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

Marion Le Cam, Thierry Rocher
Bureau de l'évaluation des élèves
DEPP, MENJVA

Remerciements à
Pascal Bessonneau (DEPP) et **Ahmed Ould Ahmed Jiddou** (ENSAE)
pour la mise en œuvre des calculs psychométriques
Jérôme Goidin et **Ginette Bourny** (DEPP)
pour la conception et le suivi du dispositif.

Le dispositif expérimental d'enseignement intégré des sciences et technologies (EIST) a fait l'objet d'une évaluation conduite par la DEPP, qui a suivi une cohorte d'élèves scolarisés en 6e à la rentrée 2008. L'analyse des progressions de ces élèves (trois temps de mesure jusqu'en fin de 5e), ne laisse apparaître aucun effet significatif sur la performance des élèves ayant bénéficié de l'EIST par rapport aux élèves de l'échantillon témoin à ce stade de l'expérimentation. Cette absence d'effet concerne aussi bien les performances cognitives que les attitudes à l'égard des sciences. Les effets liés au biais de sélection, à l'attrition, ou encore au choix des instruments de mesure sont discutés. En outre, cet article aborde précisément la méthodologie suivie pour mesurer les progressions des élèves dans le temps, à travers la construction d'une échelle de performance commune aux trois temps de mesure. Au-delà de l'évaluation de l'EIST, un résultat frappant est apparu : la baisse générale du niveau d'intérêt et de motivation à l'égard des sciences, au cours des premières années de collège.

Qu'est-ce que l'EIST ?

Depuis l'année scolaire 2006-2007, l'enseignement des sciences et de la technologie au collège fait l'objet d'une expérimentation portée par l'Académie des sciences qui préconise un rapprochement des enseignements de sciences de la vie et de la Terre (SVT), de sciences physiques et chimiques (SPC) et de technologie, lors des deux premières années du collège. Il s'agit de retarder le morcellement de ces enseignements et de promouvoir l'utilisation de la démarche d'investigation basée sur l'expérience et l'observation afin de favoriser le développement de la curiosité des élèves et de leur donner le goût pour les disciplines scientifiques et technologiques.

Les 38 collèges volontaires au départ pour expérimenter cette démarche, se sont engagés à assurer pendant quatre ans un enseignement intégré de science et technologie (EIST) pour un (ou plusieurs) groupe(s) d'élèves de niveau 6e et/ou 5e. Pour l'année 2008-2009, date de la mise en place de l'évaluation du dispositif, 36 collèges ont participé à l'expérimentation pour le niveau 6e, et 10 ont participé à l'expérimentation pour le niveau 5e.

Concrètement, la situation la plus fréquente est la suivante : les élèves de deux classes sont répartis en trois groupes. Chaque groupe reçoit pendant 3 heures 30 par semaine un enseignement intégrant les trois disciplines, mais assuré par un seul professeur de l'équipe EIST (constituée le plus souvent d'un professeur de SVT, un de SPC et un de technologie). Les trois enseignants préparent de façon concertée leurs progressions et leurs cours à partir de thèmes transversaux qui permettent d'intégrer les différents programmes disciplinaires. Ces thèmes sont encadrés par un « guide pédagogique », élaboré conjointement par l'Académie des sciences et les Inspections générales concernées.

Avec le soutien de la DGESCO (direction générale de l'enseignement scolaire), certaines conditions ont été réunies pour faciliter la mise en œuvre de cette expérimentation :

1. la réintégration d'un enseignement de SPC en 6e ;

2. l'augmentation du volume horaire de science et technologie en 6e (3 heures 30 hebdomadaire au lieu de 3 heures) ;

3. la mise en œuvre de classes à effectifs réduits pour cet enseignement (trois groupes constitués à partir de deux classes) ;

4. et enfin, la prise en compte financière d'une heure de concertation par semaine pour faciliter le travail en commun des enseignants impliqués, nécessaire au bon déroulement du projet.

En première analyse, il apparaît difficile d'être en mesure de distinguer les effets propres à chacun de ces aspects sur les progrès des élèves. Par conséquent, l'évaluation ne concerne pas seulement l'aspect « pédagogique », c'est-à-dire l'enseignement intégré de sciences et technologie, mais également ces quatre aspects qui constituent des « facteurs confondus ».

Le dispositif d'évaluation

L'évaluation des effets de l'EIST a débuté lors de l'année scolaire 2008-2009. La DEPP a été contactée pendant l'été 2008 pour une mise en place de l'évaluation dès l'automne 2008. Ce calendrier très serré explique certaines caractéristiques du dispositif d'évaluation, que nous détaillons ci-après.

Constitution des groupes

Trois groupes d'élèves, de 6e et de 5e, ont été constitués :

- Groupe 1 (EIST) : les élèves qui reçoivent un enseignement EIST (collèges de France métropolitaine uniquement, deux établissements situés outre-mer expérimentent l'EIST mais ne sont pas concernés par l'évaluation) ;
- Groupe 2 : un échantillon d'élèves qui reçoivent un enseignement « traditionnel » mais scolarisés dans des collèges qui pratiquent l'EIST dans d'autres classes ;
- Groupe 3 (témoin) : un échantillon d'élèves qui reçoivent un enseignement « traditionnel » dans des collèges non impliqués dans l'EIST.

Il est très important de signaler que les collèges engagés dans l'EIST sont des établissements volontaires. Certaines caractéristiques de ces collèges, et en particulier celles de leurs enseignants de sciences, sont susceptibles d'expliquer la participation à l'expérimentation. Certaines de ces caractéristiques, difficilement observables, comme le degré de motivation des enseignants, la cohésion des équipes pédagogiques, etc. sont potentiellement corrélées avec les progrès réalisés par les élèves en science. Un effet de sélection est donc à craindre. En d'autres termes, si un effet positif sur l'acquisition de compétences scientifiques est observé, cet effet sera-t-il attribuable au dispositif expérimental lui-même ou bien plus simplement à ces caractéristiques non observées des enseignants des collèges volontaires ?

Afin de pouvoir répondre à cette question, au moins de façon partielle, un deuxième groupe a été considéré, regroupant des élèves scolarisés dans les collèges engagés dans l'EIST mais dont la classe n'est pas concernée par l'EIST. La constitution de ce groupe a procédé par échantillonnage d'une classe entière, non concernée par l'EIST, au sein des collèges ayant des classes impliquées dans le dispositif. Si un effet positif est observé sur le groupe expérimental et également sur ce groupe particulier, il est fort probable qu'il soit dû à des caractéristiques des enseignants, qui sont en partie communes à ces deux groupes. Malheureusement, comme nous le verrons plus loin, les données recueillies pour ce groupe d'élèves se révèlent de faible qualité et rendent difficile une analyse approfondie en ce sens.

Le troisième groupe (groupe témoin) sert de référence. Il s'agit d'un échantillon représentatif des élèves des académies concernées, la population-cible étant celle des élèves scolarisés en 6e et en 5e dans les collèges publics et privés sous contrat en France métropolitaine (cf. annexe 1).

Suivi longitudinal

Les analyses présentées ici portent sur les élèves de 6e qui ont été suivis et interrogés à trois moments de leur scolarité, une première fois en début d'année scolaire (novembre 2008), une deuxième fois en fin d'année (mai 2009), et une troisième fois en fin de 5e (mai 2010)[1]. Ces élèves sont suivis et évalués chaque année, jusqu'à leur arrivée en fin de collège : en 4e (mai 2011) et en 3e (mai 2012). L'objectif de ces évaluations est de déterminer s'il existe une différence de niveau des acquis en termes de

NOTE

1. Les élèves de 5e à la rentrée 2008 ont également fait l'objet d'un suivi. Cependant, vu le faible nombre d'établissements proposant l'EIST en 5e et le manque d'informations précises quant à leur participation concrète à l'EIST, nous centrons ici l'analyse sur les élèves en 6e à la rentrée 2008. Des études ultérieures seront menées pour préciser, si possible, les effets obtenus sur l'échantillon des élèves de 5e à la rentrée 2008.

connaissances, de compétences et d'attitudes à l'égard des sciences, entre les élèves qui ont bénéficié du dispositif EIST et les élèves qui ont reçu un enseignement « traditionnel » (échantillon témoin).

Il faut noter que les élèves ayant redoublé ou ayant changé d'établissement dans l'intervalle ne sont pas suivis, ce qui aurait nécessité la mise en place de moyens beaucoup plus importants que ceux alloués à cette évaluation. Une description des phénomènes d'attrition « structurelle » est présentée dans la suite de l'article.

L'ÉVALUATION DES CONNAISSANCES ET DES COMPÉTENCES EN SCIENCES ET TECHNOLOGIE

Conception sous contraintes

Le cadre de l'évaluation des performances cognitives est celui des programmes disciplinaires et du socle commun de connaissances et de compétences. Le groupe de concepteurs, constitué d'enseignants, a été configuré de manière à assurer la représentation de chacune des disciplines, ainsi que la présence d'enseignants appartenant à la fois à des collèges EIST et à des collèges non-EIST.

Le tableau 1 présente la répartition des items[2] selon les compétences visées, pour les trois sessions d'évaluation analysées ici (début de 6e, fin de 6e, fin de 5e). En outre, la conception a veillé à respecter l'équilibre selon les différentes connaissances du socle commun : univers, matière, vivant, énergie, objets techniques, environnement et développement durable.

Tableau 1 – Répartition des items selon les compétences visées

Compétences	session 1 début 6e	session 2 fin 6e	session 3 fin 5e
Rechercher, extraire, organiser l'information	27	20	17
Réaliser, mesurer, appliquer des consignes	-	4	9
Pratiquer une démarche	17	11	14
Présenter la démarche et les résultats	-	2	4
Connaître la notion	10	25	11
Total	54	62	55

Lecture : l'épreuve de sciences de la première session comportait 54 items dont 27 relatifs à la compétence « rechercher, extraire, organiser l'information »
Source : MENJVA-DEPP

Comme nous l'avons évoqué plus haut, le calendrier très serré a contraint certains aspects du dispositif d'évaluation. Concernant la conception des épreuves cognitives, ces contraintes portent, d'une part sur les modalités d'expérimentation des instruments, d'autre part sur le format des épreuves.

Concernant l'expérimentation des instruments de mesure, rappelons qu'il s'agit d'une étape cruciale dans la démarche d'élaboration d'épreuves. Pour des dispositifs d'évaluation tels que CEDRE ou PISA par exemple, deux années sont nécessaires à la construction des outils d'évaluation avant leur utilisation finale : la première année, le travail consiste à construire les instruments et à les « cobayer » auprès de quelques classes ; la deuxième année, il s'agit d'effectuer l'expérimentation d'un grand nombre d'instruments auprès d'un échantillon représentatif, afin de sélectionner les plus pertinents, en se basant sur des critères statistiques spécifiques. Ce processus s'inscrit dans le rythme de l'année scolaire : les différentes versions doivent être passées à la même période. C'est pourquoi deux années de travail sont nécessaires en amont de l'évaluation finale. Dans le cadre de l'EIST, ce calendrier a dû être sévèrement comprimé, ce qui explique certains choix.

Une conséquence importante concerne le format des items : les épreuves des deux premières sessions – début de 6e et fin de 6e – sont constituées de questions à choix multiples (QCM), pour partie issues de l'évaluation CEDRE CM2 2007 en sciences expérimentales. Les QCM ont plusieurs avantages : la capacité à couvrir un domaine, à travers l'emploi d'un grand nombre de questions, garantissant ainsi la représentativité de l'échantillon des questions ; la simplicité et l'objectivité de la correction ; la brièveté de la réponse (les taux de non-réponse sont très bas en comparaison de ceux que l'on peut observer sur certaines questions ouvertes qui nécessitent une implication dans la rédaction de la réponse).

En revanche, ce format d'items ne permet pas d'évaluer certains aspects importants, comme la production d'une argumentation par exemple. Ce déficit n'est pas gênant si l'objectif est d'évaluer le niveau initial des élèves, essentiellement en tant que variable de contrôle. En termes de variables mesurant le produit de l'expérimentation, le format de QCM peut néanmoins apparaître trop

NOTE

2. Le mot « items » est ici à entendre au sens de l'information minimale recueillie – la question – et non au sens des items du socle commun de connaissances et de compétences « Livret personnel de compétences ».

restrictif. En particulier, l'EIST insiste sur la démarche d'investigation, sur l'expérience, qui sont des dimensions plus difficilement appréhendables sous ce format, et même plus généralement sous une forme d'évaluation papier-crayon.

C'est pourquoi les épreuves des années suivantes intègrent des questions ouvertes, qui ont pu faire l'objet d'une expérimentation préalable. En 2010, l'épreuve de 5e comportait déjà 10 questions ouvertes (QO) parmi les 55 proposées. Dans les évaluations de 4e et de 3e, la mise en œuvre d'une démarche scientifique a été accentuée : de nombreuses questions ouvertes ont été introduites et permettront d'approfondir l'analyse des effets de l'EIST, qui, par hypothèse, devraient se concentrer sur ces dimensions.

Une même échelle pour les trois temps de mesure

Nous avons souhaité nous inscrire dans une perspective de développement des acquisitions en sciences et donc pouvoir suivre dans le temps les progressions individuelles des élèves sur une échelle unique de performances dans cette matière. Cette exigence nécessite de pouvoir relier les résultats d'une évaluation à une autre. Il n'était cependant pas possible de proposer exactement les mêmes épreuves aux élèves pour les trois temps de mesure : les programmes scolaires évoluent et il convient d'évaluer les performances des élèves au regard des attendus existant au moment où ils sont évalués.

C'est pourquoi nous avons développé des évaluations différentes selon les sessions mais, afin d'assurer la comparabilité, nous avons introduit des items communs entre les épreuves de deux sessions successives : entre l'évaluation de début et de fin de 6e ; entre l'évaluation de fin de 6e et de fin de 5e. Cette architecture en forme d'escalier permet de construire une échelle de scores commune à l'ensemble des temps de mesure et donc de suivre les progrès des élèves sur cette échelle.

L'ajustement des métriques (ou *equating*) consiste à positionner les élèves sur la même échelle, quelle que soit l'évaluation qu'ils ont passée. Ce travail implique le recours à des modélisations spécifiques, présentées dans l'encadré « Création d'une échelle commune de performance » (p. 84).

Analyse des items

Préalablement au calcul du score global, nous avons procédé à l'analyse interne des épreuves dans l'objectif de consolider les instruments de mesure. Il s'agit d'identifier et d'éliminer les items qui présentent un dysfonctionnement. Premièrement, il peut s'agir d'une mauvaise propriété liée à la mesure. Deuxièmement, la construction du score global étant issue d'une modélisation statistique, certains items présentent des défauts d'ajustement au modèle employé. Enfin, parmi les items repris d'une session à l'autre, certains laissent apparaître un fonctionnement différent selon le temps de mesure.

Plus spécifiquement, ce « toilettage » des épreuves est opéré selon trois critères :

- Faible pouvoir discriminant : la réussite à l'item est peu corrélée au score obtenu aux autres items. Cette corrélation item-test indique dans quelle mesure l'item s'inscrit dans la dimension générale. Elle indique également la différence de performance constatée entre les individus qui réussissent l'item et ceux qui l'échouent ; pour un item très discriminant, les élèves qui réussissent cet item ont des scores aux autres items nettement supérieurs à ceux des élèves qui échouent l'item.
- Mauvais ajustement au modèle de réponse à l'item : un modèle de réponse à l'item à deux paramètres est appliqué pour chaque temps de mesure (*cf. encadré*). Les items s'ajustant mal à cette modélisation, c'est-à-dire les items dont le modèle prédit mal le comportement, ont été éliminés.
- Fonctionnement différentiel de l'item (FDI) : dans le but de produire une échelle commune aux trois moments de mesure, il est nécessaire que les items communs – repris d'une session à l'autre – adoptent un comportement similaire. En particulier, l'hypothèse est faite que la hiérarchie de difficulté de ces items est conservée d'une session à l'autre. Les items présentant un fonctionnement différentiel selon le moment de mesure sont éliminés.

Le tableau 2 synthétise le résultat de ce toilettage et dénombre les items identifiés pour chacune des caractéristiques présentées ci-dessus. Il apparaît que les défauts d'ajustement au modèle et les fonctionnements différentiels ne représentent qu'une petite partie des causes d'élimination des items. Le principal facteur d'élimination des items pour dysfonctionnement est celui de la mauvaise discrimination[3], conséquence directe du resserrement des phases d'expérimentation.

Ces chiffres témoignent du caractère éminemment empirique du processus de mesure et invitent à prendre

Tableau 2 – « Toilettage » des épreuves

		N total	Mauvaise discrimination	Mauvais ajustement	FDI	N final	N communs
Session 1	Début 6e	54	12	4		36	
					2		6
Session 2	Fin 6e	62	26	3		31	
					0		8
Session 3	Fin 5e	55	4	4		47	

Lecture : parmi les 54 items de l'épreuve de sciences de la première session, 12 ont été éliminés en raison d'un indice de discrimination (ou corrélation item-test) trop faible, 4 ont été éliminés car ils s'ajustaient mal au modèle de mesure (modèle de réponse à l'item) et 2 ont été éliminés car ils présentaient un fonctionnement différentiel (FDI) entre la première et la deuxième session. Au final, l'épreuve de début de 6e comporte 36 items, dont 6 sont communs avec l'épreuve de fin de 6e.

Source : MENJVA-DEPP

conscience de la temporalité nécessaire à l'aboutissement de ce processus. Toute mesure des compétences est un construit, dont il convient de vérifier empiriquement le fonctionnement au préalable.

Une analyse complémentaire concerne les FDI en fonction du fait de suivre ou non le dispositif EIST. Autrement dit, à score égal sur l'échelle commune, les élèves qui suivent le dispositif EIST ont-ils des résultats différents des autres sur certains items ? Aucun fonctionnement différentiel de ce type n'a été observé pour les trois premiers temps de mesure. Mais si, réellement, l'EIST produit un effet sur la démarche d'investigation par exemple, il sera intéressant de reconduire ce type d'analyses sur les sessions ultérieures.

L'ÉVALUATION DES ATTITUDES À L'ÉGARD DES SCIENCES

Lors de chaque session d'évaluation, les élèves ont également passé un questionnaire visant à mesurer leurs sentiments, leurs motivations et leurs attitudes vis-à-vis des sciences (variables conatives). Ce questionnaire est repris intégralement et à l'identique pour chacun des temps de mesure : contrairement aux épreuves cognitives, le questionnement de ces dimensions ne nécessite pas d'ajustement spécifique selon le niveau scolaire interrogé.

Au total, ce questionnaire comporte 26 questions de type Likert (« tout à fait d'accord », « d'accord », etc.) qui se répartissent selon six facteurs : sentiment d'efficacité en science ; les sciences dans l'avenir et dans le futur métier ; sensibilisation aux phénomènes environnementaux ; attrait pour les expériences ; engagement pour les sciences ; activités en relation avec les sciences en dehors de l'école (*cf. annexe 2*).

La construction des scores a suivi deux étapes : 1) validation des dimensions et 2) calcul des indices.

Dans un premier temps, la structure dimensionnelle est explorée et validée à l'aide d'une analyse en facteurs communs permettant d'identifier les facteurs sous-jacents mesurés par ces items. Afin de simplifier la structure des données, la méthode de rotation oblique des axes factoriels est utilisée ; autrement dit, les facteurs ne sont pas considérés comme indépendants. Finalement, un modèle satisfaisant en six facteurs est obtenu.

Dans un second temps, pour chacun des facteurs préalablement identifiés, une analyse en composantes principales (ACP) est réalisée sur les items qui le composent. L'ACP permet de synthétiser de manière optimale l'ensemble de ces items en une seule variable continue : il s'agit d'une moyenne pondérée des questions concernées. Les poids de chaque question sont calculés sur les données de la première session, puis sont repris à l'identique pour le calcul des scores des sessions suivantes.

NOTE

3. Il est frappant en effet de constater que la proportion d'items peu discriminants est très différente selon les sessions : environ 20 % pour la première session (début de 6e), plus de 40 % pour la deuxième session (fin de 6e) et seulement 7 % pour la troisième session (fin de 5e). Ces différences s'expliquent par le fait que, selon les cas, les items ont fait l'objet ou non d'une expérimentation préalable. Ainsi, pour la deuxième session, aucun des items n'a pu être testé car ils ont été construits au cours de l'année scolaire 2008/2009 pour être directement passés en fin d'année scolaire. Pour les items de fin de 5e, les items ont également été construits lors de l'année scolaire 2008/2009 et ils ont pu être testés sur un autre échantillon d'élèves de 5e en mai 2009, afin de sélectionner une épreuve de qualité à proposer aux élèves du dispositif expérimental en mai 2010, arrivés en fin de 5e. S'agissant des items de l'évaluation de début de 6e de novembre 2008, une partie d'entre eux provenait de l'évaluation CEDRE sciences expérimentales ayant eu lieu en 2006 auprès d'un échantillon d'élèves de fin de CM2. Ceci explique la situation intermédiaire de cette épreuve en termes de nombre d'items éliminés.

Création d'une échelle commune de performance

Modèle de réponse à l'item

La construction d'une échelle de scores commune aux trois épreuves cognitives s'inscrit dans le cadre des modèles de réponse à l'item (MRI). Plus précisément, nous utilisons un MRI à deux paramètres (dit aussi « 2PL, two-parameter logistic model »).
Dans ce modèle, les élèves sont uniquement définis par leur niveau de compétence, qui correspond à leur position sur la dimension latente θ. Plus leur niveau de compétence est élevé, plus la probabilité de réussite est importante.
Les items, quant à eux, sont caractérisés par deux paramètres. D'une part, le niveau de difficulté traduit le degré de succès de l'item : lorsque le niveau de difficulté d'un item est élevé, la probabilité de le réussir est réduite. D'autre part, le paramètre de discrimination, traduit la sensibilité de l'item à une variation du niveau de compétence. Ce paramètre, aussi appelé « pente » de l'item, est positif : lorsque la discrimination est élevée, une variation du niveau de compétence entraîne une variation importante de la probabilité de répondre correctement.

De manière formelle, ce modèle exprime la probabilité qu'un élève *i* réussisse l'item *j*, sous la forme suivante :

$$P(Y_i^j = 1 / \theta_i, a_j, b_j) = \frac{\exp(D\, a_j(\theta_i - b_j))}{1 + \exp(D\, a_j(\theta_i - b_j))}$$

où Y_i^j est la réponse de l'élève *i* à l'item *j* (de valeur 1 si l'item est réussi, 0 sinon), θ_i est le niveau de compétence de l'individu *i*, D est une constante qui vaut 1,7, b_j est le paramètre de difficulté de l'item *j* et a_j est le paramètre de discrimination de l'item *j*.
Ces modèles sont très largement employés dans le domaine de l'évaluation des compétences, tant au niveau international (PISA, PIRLS) qu'au niveau national (CEDRE). Pour plus de précisions sur ces modèles, le lecteur pourra consulter, en français, Dickes *et al.* (1994).

Ajustement des métriques (*equating*)

Pour chaque session, les niveaux de compétence des élèves ont été estimés selon un MRI à deux paramètres. Cependant, les paramètres a_j, b_j et θ_j ne sont pas uniques : en appliquant une transformation linéaire donnée à ces paramètres, on peut obtenir une autre solution acceptable. Il faut donc imposer des contraintes identifiantes, par exemple, une moyenne nulle et un écart-type de 1 des θ_j, pour chaque session. Dans ce cas, la comparaison des niveaux de compétences est relative à chaque session : chacun des niveaux est ramené au niveau moyen de la session correspondante. Autrement dit, les trois jeux d'estimation des θ_j ne sont pas positionnés sur la même échelle. En s'appuyant sur les items communs, repris d'une épreuve à l'autre, il est possible de relier les niveaux de compétence selon une métrique commune aux trois sessions. De nombreuses méthodes ont été proposées pour réaliser cet ajustement des métriques (ou « equating ») à partir d'items communs. Nous avons retenu ici celle de Stocking et Lord (1983).
Les paramètres du modèle sont tout d'abord estimés indépendamment pour les trois sessions. Puis, nous avons cherché à déterminer la relation linéaire qui permette de transformer les niveaux de compétences de la deuxième session sur l'échelle de compétence de la première.
Remarquons tout d'abord que si on effectue le changement d'échelle $\theta^* = A\theta + B$, alors les nouveaux paramètres d'items $b^* = Ab + B$ et $a^* = a/A$ ne modifient pas la probabilité de réussite.
Il s'agit donc de proposer une procédure pour estimer A et B de manière à ce que les paramètres des items communs soient très proches, selon qu'on les estime lors de la première session ou lors de la deuxième. C'est le principe sous-jacent à l'equating dans le cadre des MRI : les paramètres d'items donnés sont indépendants des groupes d'élèves à partir desquels ils sont estimés. Autrement dit, ce sont les niveaux de compétence des élèves qui peuvent être différents d'un groupe à l'autre mais pas les paramètres des items.
Plus précisément, soient (a_{j1}, b_{j1}) les paramètres de l'item commun *j* estimés lors de la première session, (a_{j2}, b_{j2}) les paramètres de l'item commun *j* estimés lors de la deuxième session et (a^*_{j2}, b^*_{j2}) les valeurs des paramètres estimés lors de la deuxième session et transformés sur l'échelle de la première. On cherche les « meilleurs » A et B tels que $b^*_{i2} = \mathrm{A}\, b_{i2} + \mathrm{B}$ et $a^*_{j2} = a_{j2} / \mathrm{A}$, et tels que les b^*_{j2} (respectivement les a^*_{j2}) soient proches des b_{j1} (respectivement les a_{j1}). Pour cela, la procédure consiste à minimiser l'écart entre les deux scores « vrais », d'une estimation à l'autre (i.e. d'un jeu de paramètres à l'autre). Stocking et Lord (1983) proposent ainsi de minimiser F :

$$F = \frac{1}{n} \sum_{i=1}^{n} (\xi_i - \xi_i^*)^2$$

où ξ_i et ξ_i^* sont les scores « vrais » des *n* élèves, obtenus sur les *J* items communs : $\xi_i = \sum_{j=1}^{J} P_{j1}(\theta_i)$ et $\xi_i^* = \sum_{j=1}^{J} P^*_{j2}(\theta_i)$ avec P_{j1} et P^*_{j2} les probabilités de réussite estimées respectivement avec les paramètres (a_{j1}, b_{j1}) et (a^*_{j2}, b^*_{j2}).

Cela revient à dire qu'à niveau de compétence θ donné, les scores « vrais » obtenus aux items communs aux deux sessions devraient être sinon égaux, du moins très proches, quelle que soit la session considérée. Sur ce principe, A et B sont estimés, puis appliqués à la transformation des θ de la deuxième session sur l'échelle de compétence de la première.

La même démarche est employée pour relier les métriques des deuxième et troisième sessions. Au final, par transitivité, les niveaux de compétence sont tous mis sur la même échelle, quelle que soit la session.

Résultats

Participation

Le tableau 3 présente les chiffres de la participation des élèves aux évaluations, relativement à l'ensemble des élèves visés au départ. Globalement, 3 408 élèves ont participé aux trois sessions d'évaluation et 4 241 ont participé aux deux premières sessions, sur les 4 699 élèves attendus au départ. Les valeurs manquantes lors des deux premières sessions (458 élèves) s'expliquent, soit parce que les élèves ont été absents à l'une ou l'autre de ces sessions, soit parce que certains établissements n'ont pas fait passer une des sessions (deux établissements-témoins n'ont pas fait passer la première session).

La perte observée lors de la troisième session (soit 833 élèves) est due principalement (71 % des cas) aux élèves qui ont quitté l'établissement pour différentes raisons (déménagement, mesure de discipline, etc.), mais également aux élèves qui ont redoublé la 6e (28 % des cas) ou qui ont sauté la classe de 5e (1 % des cas) et qui ne sont donc pas concernés par l'évaluation de fin de 5e. À cette perte s'ajoutent des non-réponses : un établissement expérimentant l'EIST, et 5 établissements-témoins n'ont pas fait passer la session 3.

Le taux de participation aux trois sessions est finalement de 75,7 % des élèves du groupe 1 qui expérimentent l'EIST, de 70,9 % des élèves du groupe 3 – le groupe témoin – et seulement de 67,7 % des élèves du groupe 2.

Le groupe 2, constitué d'élèves qui reçoivent un enseignement traditionnel mais dans les collèges qui pratiquent l'EIST dans d'autres classes, a donc un effectif particulièrement faible (465 élèves ayant participé aux trois sessions). Ce groupe avait déjà un effectif plus faible au départ. Il devait concerner une classe de chaque établissement expérimentant l'EIST, mais une dizaine d'entre eux n'ont pas du tout évalué de classe suivant un enseignement traditionnel. C'est pourquoi ces élèves ont été écartés par la suite des analyses présentées.

Attrition

L'attrition concerne-t-elle les mêmes élèves selon les groupes ?

Le tableau 4 présente quelques caractéristiques des élèves des groupes EIST et témoin selon deux cas : l'ensemble des élèves évalués au départ (N = 3 804) et ceux analysés ici, c'est-à-dire pour lesquels nous disposons de l'ensemble des résultats aux trois temps de mesure (N = 2 942).

L'observation des scores bruts obtenus aux épreuves de la session 1 montre que les élèves du groupe 1 (EIST) avaient au départ un niveau légèrement supérieur à celui des élèves du groupe 3 (témoin), avec un score moyen de 24,75 contre 24,27. Les deux échantillons présentent également des différences de structures, l'échantillon-témoin ayant été tiré de façon à être représentatif des 12 académies concernées par l'expérimentation, tandis que les établissements qui expérimentent l'EIST ont été sélectionnés sur la base du volontariat. Ainsi, le taux d'élèves ayant un retard scolaire à l'entrée en 6e était un peu plus élevé pour le groupe-témoin avec 20,4 % d'élèves contre 19 % pour le groupe EIST. La répartition des établissements selon le secteur est également différente, le secteur privé étant sous-représenté dans l'échantillon EIST avec seulement 4 %

Tableau 3 – Participation des élèves

		EIST (groupe 1)			Groupe 2		Témoin (groupe 3)			Total		
		Étab. EIST	Élèves		Élèves		Étab. Non EIST	Élèves		Étab.	Élèves	
2008-2009	**nb visé**	35	2 035	100 %	687	100 %	81	1 977	100 %	116	4 699	100 %
Session 1	**nb testé**	35	1 960		635		79	1 845		114	4 440	
Session 2	**nb testé**	35	1 925		612		81	1 870		116	4 407	
Participation sessions 1 et 2		35	1 852	91,0 %	593	86,3 %	79	1 796	90,8 %	114	4 241	90,3 %
2009-2010	**nb visé**	35	1 798		605		81	1 750		116	4 153	
Session 3	**nb testé**	34	1 647		515		76	1 547		110	3 709	
Participation sessions 1, 2 et 3		34	1 541	75,7 %	465	67,7 %	74	1 402	70,9 %	108	3 408	72,5 %
2010-2011	**nb visé**	35	1 629		560		81	1 584		116	3 773	
Session 4	**nb testé**	35	1 435		486		78	1 361		113	3 282	

Source : MENJVA-DEPP

Lecture : le suivi de cohorte a concerné 2 035 élèves de 6e suivant le dispositif expérimental EIST à la rentrée 2008. Parmi eux, 1 960 ont participé à l'épreuve de début 6e, 1 925 ont participé à celle de fin de 6e et 1 852 ont participé aux deux, ce qui représente 91 % des élèves visés au départ. En fin de 5e (session 3), on comptabilise 1 541 élèves qui ont participé aux épreuves des trois sessions, soit 75,7 % des élèves visés au départ (pour rappel, les élèves ayant redoublé la 6e n'ont pas été suivis).

Tableau 4 – Caractéristiques des groupes de répondants

3 804 élèves ayant passé la session 1					2 942 élèves ayant passé les trois sessions			
Scores bruts (session 1)					Scores bruts (session 1)			
	N	Moyenne	Écart-type		N	Moyenne	Écart-type	
EIST	1 959	24,75	17,40		1 540	25,26	15,30	
Témoin	1 845	24,27	13,30		1 402	24,79	12,90	
% d'élèves en retard					% d'élèves en retard			
	N				N			
EIST	1 959	19,0 %			1 540	16,7 %		
Témoin	1 845	20,4 %			1 402	18,6 %		
% selon le secteur					% selon le secteur			
	N	Public hors EP	Éducation prioritaire	Privé	N	Public hors EP	Éducation prioritaire	Privé
EIST	1 959	82,0 %	14,0 %	4,0 %	1 540	82,9 %	13,0 %	4,0 %
Témoin	1 845	65,4 %	15,7 %	18,9 %	1 402	67,9 %	14,2 %	17,9 %

Note : Les 1 540 élèves du dispositif EIST ayant participé aux trois sessions d'évaluation ont obtenu un score moyen de 25,26 à l'épreuve de science de la première session, alors que l'ensemble des élèves du dispositif EIST (1 959 élèves) évalués au départ lors de la première session ont obtenu un score moyen de 24,75. Cela signifie que les élèves « perdus » depuis la première session sont moins performants que les autres. Pour rappel, les élèves ayant redoublé la 6e n'ont pas été suivis, ce qui explique cet écart de 0,51 points entre l'ensemble de départ et ceux observés aux trois sessions. L'intérêt du tableau est de montrer que cet écart est comparable parmi les élèves de l'échantillon témoin (0,52 = 24,79 - 24,27). Par conséquent, l'attrition a bien concerné des élèves plus faibles que la moyenne, mais ce phénomène est identique pour l'échantillon EIST et pour l'échantillon témoin.

Source : MENJVA-DEPP

des élèves de ce groupe contre 18,9 % des élèves du groupe 3.

En se restreignant maintenant aux 2 942 élèves ayant passé les trois sessions, les scores moyens de la session 1 sont légèrement supérieurs, ce qui s'explique notamment par le non-suivi des élèves redoublants. Cette augmentation s'observe de manière très comparable pour les deux groupes : EIST et témoin (de 27,75 à 25,26 pour les élèves du groupe EIST, et de 27,47 à 24,79 pour les élèves du groupe témoin). De la même façon, les différences de répartition entre les deux groupes, selon le retard scolaire ou selon le secteur de scolarisation, sont conservées lorsqu'on se restreint aux élèves ayant passé les trois sessions d'évaluation. Il n'y a donc pas lieu de conclure à une attrition différentielle selon les groupes, EIST ou témoin.

Résultats descriptifs

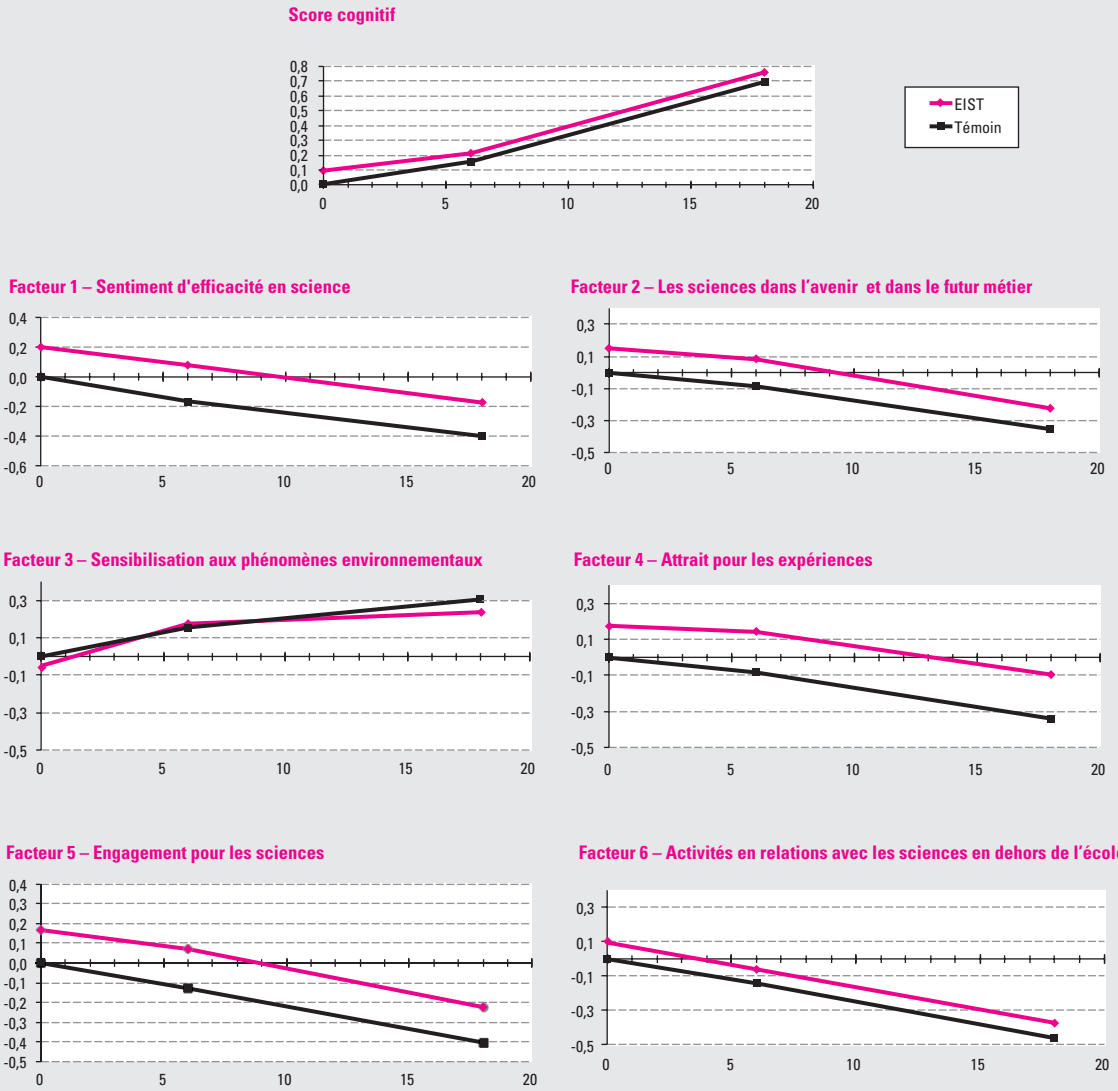
La figure 1 représente l'évolution des scores moyens obtenus par les deux groupes – EIST et témoin – en fonction des trois temps de mesure. Sept dimensions sont étudiées : le score cognitif en sciences et les six facteurs conatifs. Les scores ont été standardisés de manière à ce que, pour les élèves du groupe témoin lors de la première session, la moyenne soit de 0 et l'écart-type de 1.

Il ressort que les scores moyens évoluent en fonction du temps de manière remarquablement proches pour les deux groupes (EIST et témoin). Les élèves ayant suivi le dispositif EIST affichent globalement une supériorité, à la fois sur le score cognitif et sur les facteurs conatifs. Mais l'écart qui sépare les deux groupes reste stable au cours du temps. Cette première analyse descriptive tend à montrer l'absence d'effet lié au dispositif EIST, que ce soit en termes de performances cognitives que d'attitudes à l'égard des sciences.

S'agissant du score cognitif, il apparaît une augmentation presque constante du niveau moyen au cours du temps sur l'échelle commune de sciences. L'année scolaire de 5e correspond à une augmentation d'environ 0,55 écart-type de score, quel que soit le groupe considéré. Cette progression, associée à une année d'étude supplémentaire, est supérieure à ce que l'on peut observer dans d'autres enquêtes : dans PISA par exemple, cette progression est estimée à environ 0,4 écart-type du score. Il faut rappeler que les écarts calculés ici ne portent pas sur les élèves ayant redoublé la 6e, ce qui peut expliquer cette surestimation.

Le phénomène le plus frappant mis en exergue par ces premiers résultats est sans doute l'évolution des scores obtenus sur les variables de motivation et d'intérêt. À l'exception notable de la dimension « sensibilisation aux phénomènes environnementaux », toutes les autres affichent une tendance à la baisse du score moyen, et ce, quel que soit le groupe considéré (EIST et témoin). Une illustration de ce phénomène est fournie par le tableau 5 qui présente les réponses données à la question « Quand je serai adulte, j'aimerais continuer à faire de la science » du facteur « Les sciences dans l'avenir et dans le futur métier ». En début de 6e, 77 % des élèves du dispositif EIST répondaient « d'accord » ou « tout à fait d'accord » à cette affirmation. En fin de 5e, ces mêmes élèves n'étaient plus que 54 % à

Figure 1 – Évolution des résultats moyens selon le groupe (EIST *versus* témoin)



Note : tous les scores (cognitifs et conatifs – en ordonnées) sont standardisés, de moyenne 0 et d'écart-type 1, sur les résultats des élèves du groupe-témoin obtenus lors de la première session. Les temps de mesure (en abscisses) sont exprimés en mois : 0 mois (novembre 2008), 6 mois (mai 2008) et 18 mois (mai 2009).

donner ces réponses positives. Parmi les élèves du groupe-témoin, ce taux passe de 61 % à 47 %.

Modélisations

Afin de préciser cette première analyse descriptive et de tester statistiquement l'absence ou la présence d'effet de l'EIST, des modélisations sont engagées. Notre objectif est d'étudier les progressions des élèves,

Tableau 5 : *« Quand je serai adulte, j'aimerais continuer à faire de la science »*

% de réponse	Non-réponse	Pas du tout d'accord	Pas d'accord	D'accord	Tout à fait d'accord
EIST					
Session 1 (début 6e)	0,8	10,6	21,5	40,8	26,2
Session 2 (fin 6e)	2,1	10,1	19,0	38,7	30,1
Session 3 (fin 5e)	2,0	17,7	26,6	36,4	17,3
Témoin					
Session 1 (début 6e)	1,1	13,0	24,9	42,3	18,7
Session 2 (fin 6e)	2,9	12,8	23,0	41,5	19,8
Session 3 (fin 5e)	2,1	19,3	31,1	34,4	13,1

Lecture : parmi les élèves du dispositif EIST et ayant participé aux trois sessions d'évaluation, 26,2 % déclaraient en début de 6e être « tout à fait d'accord » avec l'affirmation « quand je serai adulte, j'aimerais continuer à faire de la science » ; ils ne sont plus que 17,3 % en fin de 5e à être dans ce cas.

c'est-à-dire la relation entre le score – cognitif ou conatif – et le temps de mesure. Cette relation est variable d'un élève à l'autre, d'une classe à l'autre et éventuellement d'un groupe à un autre, EIST et témoin.

Nous avons eu recours à un modèle multiniveau de croissance *(multilevel growth curve model)*. Ces modèles prennent en compte la structure hiérarchique des données, en l'occurrence organisée sur trois niveaux : temps de mesure/élève/collège. Ils sont très adaptés aux données recueillies ici. Pour une présentation en français de ces modèles et de leurs applications possibles dans le champ de l'éducation, le lecteur est invité à lire Bressoux (2010).

Le tableau 6 présente les résultats de la modélisation appliquée au score cognitif ainsi qu'aux six facteurs conatifs. Concernant le premier modèle, qui porte sur le score cognitif, la partie « Effets fixes » indique une relation positive et significative entre le score et le temps de mesure, ce qui traduit le développement au cours du temps des performances des élèves sur l'échelle commune de sciences. La variable indicatrice EIST (1 pour EIST, 0 pour témoin) indique une absence de différence entre EIST et témoin, lors de la première mesure. Graphiquement, il apparaît que les élèves bénéficiant de l'EIST affichent des scores supérieurs, quel que soit le temps de mesure *(figure 1)*. En fait, cette supériorité semble être due à un effet de structure, puisqu'elle devient non significative lorsqu'est introduite dans le modèle la répartition selon le type de scolarisation (privé, éducation prioritaire).

La question d'importance ici est celle de l'interaction entre le fait d'avoir bénéficié de l'EIST et le temps. Il ressort que cette interaction n'est pas significative, indiquant que la relation entre le score et le temps n'est pas différente selon que les élèves ont ou non bénéficié de l'EIST. Plus généralement, cette interaction n'est pas significative, quelle que soit la dimension interrogée.

La partie « Effet aléatoire » rapporte quant à elle les variances des différents paramètres aléatoires du modèle. La variance du niveau 1 indique la variabilité intra-élèves, celle du niveau 2 la variabilité inter-élèves et celle du niveau 3, la variabilité inter-collèges. Pour le score cognitif par exemple, des variances significatives sont observées à ces trois niveaux, et concernent à la fois les écarts de performance entre élèves ou entre collèges (les constantes) et les écarts de progressions (les pentes).

Tableau 6 – Évaluation de l'effet (modèles de croissance)

	score cognitif		facteur 1		facteur 2		facteur 3		facteur 4		facteur 5		facteur 6	
Effets fixes														
Constante	0,002	(0,061)	0,035	(0,046)	0,098	(0,039)**	0,069	(0,042)	0,023	(0,036)	-0,058	(0,039)	0,079	(0,037)**
Temps	0,039	(0,003)***	-0,022	(0,002)***	-0,02	(0,002)***	0,016	(0,002)***	-0,019	(0,002)***	-0,022	(0,002)***	-0,026	(0,002)***
EIST	0,103	(0,083)	0,2	(0,059)***	0,155	(0,047)***	-0,012	(0,054)	0,195	(0,044)***	0,185	(0,048)***	0,082	(0,044)*
EIST*Temps	-0,002	(0,005)	-0,0003	(0,004)	-0,0015	(0,003)	-0,0014	(0,002)	0,0028	(0,003)	-0,0011	(0,004)	0,0001	(0,002)
EP	-0,58	(0,108)***	0,077	(0,077)	0,067	(0,062)	-0,197	(0,069)***	0,044	(0,057)	0,06	(0,063)	0,041	(0,057)
Privé	0,033	(0,110)	0,035	(0,079)	0,01	(0,064)	0,057	(0,071)	0,051	(0,059)	0,061	(0,064)	0,036	(0,059)
Fille	0,058	(0,033)*	-0,128	(0,032)***	-0,19	(0,031)***	-0,078	(0,029)***	-0,055	(0,029)*	0,084	(0,031)***	-0,163	(0,032)***
Effets aléatoires														
Niveau 3 constante	0,124	(0,022)***	0,0463	(0,011)***	0,0203	(0,007)***	0,0372	(0,009)***	0,016	(0,006)***	0,0201	(0,007)***	0,0118	(0,005)**
(interclasses) pente	0,00047	(0,00009)***	0,0001	(0,00004)***	0,0001	(0,00003)**	0,00002	(0,00002)	0,00006	(0,00002)***	0,00013	(0,00004)***	0,00003	(0,00002)*
Niveau 2 constante	0,582	(0,021)***	0,4744	(0,020)***	0,4889	(0,019)***	0,3957	(0,017)***	0,4052	(0,016)***	0,4362	(0,019)***	0,5341	(0,020)***
(interélèves) pente	0,00103	(0,00009)***	0,0011	(0,0001)***	0,0006	(0,0001)***	0,0002	(0,0001)**	0,0002	(0,0001)**	0,0009	(0,0001)***	0,0001	(0,0001)**
Niveau 1	0,433	(0,010)***	0,5475	(0,013)***	0,5202	(0,013)***	0,6067	(0,014)***	0,504	(0,012)***	0,6016	(0,014)***	0,5238	(0,013)***
-2logL	23920.1		23651.2		23213.4		23145.3		21919.8		23998.0		23102.9	

*significatif au seuil de 10 %
**significatif au seuil de 5 %
***significatif au seuil de 1 %

Note : le tableau présente pour chacune des dimensions visées les principaux résultats des modèles multiniveaux de croissance, à savoir les valeurs des paramètres ainsi que leurs erreurs associées (entre parenthèses) et leur degré de significativité indiqué par des « * ». Dans la partie « Effets fixes », il ressort que le coefficient associé au temps de mesure est très significatif pour toutes les dimensions, ce qui traduit leur développement au cours du temps : les progressions sont positives pour les performances cognitives et le facteur 3 (sensibilisation aux phénomènes environnementaux) et négatives pour les autres dimensions (attitudes à l'égard des sciences). La variable indicatrice EIST (1 pour EIST, 0 pour témoin) indique la différence entre les élèves des deux groupes (EIST et témoin), lors de la première session. L'interaction entre la variable EIST et le temps n'est pas significative, quelle que soit la dimension : les progressions ne sont pas différentes selon que les élèves ont ou non bénéficié de l'EIST. Dans la partie « Effet aléatoire », la variance du niveau 1 indique la variabilité intra-élèves, celle du niveau 2 la variabilité interélèves et celle du niveau 3, la variabilité intercollèges.

Conclusions et perspectives

Les premières analyses de l'effet de l'expérimentation EIST sur les progressions des élèves, observées sur trois temps de mesure, du début de 6e à la fin de 5e, ne font pas ressortir d'effet significatif associé au fait d'avoir bénéficié du dispositif de l'EIST. Cette absence d'effet porte tant sur les progressions en termes d'acquis cognitifs dans le domaine des sciences qu'en termes d'attitudes à l'égard des sciences.

Plusieurs remarques importantes méritent d'être apportées pour éclairer ces premiers résultats. Tout d'abord, rappelons que les collèges engagés dans l'EIST sont des collèges volontaires, et qu'à ce titre, on pouvait craindre l'apparition d'un effet positif de l'expérimentation qui soit « artefactuel », c'est-à-dire ne traduisant pas les effets du dispositif lui-même mais celui de caractéristiques non observées des collèges, en lien avec leur adhésion à l'EIST. Cet effet n'apparaît pas, ce qui évacue mécaniquement la question de la surestimation de l'effet liée au volontariat.

Une deuxième série de commentaires concerne la temporalité de l'évaluation, qui concerne l'évolution des performances et des attitudes du début de 6e à la fin de 5e. Il est tout à fait possible que les effets ne soient pas immédiats et puissent s'observer avec un temps de latence. À cet égard, il conviendra de poursuivre les analyses de progressions, en fin de 4e et en fin de 3e. Cette question est d'autant plus importante que les instruments de mesure des performances en sciences intègrent à ces deux derniers niveaux, un volume plus important de questionnement ouvert et des éléments plus fournis sur la démarche d'investigation.

Troisièmement, nous avons traité ici sur le même plan tous les élèves ayant suivi le dispositif EIST. Mais le temps d'exposition à l'expérimentation semble variable : certains élèves auraient profité du dispositif EIST pendant toute l'année scolaire, tandis que d'autres en auraient bénéficié seulement pendant un ou deux trimestres. Cette variable sera précisée et introduite dans les futures analyses. Cependant, il est peu probable que ce phénomène soit de nature à dissimuler des effets importants.

Parallèlement, ce travail a été l'occasion de présenter des approches liées à la mesure et au suivi des dimensions, en ayant recours à des modélisations psychométriques. Ce champ sera également développé en visant une modélisation intégrée comprenant à la fois le modèle de mesure et le modèle d'évaluation de l'effet. Ceci permettra de mieux tenir compte des erreurs de mesure.

Enfin, signalons que ce suivi longitudinal a permis de constituer une base de données très riche sur le développement des performances et des attitudes à l'égard des sciences au cours des années de collège. Un premier enseignement original concerne la désaffection pour les sciences qui apparaît dès le début du collège à travers la baisse observée sur les variables d'intérêt et de motivation à l'égard des sciences. Ainsi, ces données mériteront d'être analysées plus largement, au-delà de la question de l'évaluation des effets de l'EIST. ■

BIBLIOGRAPHIE

Bressoux, P. (2010), *Modélisation statistique appliquée aux sciences sociales*, De Boeck, 2e édition, Bruxelles.

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

Stocking M.L. et Lord F.M. (1983), Developing a Common Metric in Item Response Theory, *Applied Psychological Measurement*, n° 7, pp. 201-210.

Annexe 1 : détail des groupes suivis

Groupe 1

Tous les élèves de France métropolitaine recevant un enseignement EIST sont évalués de façon exhaustive. Les collèges participant à l'EIST se situent dans les 12 académies suivantes : Clermont-Ferrand, Nice, Toulouse, Montpellier, Poitiers, Bordeaux, Orléans-Tours, Nancy-Metz, Lille, Strasbourg, Versailles, Créteil. S'agissant des élèves de 6e, cela concerne 2 035 élèves répartis dans 35 collèges *(cf. tableau 3)*.

Groupe 2

Dans les établissements participant au programme EIST, une classe suivant un enseignement « traditionnel » est tirée au hasard, et tous les élèves de cette classe participent à l'évaluation.

Groupe 3

Ce groupe constitue l'échantillon-témoin. Il regroupe des élèves d'établissements ne participant pas à l'expérimentation de l'EIST. Dans la mesure où les collèges participant à l'EIST se situent dans 12 académies, et afin de simplifier la communication avec les différents interlocuteurs, l'échantillon-témoin a été tiré dans ces 12 académies uniquement. Une analyse a d'ailleurs montré que la population des 12 académies est représentative de la population totale (France métropolitaine). Un échantillon stratifié selon le secteur de l'établissement (public hors éducation prioritaire, éducation prioritaire, privé) est alors obtenu par un tirage à deux degrés. Le premier degré consiste à tirer des établissements proportionnellement à leur taille (en nombre de classes), et le deuxième consiste à tirer, de façon aléatoire, une classe dans chacun de ces établissements. L'échantillon-témoin est constitué de manière à être de même taille que celui des élèves bénéficiant du programme d'EIST (soit environ 2 000 élèves en 6e et environ 500 élèves en 5e), et que la répartition selon les trois strates soit représentative.

Annexe 2 : questionnaire conatif

Facteur 1 : « sentiment d'efficacité en science »
Ce que je fais en science est intéressant.
La science, c'est trop difficile pour moi.
Je pense que j'ai un bon niveau en science.
Je comprends bien ce que nous faisons en science.
Je pense que je peux réussir en science.

Facteur 2 : « Les sciences dans l'avenir et dans le futur métier »
Quand je serai adulte, j'aimerais continuer à faire de la science.
Quand je serai adulte, j'aimerais exercer une profession dans le domaine scientifique.
Quand je serai adulte, j'aimerais concevoir des objets techniques.
Quand je serai adulte, j'aimerais être chercheur.

Facteur 3 : « Sensibilisation aux phénomènes environnementaux »
*L'augmentation des gaz à effet de serre dans l'atmosphère**.*
*Les pénuries d'eau**.*
*Les déchets nucléaires**.*
*La destruction des forêts**.*
*La disparition de certaines espèces de plantes et d'animaux**.*

Facteur 4 : « attrait pour les expériences »
Je refais chez moi des expériences réalisées en classe.*
J'aime les jouets à caractère scientifique.*
Je suis capable d'imaginer une expérience.
Faire des sciences au collège, c'est faire des expériences par moi-même.

Facteur 5 : « engagement pour les sciences »
Je participe en science parce que j'aime bien chercher.
Je travaille en science parce que j'aime bien cette discipline.
Faire des sciences au collège, c'est se poser des questions sur le monde qui nous entoure.
Faire des sciences au collège, c'est chercher des informations dans les livres ou sur internet.
Faire des sciences au collège, c'est expliquer ce que j'ai compris.

Facteur 6 : « Activités en relations avec les sciences en dehors de l'école »
Je regarde des émissions scientifiques à la télévision.*
J'aime lire des livres ou des revues scientifiques.*
Je cherche des documents scientifiques sur internet.*

Note : questions sans astérisque : les élèves devaient répondre « tout à fait d'accord » « d'accord », « pas d'accord » ou « pas du tout d'accord » ; questions marquées d'un astérisque « * » : les élèves devaient répondre « très souvent », « souvent », « parfois » ou « jamais » ; questions marquées de deux astérisques « ** » : les élèves devaient répondre « je connais ce sujet et je pourrais assez bien l'expliquer », « j'en ai entendu parler mais je serais incapable d'expliquer ce que c'est exactement », « je n'en ai jamais entendu parler ».