire congénérique) faisant l'hypothèse d'un seul facteur expliquant la variance commune à chacun des groupes d'items correspondants. Finalement, une structure factorielle en 4 facteurs est identifiée à partir de 13 items.

La structure factorielle trouvée est ensuite testée sur l'échantillon de *réplication*, à l'aide d'une nouvelle analyse factorielle confirmatoire. Les indicateurs d'ajustement étant satisfaisants, cette structure factorielle est celle qui est retenue.

Pour finir, nous avons interrogé la stabilité de ces construits au cours du temps, c'est-à-dire s'ils étaient bien mesurés de manière identique sur les cinq moments de mesure. Nous avons mis en œuvre une série d'analyses factorielles confirmatoires avec autocorrélation des erreurs (la réponse à une variable au temps T est corrélée avec la réponse à la même variable au temps T+1). Dans un premier temps, nous avons testé une invariance configurale, qui permet de s'assurer que le modèle de base pour chacun des construits est potentiellement spécifiable aux cinq moments de mesure (les mêmes items saturent bien sur un seul facteur). À ces cinq moments, on a autorisé l'estimation libre des moyennes et des variances d'items résiduelles. Les indices d'ajustement sont satisfaisants pour les quatre facteurs.

Le second temps a consisté à mettre en œuvre et tester un modèle d'invariance métrique (invariance faible) qui ajoute au modèle précédent la contrainte que les saturations soient égales à chaque moment de mesure. L'ajustement du modèle reste satisfaisant pour les facteurs 1 (motivation intrinsèque pour les sciences) et 2 (intérêt pour les sciences en dehors de l'école). Autrement dit, on peut avancer que pour ces deux facteurs, au cours du temps, la structure unidimensionnelle et les saturations des items correspondant à chaque construit restent identiques. Nous ne prendrons pas le risque d'avancer ce constat pour les deux autres facteurs.

Le score de motivation intrinsèque pour les sciences est calculé pour chaque élève, à chaque temps de mesure, en faisant la somme des trois items retenus pour ce facteur (« Je participe en science parce que j'aime bien chercher » ; « Je travaille en science parce que j'aime bien cette discipline » ; « Ce que je fais en science est intéressant »).

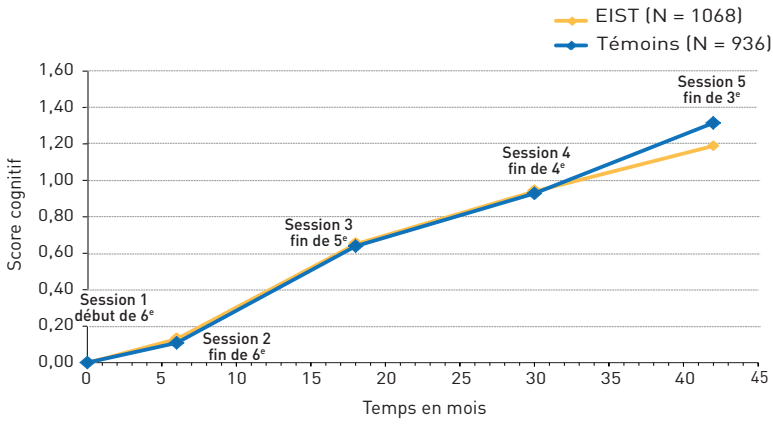
De même, le score d'intérêt pour les sciences en dehors de l'école est calculé à partir des trois items retenus pour ce facteur (« Je regarde des émissions scientifiques à la télévision » ; « J'aime lire des livres ou des revues de science » ; « Je cherche des documents scientifiques sur Internet) ».

RÉSULTATS

Progression des performances en sciences

Le niveau moyen de performance en sciences des élèves connaît une augmentation régulière au cours du temps (mesuré en mois) depuis le début de sixième (novembre 2008 : temps = 0) jusqu'en fin de troisième (mai 2012 : temps = 42). En moyenne, le niveau initial et la progression des élèves du groupe EIST sont très proches de ceux des élèves du groupe témoin, sauf en fin de troisième où le score moyen des élèves du groupe EIST est inférieur à celui des élèves du groupe témoin ▸ Figure 1. Mais ces moyennes brutes ne tiennent pas compte des différences de structures entre les deux échantillons, qui peuvent être corrélées à la performance des élèves.

▸ **Figure 1** **Évolution du score cognitif moyen selon le groupe (EIST *vs* témoin)**



Note de lecture : le score cognitif (en ordonnées) est standardisé, de moyenne 0 et d'écart-type 1 pour le groupe témoin à la première session. Le temps (en abscisses) est exprimé en mois à partir du premier temps de mesure.
Source : MENESR-DEPP.

Les données dont on dispose ici présentent une structure hiérarchisée : les élèves sont regroupés dans des établissements, dont les caractéristiques sont susceptibles d'avoir une influence sur leurs acquisitions scolaires, notamment en sciences. Les modèles multiniveaux, adaptés à ce type de données emboîtées, permettent de prendre en compte les différents niveaux dans l'analyse et de mieux estimer les erreurs standards des coefficients, et donc leur précision. Le fait d'avoir suivi les mêmes individus pendant plusieurs années, et de disposer de données longitudinales, apporte une information supplémentaire sur l'hétérogénéité intra-individuelle. Comme nous disposons de mesures répétées pour chaque élève de notre échantillon, il est possible d'étudier l'effet de l'EIST, non plus seulement sur le niveau des élèves en sciences à un moment donné, mais sur le taux de croissance, la progression, des performances en sciences des élèves durant la scolarité au collège. On le comprend, dans le cadre de cette évaluation du dispositif EIST, nous nous attendons à ce que le taux de croissance du score en sciences soit significativement plus important pour le groupe d'élèves ayant bénéficié du dispositif. Nous avons donc utilisé des modèles multiniveaux de croissance, en considérant les trois niveaux imbriqués suivants : temps de mesure > élève > établissement. Ces types de modèles, et leurs applications possibles dans le champ de l'éducation, sont présentés dans l'ouvrage de Bressoux [2010].

Du fait de l'absence de randomisation, l'étude de l'évolution du score moyen des élèves du groupe EIST et des élèves du groupe témoin a été réalisée en contrôlant les caractéristiques des élèves ainsi que celles des établissements, afin de corriger les effets de structure. Les variables introduites dans les modèles sont, en plus de la participation au dispositif EIST, le sexe, le retard scolaire à l'entrée en sixième, l'indice de position sociale de chacun des parents, calculé à partir des professions et catégories socioprofessionnelles (PCS) [Le Donné et Rocher, 2010], les deux scores conatifs, de motivation intrinsèque pour les sciences et d'intérêt pour les sciences en dehors de l'école, pour ce qui est du niveau élève, et le secteur (public/privé) ainsi que l'appartenance à l'éducation prioritaire, pour le niveau établissement.

Le **tableau 4 p. 296** présente les résultats d'une série de modèles multiniveaux dans lesquels la variable expliquée est le score cognitif. Pour les modèles 1 à 4 ces résultats ont été obtenus à partir des données de 1 894 élèves ayant participé aux cinq sessions d'évaluation et pour lesquels il a été possible de construire un score cognitif à chaque session. Le modèle 5 introduit plusieurs variables de contrôle, et a été calculé sur les 1 883 élèves pour lesquels toutes ces variables sont renseignées.

Le **modèle 1** est un modèle inconditionnel classique, c'est-à-dire sans aucune variable explicative. Il permet de décomposer la variance de la variable expliquée, le score cognitif, entre les différents niveaux et indique que 13 % de la variance se situe entre les établissements[3], 40 % entre les élèves, et 47 % au niveau intra-individuel ce qui n'est pas étonnant puisqu'on se situe sur une période assez longue, presque quatre ans, au cours de laquelle le niveau de performance de chaque élève évolue beaucoup. Le score moyen varie plus d'un élève à l'autre que d'un établissement à l'autre, il est

3. Variance inter-établissements = 0,23 / (0,23 + 0,71 + 0,85) = 0,13, variance inter-élèves = 0,71 / (0,23 + 0,71 + 0,85) = 0,40, variance intra-élève = 0,85 / (0,23 + 0,71 + 0,85) = 0,47

Tableau 4 Évaluation de l'effet de l'EIST sur les performances en science - Modèles multiniveaux de croissance

Paramètres	Modèle 1		Modèle 2	
Nombre de collèges	101		101	
Nombre d'élèves	1 894		1 894	
Nombre d'observations	9 470		9 470	
Effets fixes				
Constante	0,54	(0,05) ***	- 0,05	(0,06)
Temps			0,03	(0,001)***
EIST				
EIST*temps				
Fille				
Retard				
Indice de position sociale du père				
Indice de position sociale de la mère				
Établissement privé				
Éducation prioritaire				
Motivation intrinsèque pour les sciences				
Intérêt pour les sciences en dehors de l'école				
Motivation*temps				
Intérêt*temps				
Effets aléatoires				
Niveau 3 (inter-collèges)				
Variance des constantes	0,23	(0,04) ***	0,23	(0,04) ***
Variance des pentes				
Niveau 2 (inter-élèves)				
Variance des constantes	0,71	(0,03) ***	0,76	(0,03) ***
Variance des pentes				
Covariance constantes/pentes				
Niveau 1 (intra-élève)				
Variance des constantes	0,85	(0,01) ***	0,57	(0,01) ***
- 2LogV	28 566,74		25 557,24	
AIC	28 574,74		25 567,24	
BIC	28 585,20		25 580,32	

*** significatif au seuil de 1%, ** significatif au seuil de 5 %, * significatif au seuil de 10 %.
Lecture : le tableau présente les résultats des différents modèles multiniveaux (valeurs des paramètres, écart-type entre parenthèses et significativité indiquée par des *). Dans la partie « effets fixes » le coefficient associé au *temps* est positif et significatif : un mois supplémentaire de scolarité augmente le score en sciences de 0,02 écart-type (modèle 5).
Le coefficient de la variable indicatrice *EIST* (1 pour EIST, 0 pour témoin) indique la différence de score entre les élèves EIST et les élèves témoins au premier temps de mesure. Il n'est pas significatif. Le coefficient de l'interaction entre l'indicatrice EIST et le temps n'est pas non plus significatif : les progressions ne sont pas différentes selon que les élèves ont bénéficié ou non de l'EIST en sixième.
Source : MENESR-DEPP.

donc particulièrement important de prendre en compte les caractéristiques au niveau élève. Le **modèle 2** est un modèle inconditionnel de croissance, dans lequel la seule variable explicative est le temps (mesuré en mois à partir du premier temps de mesure). La prise en compte du temps fait diminuer la variance intra-élève de 33 % par rapport au modèle 1. La constante du modèle correspond au niveau initial moyen des élèves, au premier temps de mesure. Le coefficient de la variable temps correspond à la pente, qui représente l'évolution du niveau moyen pour une unité de temps supplémentaire (ici,

Modèle 3		Modèle 4		Modèle 5	
101		101		101	
1 894		1 894		1 883	
9 470		9 470		9 415	
- 0,05	(0,05)	- 0,06	(0,06)	- 0,07	(0,10)
0,03	(0,001) ***	0,03	(0,001) ***	0,02	(0,003) ***
		0,02	(0,10)	- 0,04	(0,08)
		- 0,001	(0,002)	- 0,001	(0,002)
				0,03	(0,04)
				- 0,83	(0,06) ***
				0,20	(0,04) ***
				0,10	(0,03) ***
				- 0,01	(0,11)
				- 0,26	(0,11) *
				0,02	(0,01) *
				0,02	(0,01) *
				0,002	(0,0003) ***
				0,0001	(0,0003)
0,17	(0,03) ***	0,17	(0,03) ***	0,10	(0,02) ***
0,0001	(0,00002) ***	0,0001	(0,00002) ***	0,0001	(0,00001) ***
0,60	(0,03) ***	0,60	(0,03) ***	0,51	(0,03) ***
0,0001	(0,00002) ***	0,0001	(0,00002) ***	0,0001	(0,00002) ***
0,004	(0,001) ***	0,004	(0,001) ***	0,002	(0,001) ***
0,53	(0,01) ***	0,53	(0,01) ***	0,52	(0,01) ***
25 271,31		25 270,80		24 666,86	
25 289,31		25 292,80		24 708,86	
25 312,84		25 321,57		24 763,78	

un mois). Dans le **modèle 3**, on autorise la pente à varier d'un établissement à l'autre, et d'un élève à l'autre. Les variances des pentes indiquent que l'évolution du score peut être significativement différente d'un élève à l'autre, et d'un établissement à l'autre. Le **modèle 4** intègre en plus l'indicatrice d'appartenance au groupe EIST, ainsi que son interaction avec le temps. Le coefficient de la variable EIST correspond à l'écart de niveau moyen entre les élèves EIST et les élèves témoins au premier temps de mesure (lorsque temps = 0), et n'est pas significativement différent de zéro. Le coefficient de l'interaction de l'EIST avec le temps (- 0,001, non significatif) nous intéresse particulièrement, puisqu'il correspond à la différence d'évolution du niveau moyen de performance en sciences des élèves EIST par rapport aux élèves témoins. Ce coefficient n'est pas significatif non plus, indiquant que le taux de croissance du score en sciences des élèves ayant bénéficié de l'EIST n'est pas différent de celui des élèves témoins. Le développement des compétences en sciences est similaire dans les deux groupes. Les indices d'ajustement montrent que ce modèle n'est d'ailleurs pas mieux ajusté que le précédent, en d'autres

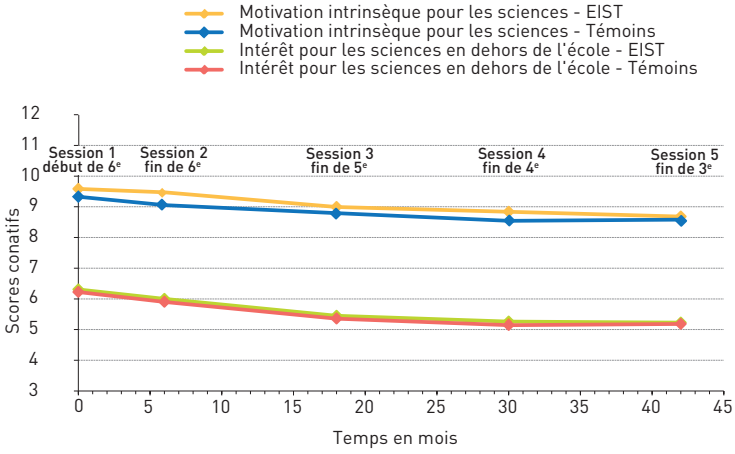
termes ces variables explicatives n'apportent aucune information supplémentaire.
Dans le **modèle 5**, on ajoute les caractéristiques aux niveaux élèves et établissements, ainsi que les deux scores conatifs, et leurs interactions avec le temps. On ne trouve pas d'effet, en moyenne, de l'EIST sur l'évolution des performances des élèves en sciences. Les caractéristiques qui expliquent les différences de développement des performances en sciences sont le retard scolaire et les positions sociales des parents, ainsi que l'évolution de la motivation intrinsèque pour les sciences (en moyenne, sur un mois de scolarité, une augmentation d'un point de score de motivation correspond à une augmentation de 0,17 % d'écart-type du score cognitif, toutes choses égales par ailleurs).

Évolution de la motivation pour les sciences, et de l'intérêt pour les sciences en dehors de l'école

Sur la base du questionnaire conatif passé à chaque temps de mesure par les élèves, nous avons identifié deux dimensions pour lesquelles la qualité de l'ajustement des données au cours du temps est satisfaisante. Nous avons donc calculé deux scores conatifs : un score mesurant la motivation intrinsèque des élèves pour les sciences, et un score mesurant l'intérêt des élèves pour les sciences en dehors de l'école. Dans la mesure où une invariance métrique a été mise en évidence au cours du temps, chacun de ces deux scores est calculé à chaque temps de mesure en sommant les réponses des élèves aux trois items retenus pour chacune des deux dimensions. Les réponses des élèves aux items étant codées de 1 à 4, ces scores conatifs peuvent varier, à chaque temps de mesure, de 3 à 12.

Les données recueillies mettent en évidence une baisse générale de l'intérêt et de la motivation pour les sciences, depuis le début de sixième et au moins jusqu'en fin de cinquième, et cela quel que soit le groupe considéré. Les élèves du groupe EIST ont en moyenne un score un peu plus élevé au départ que les élèves du groupe témoin, en particulier en termes de motivation pour les sciences. Cet écart reste stable jusqu'en fin de quatrième, et semble se réduire au cours de la troisième ▸ Figure 2.
Afin d'apprécier l'effet de l'EIST sur l'évolution des scores de motivation intrinsèque pour

▸ **Figure 2** **Évolution des scores conatifs moyens selon le groupe (EIST *vs* témoin)**



Note de lecture : les score conatifs (en ordonnées) sont mesurés sur une échelle allant de 3 à 12. Le temps (en abscisses) est exprimé en mois à partir du premier temps de mesure.
Source : MENESR-DEPP.

les sciences, et d'intérêt pour les sciences en dehors de l'école, nous avons également utilisé une modélisation multiniveaux de croissance. Seul le modèle final pour chacun de ces deux scores est reporté dans le **tableau 5**, il inclut comme variables explicatives le temps, l'indicatrice *EIST* et son interaction avec le temps, les caractéristiques au niveau établissements (secteur privé et éducation prioritaire) et au niveau élève (sexe, retard scolaire à l'entrée en sixième, indice de position sociale de chacun des parents), ainsi que le score cognitif et le deuxième score conatif à chacun des temps de mesure.

Tableau 5 Évaluation de l'effet de l'EIST sur les scores conatifs

Paramètres	Motivation intrinsèque pour les sciences			Intérêt pour les sciences en dehors de l'école		
Nombre de collèges	101			101		
Nombre d'élèves	1 883			1 883		
Nombre d'observations	9 415			9 415		
Effets fixes						
Constante	7,57	(0,11)	***	2,90	(0,15)	***
Temps	- 0,04	(0,004)	***	- 0,01	(0,01)	**
EIST	0,25	(0,1)	**	0,02	(0,08)	
EIST* temps	- 0,003	(0,004)		0,001	(0,002)	
Fille	- 0,06	(0,05)		0,36	(0,06)	***
Retard	- 0,12	(0,08)		- 0,08	(0,09)	
Indice de position sociale du père	0,06	(0,04)		0,03	(0,05)	
Indice de position sociale de la mère	0,00	(0,04)		- 0,13	(0,05)	***
Établissement privé	- 0,07	(0,12)		0,13	(0,10)	
Éducation prioritaire	0,20	(0,12)		- 0,03	(0,10)	
Score cognitif	0,10	(0,03)	***	0,03	(0,03)	
Score cognitif* temps	0,005	(0,001)	***	0,002	(0,001)	***
Intérêt pour les sciences en dehors de l'école	0,29	(0,01)	***			
Intérêt* temps	0,003	(0,003)	***			
Motivation intrinsèque pour les sciences				0,32	(0,01)	***
Motivation* temps				- 0,002	(0,001)	***
Effets aléatoires						
Niveau 3 (inter-collèges)						
Variance des constantes	0,12	(0,03)	***	0,02	(0,02)	
Variance des pentes	0,0002	(0,00004)	***	0,00003	(0,00002)	*
Niveau 2 (inter-élèves)						
Variance des constantes	0,79	(0,06)	***	1,58	(0,08)	***
Variance des pentes	0,001	(0,0001)	***	0,001	(0,0001)	***
Covariance constantes/pentes	- 0,01	(0,002)	***	- 0,02	(0,002)	***
Niveau 1 (intra-élève)						
Variance des constantes	1,92	(0,04)	***	1,51	(0,03)	***

*** significatif au seuil de 1 %, ** significatif au seuil de 5 %, * significatif au seuil de 10 %

Lecture : le tableau présente les résultats des différents modèles multiniveaux (valeurs des paramètres, écart-type entre parenthèses et significativité indiquée par des *). Concernant le score de motivation intrinsèque, dans la partie « effets fixes » le coefficient associé au *temps* est négatif et significatif : un mois supplémentaire de scolarité correspond à une baisse du score de motivation de 0,04 écart-type. Le coefficient de la variable indicatrice *EIST* (1 pour EIST, 0 pour témoin) indique la différence de score entre les élèves EIST et les élèves témoins au premier temps de mesure. Il est significatif, et positif, concernant la motivation, mais pas concernant l'intérêt pour les sciences en dehors de l'école. Le coefficient de l'interaction entre l'indicatrice EIST et le temps n'est pas significatif : les évolutions ne sont pas différentes selon que les élèves ont bénéficié ou non de l'EIST en sixième, aussi bien pour la motivation intrinsèque que pour l'intérêt pour les sciences en dehors de l'école.

Les modélisations confirment que le niveau moyen de motivation intrinsèque pour les sciences au premier temps de mesure est un peu plus élevé pour les élèves du groupe EIST que pour ceux du groupe témoin, de 0,25 point (sur une échelle de 3 à 12). Cette différence peut provenir de la façon dont les élèves ont été sélectionnés pour faire partie des classes qui expérimentent l'EIST, ou plus simplement du fait que faire partie d'un dispositif particulier donne aux élèves *« le sentiment qu'on leur accorde une sorte de privilège ou du moins un égard particulier et* [qu']*ils s'en sentent valorisés »* [Perrot, Pietryk, Rojat, 2009], ce qui peut impacter leur façon de répondre au questionnaire de motivation. Par ailleurs, les élèves du groupe EIST ne présentent pas un score initial d'intérêt pour les sciences en dehors de l'école significativement différent de celui des élèves du groupe témoin. Ce qui nous intéresse le plus ici concerne la comparaison des évolutions dans le temps des attitudes des élèves des deux groupes. Comme dans le cas du score cognitif, on constate que le coefficient de l'interaction avec le temps de la variable indicatrice indiquant la participation à l'EIST (mesuré en mois à partir du premier temps de mesure) n'est pas significativement différent de zéro pour les deux scores conatifs, ce qui signifie que l'évolution de ces deux scores tout au long de la scolarité au collège est tout à fait similaire, en moyenne, pour les élèves du groupe EIST et pour les élèves du groupe témoin. On ne peut donc pas conclure ici à un effet de l'EIST sur l'évolution de la motivation des élèves et de leur intérêt pour les sciences.

LIMITES

D'un point de vue méthodologique, les évaluations d'expérimentations se heurtent bien souvent à de nombreuses difficultés, qui doivent être identifiées car elles peuvent avoir des conséquences sur les résultats des analyses réalisées [Fougère, 2012].

Dans le cas de l'évaluation de l'EIST, la première difficulté réside dans le fait que l'affectation des établissements au dispositif était basée sur le volontariat. Il est probable que tous les établissements n'aient pas eu le même accès à l'information concernant l'existence du dispositif, voire que certains établissements aient été plus ou moins sollicités par les IA-IPR pour participer. De la même façon, il est possible que les enseignants engagés dans l'expérimentation aient des caractéristiques particulières. Les informations dont on dispose indiquent que les enseignants ont souvent été sollicités pour participer, plutôt que d'être eux-mêmes volontaires, mais ils ne l'ont probablement pas été au hasard. Enfin, les élèves participant peuvent également avoir des caractéristiques particulières. Cette sélection à différents niveaux peut avoir un impact sur les résultats des analyses, en particulier si elle s'est faite sur des critères fortement corrélés à nos variables d'intérêt, à savoir la performance en sciences des élèves, et leurs attitudes envers les sciences. Il s'avère que le niveau des élèves n'était pris en compte pour sélectionner les participants que dans un nombre restreint d'établissements, dans la majorité des cas la sélection étant aléatoire, ou concernant tous les élèves de sixième de l'établissement. Pour ce qui est des enseignants, on peut supposer qu'ils ont généralement été sollicités car ils étaient considérés comme susceptibles de faire bénéficier au mieux les

élèves de ce dispositif. Le risque est alors d'attribuer à l'EIST ce qui ne serait qu'un « effet enseignant ». À noter qu'il n'y a pas eu de questionnaire pour les chefs d'établissement et les enseignants du groupe témoin, ce qui ne permet pas d'évaluer dans quelle mesure les enseignants engagés dans l'EIST en sixième peuvent être différents de ceux des classes de sixième du groupe témoin. Dans la mesure où les élèves ont été suivis sur quatre années, au cours desquelles ils ont connu différents enseignants dans les matières scientifiques, on peut penser que l'effet, sur les résultats de nos modélisations, des caractéristiques des enseignants engagés dans l'EIST intervenant en sixième, s'il existe, reste limité. En tout état de cause, il ne semble pas suffisant pour faire apparaître un effet de l'EIST sur la progression des performances des élèves en sciences.

L'attrition est un autre problème méthodologique auquel nous sommes confrontés dans le cadre de ce dispositif d'évaluation, et qui peut également être la source d'un biais dans les résultats obtenus, surtout s'il touche de façon différentielle les groupes expérimentateur et témoin. Ici, l'échantillon final ne compte que 50 % des élèves sélectionnés au départ pour faire partie de la cohorte, et les causes de cette perte sont multiples (non-réponse au niveau de l'établissement, redoublements ou changements d'établissement des élèves, etc.).

Une autre difficulté réside dans le fait que les conditions d'application et la mise en œuvre de l'EIST peuvent largement varier d'un établissement et d'un enseignant à l'autre dans la mesure où l'EIST constitue un *« laboratoire de réflexion pédagogique »* [Perrot, Pietryk, Rojat, 2009]. Cela rend difficile l'identification de ce qui est réellement évalué.

Enfin, la mise en place du dispositif d'évaluation a dû se faire dans des délais très courts (durant l'été 2008 pour une première prise d'information à la rentrée suivante) ce qui peut se ressentir sur la qualité des instruments d'évaluation, en particulier lors des deux premiers moments de mesure. Malgré quelques items écartés, le score cognitif calculé présente de bonnes qualités psychométriques et semble pertinent pour mesurer les performances en sciences des élèves. En ce qui concerne le questionnaire conatif, la situation est moins favorable : nous avons dû nous restreindre finalement à étudier seulement deux dimensions, mesurées chacune par trois items.

CONCLUSION

L'évaluation de l'EIST ne montre pas que ce dispositif produise un effet significativement différent de celui d'un enseignement disciplinaire traditionnel sur l'évolution des performances des élèves en sciences et sur l'évolution de leur intérêt et de leur motivation pour ces dernières. Ces résultats viennent confirmer les premiers résultats, obtenus à partir des trois premiers temps de mesure [Le Cam et Rocher, 2012]. Entre le début de sixième et la fin de troisième, les performances des élèves progressent en sciences, et leur intérêt et leur motivation à l'égard des sciences tendent à diminuer de manière générale. Les élèves ayant bénéficié du dispositif

d'enseignement intégré de sciences et technologie avaient un niveau de motivation pour les sciences en début de sixième légèrement supérieur en moyenne à celui des élèves du groupe témoin, cette différence se maintient jusqu'en fin de collège.

Des études montrent que les compétences et la qualité des enseignants sont déterminantes dans l'attitude des élèves envers les sciences, peut-être même plus que les contenus des programmes [Osborne, 2003]. Ainsi, un enseignement intégré de science et technologie, pour être efficace en termes d'amélioration de la performance en sciences des élèves et de leur motivation pour les sciences doit reposer sur des enseignants qui maîtrisent leur sujet. Dans le questionnaire auquel ils ont eu à répondre, les enseignants impliqués dans l'EIST soulignent particulièrement la difficulté à enseigner des contenus différents, mais également la précipitation dans laquelle s'est faite l'entrée dans le dispositif, et le manque de temps de préparation avant le début des cours. On peut penser qu'une action sur la formation des enseignants, largement en amont de la mise en œuvre d'un tel dispositif, pourrait en augmenter les effets.

Enfin, notre étude révèle un paradoxe couramment rencontré dans l'évaluation des dispositifs d'intervention, montrant que, bien que les données recueillies ne fassent pas apparaître d'effet de l'expérimentation sur les performances et attitudes des élèves, la plupart des acteurs engagés (enseignants, IA-IPR, élèves, parents, etc.) ont un ressenti plutôt positif [Perrot, Pietryk, Rojat, 2009]. Doit-on pour autant avancer que ce dernier aspect constitue le seul point positif du dispositif ? Au regard de la grande souplesse de mise en œuvre du dispositif expérimental et du grand nombre de facteurs confondus susceptibles d'avoir eu un impact sur les dimensions mesurées, il semble difficile de conclure avec certitude sur l'existence d'un impact positif du dispositif sur les performances et attitudes des collégiens en sciences. Rien, en tout cas, dans les résultats de cette évaluation, ne le fait apparaître.

BIBLIOGRAPHIE

Abrahams I., Millar R., 2008, "Does practical work really work? A study of the effectiveness of practical work as a teaching and learning method in school science", *International Journal of Science Education*, vol. 30, No. 14, p. 1945-1969.

Abrahams I., Reiss M., 2012, "Practical work : Its effectiveness in primary and secondary schools in England", *Journal of Research in Science Teaching*, vol. 49, No. 8, p. 1 035-1 055.

Alfieri L., Brooks P., Aldrich N., Tenenbaum H., 2011, "Does discovery-based instruction enhance learning?" *Journal of Educational Psychology*, vol. 103, No. 1, p.1-18.

Banerjee A., Duflo E., 2009, "The experimental approach to development economics", *Annual Review of Economics*, vol. 1, p. 151-178.

Benhaim-Grosse J., 2012, « L'enseignement intégré de science et de technologie (EIST) en 2008-2009 : ressenti et pratiques des enseignants », *Les dossiers,* MENJVA-DEPP, n° 200.

Bressoux P., 2010, *Modélisation statistique appliquée aux sciences sociales* (2e édition), Bruxelles, De Boeck.

Delserieys-Pedregosa A., Boilevin J-M., Brandt-Pomares P., Givry D., Martin P., 2010, « Enseignement intégré de science et technologie, quels enjeux ? », *Review of Science, Mathematics and ICT Education,* vol. 4, No. 2, p. 9-28.

Eurydice, 2006, *L'enseignement des sciences dans les établissements scolaires en Europe – État des lieux des politiques et de la recherche,* Bruxelles, 112 p.

Fondation La Main à la Pâte, 2011, *Un enseignement intégré de science et technologie au collège 6e et 5e, Guide de découverte*, 128 p.

Fougère D., 2012, « Les méthodes d'expérimentation en question », *Éducation & Formations,* n° 81, MENJVA-DEPP, p. 41-47.

Gago J., Caro P., Constantinou C., Davies G., Parchmann I., Rannikmae M., Sjoberg S., Ziman J., 2004, *Europe needs more scientists*, Bruxelles, Commission Européenne, Rapport par le groupe de haut niveau sur l'accroissement des ressources humaines pour la science et la technologie, 24 p.

Givord P., 2010, *Méthodes économétriques pour l'évaluation de politiques publiques,* Document de travail, G2010-08, Insee.

Gurgand M., Valdenaire M., 2012, « Le fonds d'expérimentation pour la jeunesse et les politiques éducatives : premier retour d'expérience », *Éducation & Formations*, n° 81, MENJVA-DEPP, p. 27-37.

KIRCHNER P., SWELLER J., CLARK R., 2006, "Why Minimal Guidance During Instruction Does Not Work: An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and Inquiry-Based Teaching", *Educational Psychologist*, vol. 41, No. 2, p. 75-86.

OCDE, 2006, *Évolution de l'intérêt des jeunes pour les études scientifiques et technologiques – Rapport d'orientation*, Paris, 4 mai 2006, 21 p.

LE CAM M., ROCHER T., 2012, « Évaluation de l'effet du dispositif d'enseignement intégré de science et technologie (EIST) – Premiers résultats de l'analyse des progressions des élèves sur trois temps de mesure », *Éducation & Formations,* n° 81, MENJVA-DEPP, p. 79-90.

LE DONNÉ N., ROCHER T., 2010, « Une meilleure mesure du contexte socio-éducatif des élèves et des écoles – Construction d'un indice de position sociale à partir des professions des parents », *Éducation & Formations,* n° 79, MENJVA-DEPP, p. 103-115.

L'HORTY Y., PETIT P., 2011, « Évaluation aléatoire et expérimentations sociales », *Revue française d'économie*, vol. 26, n°1, p. 13-48

OSBORNE J., 2003, "Attitudes towards science: a review of the literature and its implications", *International Journal of Science Education*, vol. 25, No. 9, p. 1049-1079.

OSBORNE J., DILLON J., 2008, *Science education in Europe : Critical reflections*, A report to the Nuffield Foundation, London, 32 p.

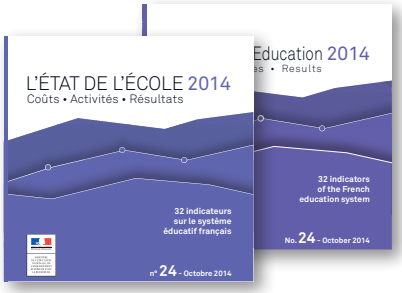
PERROT N., PIETRYK G., ROJAT D., 2009, *L'enseignement intégré de science et technologie (EIST)*, Rapport de l'inspection générale de l'éducation nationale.

ROCARD M., CSERMELY P., JORDE D., LENZEN D., WALBERG-HENRIKSSON H., HEMMO V., 2007, *L'enseignement scientifique aujourd'hui : une pédagogie renouvelée pour l'avenir de l'Europe*, Bruxelles, Commission Européenne, Rapport par le groupe de haut niveau sur l'enseignement scientifique, 28 p.

STOCKING M., LORD F., 1983, "Developing a common metric in item response theory*", Applied Psychological Measurement*, vol. 7, p. 201-210.

LES PUBLICATIONS DE LA DEPP

La direction de l'évaluation, de la prospective et de la performance (DEPP) du ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche propose diverses publications. Elles présentent les données statistiques complètes résultant d'enquêtes systématiques, mais aussi des séries d'indicateurs analytiques, des articles méthodologiques ou de synthèse, des résultats d'études ou de recherches. Elles permettent, par-delà les données succinctes contenues dans la revue *Éducation & formations*, d'aborder de façon plus approfondie le système éducatif de notre pays.



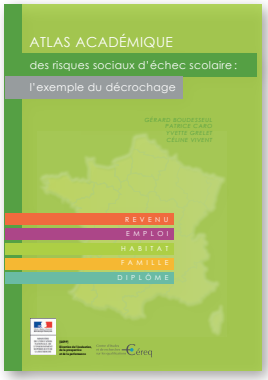
L'état de l'École expose les principales données du système éducatif mises à jour annuellement : une analyse synthétique des coûts, des activités et des résultats de l'École, qui couvre tous les niveaux du système éducatif.
Des indicateurs internationaux aident à mieux situer la France par rapport aux autres pays.
The State of Education, l'état de l'École en langue anglaise.

Repères et références statistiques présente toute l'information statistique disponible sur le système éducatif et de recherche français, déclinée en plus de 180 thématiques.
Ce vaste ensemble de données contribue à étayer le débat sur le fonctionnement et les résultats de l'École.

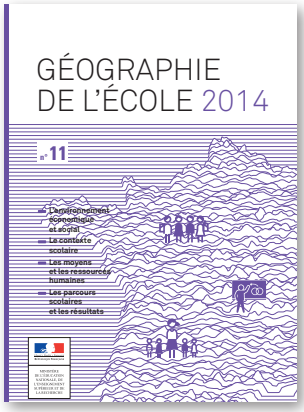
L'Éducation nationale en chiffres synthétise les caractéristiques et les tendances du système éducatif français et présente chaque année les chiffres-clés pour l'année scolaire écoulée.

Filles et garçons sur le chemin de l'égalité regroupe les principales statistiques sur les parcours scolaires comparés des filles et des garçons : résultats scolaires, choix d'orientation, poursuites d'études après le baccalauréat, insertion professionnelle.
Des indicateurs internationaux situent la France au niveau européen et au sein de l'OCDE.

Tous les contenus sont accessibles gratuitement en ligne.
La plupart proposent le téléchargement d'un format imprimable et de tableaux de données chiffrées :
www.education.gouv.fr/statistiques-catalogue-publications



Atlas académique des risques sociaux d'échec scolaire
L'ouvrage décrit finement la situation de chaque académie d'un double point de vue : il analyse et cartographie, à l'échelon cantonal, les facteurs plus ou moins propices à la réussite scolaire et la difficulté scolaire qui peut conduire à l'abandon précoce des études.



Géographie de l'École présente les principales données du système éducatif dans leurs disparités géographiques : offre de formation, moyen et ressources humaines, parcours scolaires et résultats.

Les Notes d'Information font le point sur un des aspects récurrents ou ponctuels du système éducatif et donnent sous forme synthétique l'essentiel des dernières exploitations d'enquêtes et d'études.

Les articles de la revue *Éducation & formations*, au travers d'études menées par des spécialistes, traitent des grands enjeux de l'éducation, de la formation professionnelle ou de la recherche.

Chaque année, le *Bilan social* dresse un portrait de l'ensemble des personnels, enseignants et autres personnels de l'Éducation nationale et de l'Enseignement supérieur. Il présente les indicateurs utiles au pilotage des ressources humaines contribuant au fonctionnement du système éducatif : effectifs détaillés et caractéristiques des personnels, carrières, conditions de travail.

LES STATISTIQUES DU MINISTÈRE

www.education.gouv.fr/statistiques

www.enseignementsup-recherche.gouv.fr/statistiques

Sur les sites Internet du ministère de l'Éducation nationale de l'Enseignement supérieur et de la Recherche, retrouvez l'ensemble des **données publiques** couvrant tous les aspects structurels de l'éducation et de la recherche :

- ✓ les derniers résultats d'enquêtes ;
- ✓ les publications et rapports de référence ;
- ✓ des données détaillées et actualisées ;
- ✓ des répertoires, nomenclatures et documentation.

Vous recherchez une information statistique ?
Contactez le centre de documentation (61-65, rue Dutot – 75732 Paris cedex 15)
par téléphone au : 01 55 55 73 58 (les **lundis**, **mercredis** et **jeudis** de 14 h à 16 h 30)
ou par courriel : **depp.documentation@education.gouv.fr**

Achevé d'imprimer en mai 2015
sur les presses de l'imprimerie Ovation
4, rue du Docteur Leray – 95880 Enghien-les-Bains

ÉDUCATION & FORMATIONS n°86-87

ÉVALUATION DES ACQUIS : PRINCIPES, MÉTHODOLOGIE, RÉSULTATS

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

[DEPP]
Direction de l'évaluation de la prospective et de la performance

26 €
Téléchargeable sur www.education.gouv.fr
ISBN 978-2-11-138951-9