CEDRE

Cycle des Évaluations Disciplinaires Réalisées sur Échantillons

Rapport technique

Sciences expérimentales 2013

École

Auteurs :
Sandra ANDREU
Yann ETEVE
Emilie GARCIA
Thierry ROCHER
Ronan VOURC'H

Bureau de l'évaluation des élèves
DEPP - Direction de l'évaluation, de la prospective et de la performance
Ministère de l'éducation nationale

Février 2015

Table des matières

Introduction

La DEPP met en place des dispositifs d'évaluation des acquis des élèves reposant sur des épreuves standardisées. Elle est également maître d'œuvre pour la France de diverses évaluations internationales. Ces programmes d'évaluations dites « bilans » sont des outils pour le pilotage d'ensemble du système éducatif. Ainsi, les évaluations du CEDRE révèlent, en référence aux programmes scolaires, 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.

Leur méthodologie de construction s'appuie sur les méthodes de la mesure en éducation et les modèles psychométriques. Ces évaluations concernent de larges échantillons représentatifs d'établissements, de classes et d'élèves. Elles permettent d'établir des comparaisons temporelles afin de suivre l'évolution des performances des systèmes éducatifs.

Ce rapport présente l'ensemble des méthodes qui sont employées pour réaliser les évaluations du cycle CEDRE, en balayant des aspects aussi divers que la construction des épreuves, la sélection des échantillons ou bien la modélisation des résultats. L'objectif est de rendre accessible les fondements méthodologiques de ces évaluations, dans un souci de transparence. La publication de ce rapport fait d'ailleurs partie des engagements pris par la DEPP dans le cadre du processus de certification des évaluations du cycle CEDRE.

1 Cadre d'évaluation

1.1 Objectifs

Le cycle des évaluations disciplinaires réalisées sur échantillons (CEDRE) 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 scolaires. 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.

Ces évaluations apportent un éclairage qui intéresse tous les niveaux du système éducatif, des décideurs aux enseignants sur le terrain, en passant par les formateurs : elles informent, bien sûr, sur les compétences et les connaissances des élèves à la fin d'un cursus, mais elles éclairent également sur l'attitude et la représentation des élèves à l'égard de la discipline. Elles interrogent les pratiques d'enseignement au regard des programmes et elles contribuent à enrichir la réflexion générale sur l'efficacité et la performance de notre système éducatif.

Ces évaluations étant passées auprès d'échantillons statistiquement représentatifs de la population scolaire de France métropolitaine, aucun résultat par élève, établissement ni même par département ou académie ne peut être calculé.

CEDRE a débuté en 2003 avec l'évaluation des compétences générales. Afin d'assurer une comparabilité dans le temps, l'évaluation est reprise pour chaque discipline selon un cycle de six ans jusqu'en 2012, et de cinq ans depuis 2012 (tableau 1)

Tableau 1 – Cycle des évaluations disciplinaires réalisées sur échantillons (CEDRE) depuis 2003

Discipline évaluée	**Début du cycle**	**Reprises**	
Maîtrise de la langue et compétences générales	2003	2009	2015
Langues étrangères	2004	2010	2016
Attitude à l'égard de la vie en société	2005	–	–
Histoire, géographie et éducation civique	2006	2012	2017
Sciences	2007	2013	2018
Mathématiques	2008	2014	2019

1.2 Compétences visées

La modification des programmes en 2008, avec l'intégration du socle commun, a modifié l'approche de l'évaluation en sciences expérimentales et en technologie dans le cadre de CEDRE en fin d'école. En 2007, les items qui mesuraient les connaissances étaient bien distincts de ceux qui abordaient les compétences. Depuis la mise en place du socle commun, les compétences sont définies comme une combinaison de connaissances, de capacités et d'attitudes. C'est au cycle 3 de l'école primaire (CE2, CM1, CM2) que les sciences sont abordées en tant que disciplines à part entière. Depuis les programmes de 2008, elles font partie de la compétence 3, définie comme les principaux éléments de mathématiques et la culture scientifique et technologique.

Voici un extrait du Bulletin Officiel de 2008 portant sur les horaires et programmes d'enseignement de l'école primaire :

« Les sciences expérimentales et les technologies ont pour objectif de comprendre et de décrire le monde réel, celui de la nature et celui construit par l'Homme, d'agir sur lui, et de maîtriser les changements induits par l'activité humaine. Leur étude contribue à faire saisir aux élèves la distinction entre faits et hypothèses vérifiables d'une part, opinions et croyances d'autre part. Observation, questionnement, expérimentation et argumentation pratiqués,[...] sont essentiels pour atteindre ces buts ; c'est pourquoi les connaissances et les compétences sont acquises dans le cadre d'une démarche d'investigation qui développe la curiosité, la créativité, l'esprit critique et l'intérêt pour le progrès scientifique et technique. Familiarisés avec une approche sensible de la nature, les élèves apprennent à être responsables face à l'environnement, au monde vivant, à la santé. Ils comprennent que le développement durable correspond aux besoins des générations actuelles et futures. En relation avec les enseignements de culture humaniste et d'instruction civique, ils apprennent à agir dans cette perspective. Les travaux des élèves font l'objet d'écrits divers consignés, par exemple, dans un carnet d'observations ou un cahier d'expériences. » (BO n°3 du 19 juin 2008)

Deux changements sont intervenus entre 2007 et 2013. Le premier concerne les « sciences » et le second la « maîtrise de la langue ».

1.2.1 Les sciences

L'évaluation 2013 porte sur les huit domaines définis par les programmes : « Le ciel et la Terre », « La matière », « L'énergie», « Le fonctionnement du vivant et unité et diversité du vivant», « Le fonctionnement du corps humain et la santé », « Les êtres vivants dans leur environnement », « Les objets techniques », plus un domaine défini par le socle « Environnement et développement durable ».

Les TIC (compétences informatiques) étaient un chapitre des sciences dans les programmes de 2002. Elles ont été exclues de l'évaluation dans les épreuves de 2013, étant regroupées maintenant sous la compétence 4 du socle commun : la maîtrise des techniques usuelles de l'information et de la communication.

1.2.2 La maîtrise de la langue

Lors de l'évaluation 2007, nous avions noté la place importante que tenait la maîtrise de la langue dans le domaine scientifique. Pour acquérir des notions et progresser les élèves doivent développer conjointement leurs compétences. La confrontation avec des écrits leur apporte de nouveaux mots, ce lexique étant constitutif de notions. Ces dernières viennent s'intégrer à ce qui est déjà acquis et demande à être à nouveau interrogé lors de la lecture de nouveaux textes. Les programmes de 2002 se focalisaient sur la maîtrise de la langue transcendant toutes les disciplines. La production d'écrit y prenait une part notable. En effet, c'est lors du passage à l'écrit que l'élève rassemble, organise et clarifie sa pensée. Ce travail difficile permet d'intégrer de nouveaux savoirs tout en accédant à une meilleure maîtrise de la langue. En 2013, les aspects de la production d'écrit sont positionnés dans la compétence 1 ; néanmoins, les concepteurs des évaluations ont créé des items monopolisant cette compétence afin d'observer à quel niveau les réussites seraient observées. Les compétences permettant de cerner les acquis des élèves ont été retenues selon les finalités assignées à l'enseignement des sciences expérimentales et de la technologie préconisées en introduction des programmes de l'école. Une évaluation dans ce domaine a pour objet de mesurer la capacité de l'enseignement dispensé dans cette discipline à mettre en oeuvre ses finalités.

1.2.3 Compétences évaluées

Deux grandes compétences ont été évaluées : les unités proposées se composaient donc d'items visant à mesurer les **connaissances mémorisées** d'une part **(connaître)**, et, d'autre part la capacité des élèves à **traiter les documents** mis à leur disposition **(raisonner)**.

De façon transversale, s'agissant de la **démarche d'investigation**, telle que définie par les programmes 2008, en classe, elle peut recourir à diverses formes de travail :
- expérimentation directe conçue et réalisée par les élèves ;
- réalisation matérielle (recherche d'une solution technique) ;
- observation directe ou assistée par un instrument, avec ou sans mesure ;
- recherche documentaire.

Dans cette évaluation, la démarche d'investigation a été évaluée de deux manières :
– à travers une des étapes de l'expérimentation (manipulation) ;
– à travers la recherche documentaire.
Cette évaluation comporte trois sujets destinés à mettre en œuvre les étapes de la démarche d'investigation et à tester les connaissances dans deux des huit domaines définis par les programmes.

C'est dans ce contexte, que des modules de situations dites complexes ont été construits. Les trois modules ont été élaborés à partir des travaux du groupe PRESTE de l'académie de Toulouse. Dans ces situations, la démarche d'investigation est mise en œuvre à partir de recherches documentaires sans recours à l'expérimentation directe.

La compétence 1 - « La maîtrise de la langue française » est en jeu dans les trois situations proposées, mais n'a pas fait l'objet d'une évaluation directe dans ce test, ainsi que des compétences dans le domaine associé des mathématiques.

Bien entendu, les deux grandes compétences (connaître et raisonner) ne sont pas hermétiquement étanches entre elles, ni hiérarchisées. Leur présentation distincte tente de clarifier des opérations intellectuelles demandées à l'élève pendant un apprentissage plus global. La déclinaison de ces domaines de compétences en composantes a permis de construire l'évaluation en lien avec les programmes en prenant en compte ce que chaque composante apporte à la maîtrise de la compétence.

Dans le cadre d'une évaluation bilan, cette présentation permet de positionner l'élève selon différents niveaux d'acquisition dans les deux compétences. Mais elle permet également de faire apparaître des regroupements de niveaux de savoir-faire, d'établir des liens entre la maîtrise de plusieurs compétences qui permettent de mieux cerner les comportements d'acquisition et de donner du sens aux étapes de la construction du savoir...

Néanmoins, dans tout type d'évaluation il faut être sûr que l'on évalue bien la compétence visée, c'est à dire qu'il faut aussi tenir compte du poids de la connaissance déjà acquise par l'élève, du poids de la tâche qu'on lui demande d'effectuer pour mettre en œuvre la compétence visée et du poids du document proposé, soit en tant que support, soit en tant qu'objet d'évaluation. En classe, l'enseignant fait souvent une appréciation globale du travail de l'élève où l'intuitif joue un rôle d'autant plus important qu'il connaît les élèves, qu'il vient de faire ses cours, qu'il s'est fixé des objectifs. L'évaluation de l'élève au cours d'une année scolaire se fait sur le court terme, à la suite d'une leçon, d'un chapitre ou d'une séquence d'enseignement, même si l'enseignant repère aussi à ce moment

là les compétences acquises ou en construction. Dans le cadre de cette opération, il a fallu se confronter à une « démultiplication » de l'évaluation puisqu'il s'agissait de faire un bilan des acquis sur des connaissances et sur différents types de compétences acquis au cours de cinq années scolaires.

1.3 Construction du test

Dans le cadre d'une évaluation sur support papier, l'évaluation se compose d'un ensemble de cahiers, constitués de blocs, qui sont eux-mêmes composés d'unités (ensemble d'items). La préparation des cahiers et de leur contenu fait intervenir des concepteurs, qui sont le plus souvent des professeurs. Dans chaque discipline, les enseignants sont coordonnés par un chargé d'étude, personnel du bureau de l'évaluation des élèves de la DEPP, sous la responsabilité du chef du bureau.

1.3.1 Elaboration des questionnaires

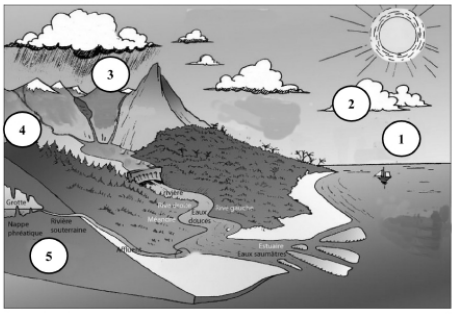
Les items sont le fruit d'un travail collectif des concepteurs, encadré par le chargé d'étude, l'inspection et l'inspection générale. Un item proposé par un concepteur, pédagogue de terrain ayant une bonne connaissance des pratiques de classe, fait l'objet d'une discussion contradictoire jusqu'à aboutir à un consensus, au final validé par le chargé d'étude et l'inspection. L'item fait alors l'objet d'un cobayage, c'est-à-dire d'une passation auprès d'une ou plusieurs classes pour estimer la difficulté de l'item et recueillir les réactions des élèves.

Une application *ad hoc* est utilisée en interne pour faciliter la création des items, ainsi que leur édition, leur stockage et la gestion des évaluations (cf. plus loin l'encadré « GEODE »).

Exemples d'items

Exemple 1 : série de vrai-faux

Voici le dessin d'un paysage.

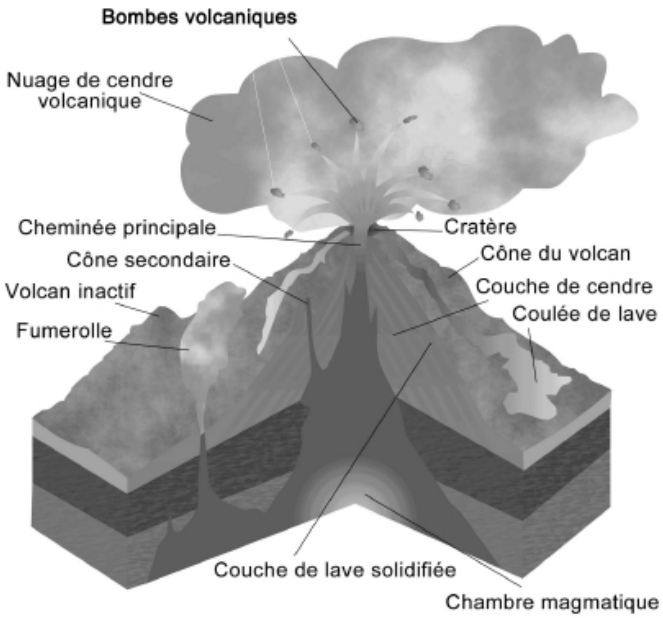


Pour chaque phénomène, coche la bonne réponse.

	vignette 1	vignette 2	vignette 3	vignette 4	vignette 5	
Condensation	☐ 1	■ 2	☐ 3	☐ 4	☐ 5	E5XME460201
Évaporation	■ 1	☐ 2	☐ 3	☐ 4	☐ 5	E5XME460202
Infiltration	☐ 1	☐ 2	☐ 3	☐ 4	■ 5	E5XME460203
Précipitation	☐ 1	☐ 2	■ 3	☐ 4	☐ 5	E5XME460204
Ruissellement	☐ 1	☐ 2	☐ 3	■ 4	☐ 5	E5XME460205

Exemple 2 : texte à trous

Lis le document suivant.



Replace les différentes légendes du schéma dans le texte à trous ci-dessous.

Les [ ] et le [ ] s'envolent dans le ciel au dessus du [ ] .

Une [ ] s'écoule sur le [ ] .

[ ] est situé au sommet de la [ ] .

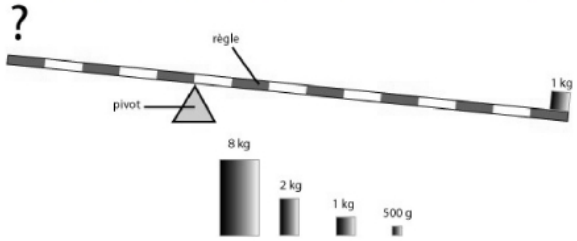
Le cône volcanique est composé de couches de [ ] .

L'ensemble du cône volcanique est recouvert par de la [ ] .

E5XCE970101
E5XCE970102
E5XCE970103
E5XCE970104
E5XCE970105
E5XCE970106
E5XCE970107
E5XCE970108
E5XCE970109

Exemple 3 : question à choix multiples (QCM)

On réalise le dispositif suivant avec une règle, un pivot et une masse de 1kg.



On souhaite équilibrer le dispositif, quelle masse faut-il placer à l'autre extrémité de la règle ?

Coche la bonne réponse.

1 ☐ 500 g
2 ☐ 1 kg
3 ■ 2 kg
4 ☐ 8 kg

E5XOE1560101

Quels sont les équilibres de l'évaluation CEDRE ?

La comparaison est effectuée sur un « noyau dur » d'items qui représentent des connaissances et des compétences dans les différents domaines des sciences expérimentales et de la technologie des programmes 2008. Durant la période 2007-2013, un nouveau programme a été mis en place. Il intègre, en tant qu'attendu de fin d'école les compétences du socle commun. L'originalité de l'évaluation 2013 par rapport à la précédente réside dans la présence de situations complexes. Elle comportait trois sujets destinés à mettre en œuvre les étapes de la démarche d'investigation et à tester les connaissances et les compétences des élèves dans deux des huit domaines définis par les programmes 2008. Une situation complexe est définie comme une tâche mobilisant à la fois des ressources internes (culture scientifique, concepts, capacités, connaissances...) et des ressources externes (protocoles, ressources documentaires...) ; elle fait donc directement référence à la notion de compétence. Ces situations sont scénarisées (organisées) de manière à inciter l'élève à utiliser des connaissances et des capacités, à la croisée des éléments du programme et du socle commun de connaissances et de compétences. L'évaluation 2013 a pris en compte les nouvelles orientations. Elle permet d'observer, pour cette deuxième prise d'informations son positionnent dans l'échelle des acquis des élèves.

L'évaluation CEDRE sciences 2013 se présente sous la forme de QCM et de questions ouvertes (voir l'encadré « Exemples d'items »). Elle est composée d'items repris à l'identique par rapport à 2007 (98 items soit environ 60 % du test) et d'items nouveaux (81 items).

1.3.2 Constitution des cahiers

L'évaluation CEDRE sciences 2013 en CM2 est constituée de 13 cahiers tournants intégrant un ensemble de 13 blocs d'évaluations contenant des items de 2007 repris à l'identique pour assurer une comparaison diachronique et de nouveaux items qui ont fait l'objet d'une expérimentation en 2012 (tableau 2).

Tableau 2 – Répartition des blocs dans les cahiers pour l'évaluation CEDRE Sciences 2013

Cahier	*Séquence 1*	*Séquence 2*	*Séquence 3*	*Séquence 4*	
E01	*B16*	*B18*	*B20*	*B26*	BS1
E02	**B17**	*B20*	**B19**	**B24**	BS1
E03	*B18*	**B19**	**B21**	**B25**	BS1
E04	*B20*	**B21**	**B22**	**B23**	BS1
E05	**B19**	**B22**	*B26*	**B27**	BS1
E06	**B21**	*B26*	**B24**	**B28**	BS1
E07	**B22**	**B24**	**B25**	*B16*	BS1
E08	*B26*	**B25**	**B23**	**B17**	BS1
E09	**B24**	**B23**	**B27**	*B18*	BS1
E10	**B25**	**B27**	**B28**	*B20*	BS1
E11	**B23**	**B28**	*B16*	**B19**	BS1
E12	**B27**	*B16*	**B17**	**B21**	BS1
E13	**B28**	**B17**	*B18*	**B22**	BS1

Blocs composés d'items de 2007
Blocs composés de nouveaux items 2013
BS1= Questionnaire de contexte

La méthodologie des cahiers tournants permet d'évaluer un nombre important d'items sans allonger le temps de passation. Les items sont ainsi répartis dans des blocs d'une durée de 30 minutes et les blocs sont ensuite distribués dans les cahiers tout en respectant certaines contraintes (chaque bloc doit se retrouver un même nombre de fois au total et chaque association de blocs doit figurer au moins une fois dans un cahier). Ce dispositif, couramment utilisé dans les évaluations-bilans, notamment les évaluations internationales, permet d'estimer la probabilité de réussite de chaque élève à chaque item sans que chaque élève ait passé l'ensemble des items.

Au final, pour l'évaluation CEDRE 2013, chaque cahier comprend quatre séquences cognitives de 30 minutes chacune. Elles sont complétées par une cinquième séquence (questionnaire de contexte), identique dans tous les cahiers, dans laquelle l'élève doit renseigner plusieurs éléments concernant l'environnement familial dans lequel il évolue, sa perception des sciences et de son environnement scolaire, sa motivation et son autonomie face au travail scolaire.

GEODE (Gestion électronique d'outils et documents d'évaluation) : un outil de création et de stockage des évaluations

Objectifs

Le bureau de l'évaluation des élèves coordonne chaque année plusieurs évaluations afin d'apprécier le niveau de connaissances et de compétences des élèves en référence aux programmes officiels. Ces évaluations utilisent des livrets d'évaluation sur format papier et/ou électroniques.

L'application GEODE (gestion électronique d'outils et documents d'évaluation) est une application de création et de gestion dématérialisées des évaluations. Développée en 2009, elle a pour objectif de soutenir de bout en bout le processus de création des exercices et de constitution des cahiers et supports électroniques, allant jusqu'au bon à imprimer pour les évaluations papiers ou la génération d'une maquette de site web pour l'évaluation électronique.

L'application permet la conservation, l'indexation et la recherche des documents ou fichiers joints. Une partie des données textuelles, images, sons ou vidéos y est donc stockée que ce soit pour les évaluations papiers (cahier d'évaluations) ou les évaluations électroniques (outil de maquettage).

Principes fonctionnels

GEODE permet ainsi l'harmonisation des pratiques et formats de documents. La dématérialisation des documents rend indépendant l'éditeur (OpenOffice, Word,...) tout en permettant des variantes selon les disciplines. L'application dispose d'une GED (gestion électronique de documents) intégrée capable de gérer du texte, des images, du son et de la vidéo sous forme d'objets. Les cahiers sont générés au format Open Office principalement pour le format « papier », l'utilisation de la même technologie permet de générer du HTML pour la partie évaluation électronique (outil de maquettage).

1.4 Passation des évaluations

La passation de l'évaluation finale a eu lieu en mai 2013. Comme en 2007, cette évaluation a été précédée d'une expérimentation l'année -1 de façon à tester un grand nombre d'items auprès d'un échantillon réduit d'écoles (environ 130).

Dans chaque école, une personne a été désignée comme étant le coordinateur de l'évaluation, son rôle étant de veiller au strict respect de la procédure à suivre pour que l'épreuve soit administrée dans les mêmes conditions quel que soit l'école. La collecte de l'information s'est faite par questionnaires papier crayon.

La passation a été organisée en quatre séquences séparées par des pauses ; les quatre séquences ne devant être passées dans la même journée (tableau 3). La durée de l'évaluation était estimée à deux heures par élève auxquelles il était important d'ajouter les temps d'explication et de gestion du matériel.

Tableau 3 – Passation des épreuves

Jour	**Séquence**	**Durée**
Jour 1	Séquence 1	Environ 45 min.
Jour 2	Séquence 2	Environ 30 min.
Jour 3	Séquence 3	Environ 30 min.
Jour 4	Séquence 4	Environ 45 min.

La première séquence était constituée d'une série d'exemples réalisée avec l'ensemble de la classe (15 minutes), puis d'une phase de travail individuel sur le cahier, le bloc n°1 (environ 30 minutes). La dernière séquence est consacrée au bloc n°4 et au questionnaire de contexte (15 minutes) Entre chaque séquence, l'administrateur devait relever les cahiers, qui ne devaient pas être gardés par les élèves.

Les enseignants de la classe ou des classes concernées ont également dû renseigner un questionnaire de contexte. L'anonymat des élèves et des personnels a été respecté, chaque cahier étant repéré par un numéro allant de 1 au nombre total d'élèves ou d'enseignants ayant répondu dans l'école. Une fois l'évaluation terminée, les cahiers et questionnaires devaient être renvoyés dans des conditionnements prévus à cet effet, pré affranchis et pré étiquetés. Aucun travail de correction n'a été demandé aux enseignants.

2 Sondage

2.1 Méthodes

2.1.1 Sondage par grappes stratifié

Dans le premier degré, nous ne disposons pas des informations auxiliaires présentes dans les bases de sondage de la DEPP, telle que la PCS des parents par exemple. Il n'est donc pas possible de réaliser un tirage équilibré comme c'est le cas pour les évaluations CEDRE en 3e.

Le tirage consiste donc simplement en un sondage par grappes stratifié. La stratification porte généralement sur la zone de scolarisation et tous les élèves de CM2 des écoles sélectionnés participent. Le choix de sondages par grappe est motivé par la facilité de gestion. En effet, le fait de sélectionner tous les élèves d'une école permet d'éviter de mettre en place des procédures de tirage au sort d'élèves une fois les écoles tirées.

Par ailleurs, au moment du tirage de l'échantillon, les écoles ayant déjà été sélectionnée pour une autre évaluation la même année sont exclues de la base de sondage. Les probabilités d'inclusion sont donc recalculées pour tenir compte de ces exclusions tout en gardant une représentativité nationale (cf. encadré « tirage après élimination de la base des échantillons précédemment tirés »).

2.1.2 Redressement de la non réponse : calage sur marges

Comme toute enquête réalisée par sondage, les évaluations des élèves sont exposées à la non-réponse. Bien que les 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 connait 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 :

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

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 1.

Tirage 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 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) dans un cadre de tirage équilibré est en cours de développement avec une application à des données issues d'évaluations standardisées (Christine & Rocher, 2012).

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 $\lambda_j = \frac{\pi_j^2}{1-\pi_j^1}$

Condition fondamentale

Comme il s'agit d'une probabilité, la condition fondamentale est que $\lambda_j \in [0,1]$. Comme $\lambda_j = \frac{\pi_j^2}{1-\pi_j^1}$, 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 : Eclair), cette condition pourrait ne pas être satisfaite. Cependant, de façon concrète, la condition a toujours été respectée dans les plans de sondage réalisés.

2.1.3 Calcul de précision : méthode

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, appelées erreurs standard.

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.

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 $g_i e_i$.

2.2 Echantillonnage

Le champ des évaluations CEDRE à l'école est celui des élèves de CM2 scolarisés dans des écoles publiques et privées sous contrat de France métropolitaine. Pour des raisons de coût, les écoles ayant moins de 6 élèves de CM2 sont exclues du champ.

Comme nous l'avons dit, la base de sondage est relativement pauvre en informations dans le premier degré. Nous disposons cependant d'informations sur les établissements scolaires, comme le secteur d'enseignement.

2.2.1 Echantillon 2007

Modalités de sélection

Le tirage des écoles est à allocation proportionnelle selon six strates. Ensuite, tous les élèves de CM2 des écoles sélectionnées sont interrogés. Préalablement au tirage, les échantillons de CEDRE histoire-géographie 2006 et de l'expérimentaion CEDRE Sciences 2006 ont été retirés de la base de sondage.

Stratification

La stratification prend en compte à la fois la taille et le secteur d'enseignement de l'école :

1. Écoles publiques hors ZEP (plus de 15 élèves)
2. Écoles privées (plus de 15 élèves)

3. Écoles publiques en ZEP (plus de 15 élèves)
4. Écoles publiques hors ZEP (moins de 15 élèves)
5. Écoles privées (moins de 15 élèves)
6. Écoles publiques en ZEP (moins de 15 élèves)

On vise environ 6 000 élèves.

Base de sondage

Le tableau 4 présente la répartition de la population ciblée dans les différentes strates.

Tableau 4 – Répartition dans la base de sondage - Base CEDRE CM2

	Nb écoles	Nb élèves
Écoles publiques hors ZEP (plus de 15 élèves)	13 344	428 604
Écoles privées (plus de 15 élèves)	2 839	96 742
Écoles publiques en ZEP (plus de 15 élèves)	2 532	89 605
Écoles publiques hors ZEP (moins de 15 élèves)	6 970	64 279
Écoles privées (moins de 15 élèves)	1 210	11 358
Écoles publiques en ZEP (moins de 15 élèves)	352	3 343
Total	**27 247**	**693 931**

Échantillon

Le tableau 5 présente la répartition de l'échantillon dans les différentes strates. Au total, 219 écoles ont été sélectionnées.

Tableau 5 – Répartition dans l'échantillon - CEDRE Sciences CM2

	Nb écoles	Nb élèves
Écoles publiques hors ZEP (plus de 15 élèves)	52	1644
Écoles privées (plus de 15 élèves)	49	1666
Écoles publiques en ZEP (plus de 15 élèves)	53	1855
Écoles publiques hors ZEP (moins de 15 élèves)	32	288
Écoles privées (moins de 15 élèves)	28	252
Écoles publiques en ZEP (moins de 15 élèves)	5	45
Total	**219**	**5750**

2.2.2 Echantillon 2013

Modalités de sélection

Le tirage des écoles est à allocation proportionnelle selon trois strates. Ensuite, tous les élèves de CM2 des écoles sélectionnées sont interrogés.

Stratification

Une stratification est réalisée en fonction du secteur d'enseignement :

1. Public hors Éducation Prioritaire (PU)
2. Public en Éducation Prioritaire (EP)
3. Privé (PR)

On vise environ 6 250 élèves ce qui correspond à 178 écoles dans la strate 1, 25 écoles pour la strate 2 et 35 écoles pour la strate 3.

Champ et exclusions

Pour l'année 2013, nous documentons le champ de l'évaluation qui est l'ensemble des élèves de CM2 de France métropolitaine (tableau 6).

Tableau 6 – Exclusions pour la base de sondage - CEDRE Sciences CM2

	Ecoles	**Elèves**
Ecoles accueillant des CM2 hors TOM	3 772	809 218
On retire les écoles hors contrat	32 528	806 801
On retire les écoles spécialisées	32 509	806 269
On retire les petites écoles (<6 CM2)	29 916	797 076
Exclusion des DOM	29 005	757 822
Base CM2 CEDRE sciences	**29 005**	**757 822**

Base de sondage

Le tableau 7 présente la répartition de la population ciblée dans les différentes strates.

Échantillon

Le tableau 8 présente la répartition de l'échantillon dans les différentes strates. Au total, 239 écoles ont été sélectionnées et 6 271 élèves sont attendus.

Tableau 7 – Répartition dans la base de sondage - CEDRE sciences CM2

	Ecoles	**Elèves**	**Nb moyen de CM2 par école**
1. Public hors EP	21 737	545 340	25.1
2. EP	3 078	97 714	31.7
3. Privé	4 190	114 768	27.4
Total	**29 005**	**757 822**	

Tableau 8 – Répartition dans l'échantillon - CEDRE sciences CM2

	Nb écoles	**Nb élèves attendus**
1. Public hors EP	179	4 442
2. EP	25	848
3. Privé	35	981
Total	**239**	**6 271**

2.3 Etat des lieux de la non-réponse

2.3.1 Non-réponse totale

Parmi la non-réponse totale, nous distinguons selon la non-réponse d'écoles entières ou la non-réponse d'élèves dans les écoles participantes. Les chiffres suivants ont été observés pour 2013. Tout d'abord, 93,7 % des écoles de l'échantillon ont répondu à l'évaluation. Les 15 écoles non répondantes représentent 487 élèves (tableau 9).

Tableau 9 – Non-réponse des écoles

strate	N écoles attendues	N écoles répondantes	% de écoles répondantes	N élèves non répondants
1- public hors EP	179	169	94,4%	297
2- EP	25	24	96%	60
3- privé	35	31	88,6%	130
Total	**239**	**224**	**93,7%**	**487**

Parmi les écoles ayant répondu, 93,7 % des élèves ont participé à l'évaluation (tableau 10).

Au final, 90,8 % des effectifs attendus ont participé (tableau 11).

Tableau 10 – Non-réponse des élèves

strate	N élèves attendus classes répondantes	N élèves répondants classes répondantes	% élèves répondants
1- public hors EP	4 363	4 088	93,7%
2- EP	818	757	92,5%
3- privé	895	850	95%
Total	**6 076**	**5 695**	**93,7%**

Tableau 11 – Non-réponse globale (écoles et élèves)

strate	N élèves attendus	N élèves répondants	% élèves répondants
1- public hors EP	4 442	4 088	92%
2- EP	848	757	89,3%
3- privé	981	850	86,6%
Total	**6 271**	**5 695**	**90,8%**

2.3.2 Valeurs manquantes et imputation

Dans le cas où certaines données sont manquantes, nous procédons à des imputations. Cela concerne uniquement les variables sexe et année de naissance, afin de pouvoir réaliser des statistiques selon ces variables sur l'échantillon complet, quelle que soit l'analyse. Nous imputons aléatoirement les valeurs manquantes de ces deux variables, de manière à respecter la répartition des répondants.

2.3.3 Non-réponse partielle et terminale

Lorsque des non-réponses sont observées aux items, nous distinguons les cas suivants :

– La non-réponse partielle : un élève n'a pas répondu à certains items dans le cahier.
– La non-réponse terminale : un élève s'est arrêté avant la fin du cahier soit par manque de temps soit par abandon.

Dans le premier cas, les non-réponses sont traitées comme des échecs (code "0"). Le second cas conduit à déterminer des règles. Nous considérons que si un élève a passé moins de 50 % d'une séquence, il n'a pas vu la séquence, les valeurs manquantes sont donc traitées de manière structurelle (code "s").

2007

En 2007, les cahiers étaient composés de deux séquences. La non réponse terminale a été étudiée par séquence et par cahier. Parmi les élèves ayant de la non réponse terminale, il y en a en moyenne 9,3 pour la 1ère séquence et 12 pour la 2ème séquence.

Au final, pour 2007, on considère que :

- 87 élèves n'ont pas vu la séquence 1 dont :
  - 56 n'ont répondu à aucun item de la séquence
  - 31 ont répondu à moins de 50 % de la séquence
- 95 élèves n'ont pas vu la séquence 2 dont :
  - 67 n'ont répondu à aucun item de la séquence
  - 28 ont répondu à moins de 50 % de la séquence

Les élèves dont les deux séquences sont codées en "s" sont considérés comme de la non réponse totale. C'est le cas pour **28 élèves**. Au final, l'échantillon de 2007 est composé de 4 128 élèves.

2013

Les cahiers élèves sont composés de quatre séquences. La non réponse terminale a été étudiée par séquence et par cahier. Parmi les élèves ayant de la non réponse terminale, il y en a en moyenne entre 5 et 5,5 selon les séquences.

Au final, on considère que :

- 94 élèves n'ont pas vu la séquence 1 dont :
  - 82 n'ont répondu à aucun item de la séquence
  - 12 ont répondu à moins de 50 % de la séquence
- 108 élèves n'ont pas vu la séquence 2 dont :
  - 100 n'ont répondu à aucun item de la séquence
  - 8 ont répondu à moins de 50 % de la séquence
- 116 élèves n'ont pas vu la séquence 2 dont :
  - 107 n'ont répondu à aucun item de la séquence
  - 9 ont répondu à moins de 50 % de la séquence
- 176 élèves n'ont pas vu la séquence 2 dont :
  - 156 n'ont répondu à aucun item de la séquence
  - 20 ont répondu à moins de 50 % de la séquence

Les élèves dont les quatre séquences sont codées en " s " sont considérés comme de la non réponse totale. C'est le cas pour **18 élèves**.

2.4 Redressement

Pour tenir compte de la non réponse, l'échantillon a été redressé à l'aide d'un calage sur marge. Préalablement au calage, on effectue tout d'abord une post-stratification. Puis, deux variables de calage sont utilisées :

– la répartition selon le sexe dans la population ;
– la répartition selon le retard scolaire.

Le tableau 12 montre que l'ampleur du calage est très réduit.

Tableau 12 – Comparaison entre les marges de l'échantillon et les marges dans la population

	Modalité ou variable	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population
Retard	1	91 918.35	90 939	12.13	12
	2	665 903.66	666 883	87.87	88
Sexe	1	379 626.37	380 427	50.09	50.2
	2	378 195.64	377 395	49.91	49.8
Strate	1	545 340	545 340	71.96	71.96
	2	97 714	97 714	12.89	12.89
	3	114 768	114 768	15.14	15.14

2.5 Précision

L'erreur standard (*se*) peut être calculée sur le score moyen de chaque année (tableau 13).

Tableau 13 – Scores moyens en sciences et erreurs standard associées

Année	Score moyen	Erreur standard
2007	250	1.49
2013	249	1.27

Pour savoir si l'évolution entre 2007 et 2013 est significative, il faut donc calculer la valeur suivante :

$$\frac{\left|\hat{\bar{Y}}_{2013} - \hat{\bar{Y}}_{2007}\right|}{\sqrt{se^2_{\hat{\bar{Y}}_{2013}} + se^2_{\hat{\bar{Y}}_{2007}}}} \tag{2}$$

Avec une valeur de 0,38 (inférieure à 1,96), cela signifie que la baisse du score moyen observée entre 2007 et 2013 n'est pas statistiquement significative.

Les erreurs standards sont également calculées pour les répartitions dans les différents groupes de niveaux (tableaux 14 et 15).

Tableau 14 – Répartition en % dans les groupes de niveaux en sciences

Année	Groupe < 1	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5
2007	1.9	13.1	29.1	27.8	18.2	10.0
2013	2.4	13.3	28.4	29.0	17.1	9.9

Tableau 15 – Erreurs standards des répartitions en % dans les groupes de niveaux en sciences

Année	Groupe < 1	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5
2007	0.3	0.7	0.9	0.9	0.7	0.9
2013	0.2	0.6	0.7	0.7	0.6	0.6

Design effect

L'effet du plan de sondage (*Design Effect*) permet de rapporter l'erreur de mesure faite par un tirage spécifique à l'erreur de mesure qui aurait été faite en procédant à un sondage aléatoire simple (SAS) du même nombre d'élèves. Pour la moyenne d'une variable Y et un plan de sondage complexe P, il est défini par :

$$D_{eff} = \frac{V_P(\hat{\bar{Y}})}{V_{SAS}(\hat{\bar{Y}})} \tag{3}$$

Dans le cas d'un sondage en grappes, la précision est dégradée en comparaison d'un sondage aléatoire simple (tableau 16). Cela signifie qu'en 2013, un sondage aléatoire simple avec un effectif 3,4 fois moins important aurait conduit au même niveau de précision.

Tableau 16 – Effet du plan de sondage

Année	Erreur Standard	Erreur SAS	*Design Effect*
2007	1.49	0.78	3.67
2013	1.27	0.69	3.43

3 Analyse des items

3.1 Méthodologie

3.1.1 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) :

$$Y_i^j = \begin{cases} 1 & \text{si l'élève } i \text{ réussit l'item } j \\ 0 & \text{si l'élève } i \text{ échoue à l'item } j \end{cases} \tag{4}$$

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

$$p_j = \frac{1}{n} \sum_{i=1}^{n} Y_i^j \tag{5}$$

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.

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

$$S_i = \sum_{j=1}^{J} Y_i^j \tag{6}$$

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, cf. par exemple Laveault et Gregoire, 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 [1].

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 :

$$\alpha = \frac{J\bar{r}}{1 + (J-1)\bar{r}} \tag{7}$$

où $\bar{r}$ 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 [2].

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 l'échouent. En effet, on peut montrer que

$$r_{bis-point}(j) = corr(Y^j, S) = \frac{\bar{S}_{(j1)} - \bar{S}_{(j0)}}{\sigma_S}\sqrt{p_j(1-p_j)} \tag{8}$$

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

1. 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.

2. 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 longueur 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).

3.1.2 Analyse factorielle des items

L'analyse factorielle permet d'étudier la structure des données et, plus particulièrement, la structure des corrélations entre les variables observées (ou manifestes) [3]. Il s'agit d'identifier les différentes dimensions sous-jacentes aux réussites observées et surtout d'évaluer le poids de la dimension principale, dans la mesure où c'est une optique unidimensionnelle qui sera envisagée lors de la modélisation.

Dans le cas où les items sont dichotomiques, la matrice des corrélations entre items est en fait la matrice des coefficients ϕ, qui sont bornés selon les taux de réussite aux items (Rocher, 1999). Une analyse factorielle basée sur cette matrice peut donc montrer quelques faiblesses : des facteurs « artefactuels » sont susceptibles d'apparaître, en lien avec le niveau de difficulté des items et non avec les dimensions auxquelles ils se rapportent. De plus, d'un point de vue théorique, certaines hypothèses utiles pour l'estimation, comme la normalité des variables, ne sont pas envisageables.

L'optique retenue est alors de se ramener à un modèle linéaire : les variables observées catégorielles sont considérées comme la manifestation de variables latentes continues. Les réponses à un item dichotomique sont définies de la manière suivante :

$$y_{ij} = \begin{cases} 0 & \text{si } z_{ij} \leq \tau_j \\ 1 & \text{si } z_{ij} > \tau_j \end{cases} \tag{9}$$

La réponse y_{ij} de l'élève i à l'item j est incorrecte tant que la variable latente Z_j reste en deça d'un certain seuil τ_j, qui dépend de l'item. Au-delà de ce seuil, la réponse est correcte.

L'analyse factorielle des items consiste donc en une analyse factorielle linéaire sur les variables continues Z_j. Deux modèles sont donc considérés. D'une part, une variable latente continue et conditionnant la réponse à l'item est fonction linéaire de facteurs communs et d'un facteur spécifique. D'autre part, un modèle de seuil représente la relation non linéaire entre la variable latente et la réponse

3. Notons qu'il s'agit ici d'analyse factorielle en facteurs communs et spécifiques et non d'analyse factorielle géométrique de type ACP ou ACM (pour des détails, consulter Rocher, 2013)

à l'item. Ce procédé permet de se ramener à une analyse factorielle linéaire, à la différence que les variables Z_j ne sont pas connues. Il s'agit donc d'estimer la matrice de corrélation de ces variables, sous certaines hypothèses.

Considérons le lien entre deux items j et k. Si les variables latentes correspondantes Z^j et Z^k sont distribuées selon une loi normale bivariée, il est possible d'estimer le coefficient de corrélation linéaire de ces deux variables à partir du tableau croisant les deux items. C'est le coefficient de corrélation tétrachorique – ou polychorique dans le cas d'items polytomiques. L'estimation de ce coefficient par le maximum de vraisemblance requiert la résolution d'une double intégrale (pour les détails de l'estimation pour deux items dichotomiques, cf. Rocher, 1999). Pour plus de deux items, il devient difficile d'estimer de la même manière les coefficients de corrélation à partir de la distribution conjointe des items qui est une loi normale multivariée. C'est pourquoi les coefficients de corrélation tétrachorique sont estimés séparément pour chaque couple d'items. Ce procédé a le désavantage de conduire à une matrice de covariances qui n'est pas nécessairement semi-définie positive, donc potentiellement non inversible.

3.2 Codage des réponses aux items

3.2.1 Valeurs manquantes

Trois types de valeurs manquantes sont distinguées :

- Valeurs manquantes structurelles : l'élève n'a pas vu l'item. C'est le cas pour les cahiers tournants, où les élèves ne voient pas tous les items. Dans ce cas, on considère l'item comme *non administré*, l'absence de réponse n'est alors pas considérée comme une erreur.
- Absence de réponse : l'élève a vu l'item mais n'y a pas répondu. L'absence de réponse est alors considérée comme une erreur de la part de l'élève.
- Non-réponse terminale : l'élève s'est arrêté au cours de l'épreuve, potentiellement en raison d'un manque de temps. Des choix sont effectués pour déterminer le traitement de ces valeurs. Nous considérons que si un élève a passé moins de 50 % d'une séquence, il n'a pas vu la séquence, les valeurs manquantes sont alors traitées de manière structurelle. Sinon, elles sont traitées comme des échecs.

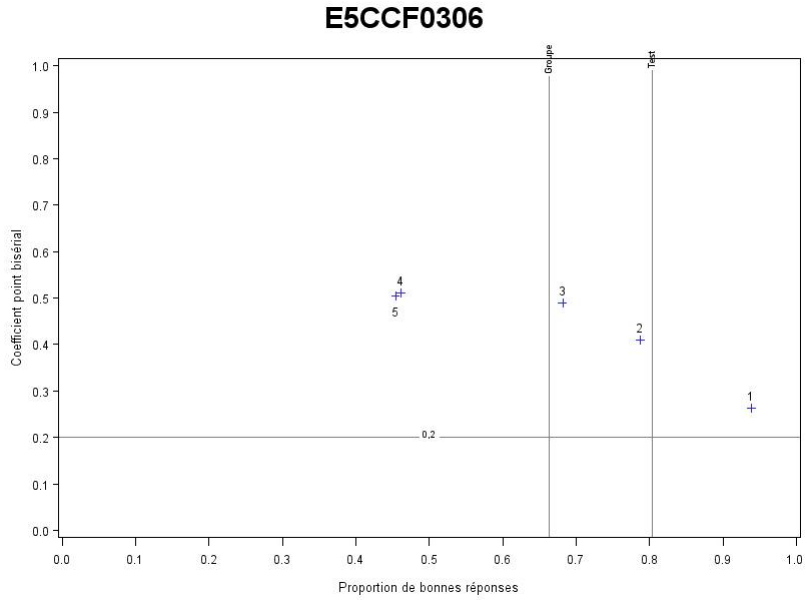
3.2.2 Regroupement des items

Les séries d'items comportant seulement deux réponses, comme les Vrai/Faux par exemple, font l'objet d'un traitement spécifique : les items de ce type sont

regroupés pour former un seul item à réponse binaire (réussite ou échec). En effet, la plus forte potentialité de réponse au hasard et l'inter-dépendance des items fragilisent leur utilisation individuelle.

Le regroupement de ces items consiste à faire la somme des indicatrices de réussite et à déterminer un seuil de maîtrise. Une visualisation graphique est utilisée pour fixer les scores « seuils » (cf. figure 1). Ce graphique représente le taux de réussite pour chaque seuil possible en fonction de la discrimination obtenu pour le seuil. Il permet de choisir la combinaison la mieux adaptée. Le score seuil doit préserver la discrimination de l'item regroupé et la difficulté peut être modulée en fonction des objectifs.

Figure 1 – Représentation graphique utilisée pour le regroupement d'items



Comme nous l'évoquerons, ce type de données pour être analysé de manière polytomique mais la modélisation considérée par la suite n'envisage pour l'instant que des items dichotomiques.

3.2.3 Traitement des données et correction des questions ouvertes

Tous les cahiers recueillis dans le cadre de cette opération ont été scannés par une société extérieure. Les réponses aux questions à choix multiples ainsi que les grilles d'évaluation remplies par les professeurs lors des séquences de travaux pratiques ont été numérisées et les codes de réponses stockés dans un fichier.

En ce qui concerne les questions ouvertes, demandant une rédaction plus ou moins longue de la part des élèves (explication, schématisation...), elles ont été découpées en « imagettes » puis transmises au ministère afin d'être intégrées dans un logiciel de correction à distance (cf. encadré « AGATE »). Celui-ci nécessite la formation technique des correcteurs et l'élaboration d'un cahier des charges strict de corrections pour limiter la subjectivité des corrections. Une fois la correction terminée, les codes saisis par les correcteurs ont été stockés dans un fichier puis associés à ceux issus des réponses aux QCM.

AGATE : un outil de correction à distance des questions ouvertes

Objectifs

Le logiciel AGATE, qui a été développé par les informaticiens de la DEPP, permet une correction à distance des questions ouvertes. Le principe général du logiciel est de soumettre un lot d'imagettes (image scannée de la réponse d'un élève) à un groupe de correcteurs tout en paramétrant des contraintes de double correction et/ou d'auto-correction. Lorsque deux correcteurs corrigent la même imagette, il arrive parfois qu'il y ait une différence de codage. Cette imagette est alors proposée au superviseur qui arbitre et valide l'un des deux codages. Ce jeu de codages multiples incrémente des compteurs (temps de connexion, avancement général et taux d'erreur) qui sont autant d'indicateurs pour suivre la correction. A noter qu'un processus de déconnexion automatique d'un correcteur existe si le superviseur se rend compte d'un trop grand nombre d'erreurs de correction. Ce logiciel est utilisé depuis 2004 par le bureau des évaluations de la DEPP. Il a permis d'intégrer des questions ouvertes dans des évaluations à grandes échelles, aussi bien aux évaluations nationales qu'aux évaluations internationales telles PISA, TIMSS ou PIRLS. Les correcteurs n'ont plus à manipuler un nombre très important de cahiers et peuvent travailler de manière autonome lorsqu'ils le souhaitent, tout en maintenant un contact entre eux et les responsables de l'évaluation afin d'assurer une meilleure fiabilité de la correction

Principes fonctionnels

Le chef de projet paramètre la session de correction. Il définit les groupes de correcteurs et supervise chaque groupe. Il intègre et vérifie les items mis en correction et ajuste les paramètres de double correction. Son rôle consiste également à répondre aux questions des correcteurs par le biais d'une messagerie intégrée au logiciel et à communiquer sa réponse également aux autres correcteurs. Le superviseur gère son groupe de correcteurs.

Il anime la session de formation, qui consiste d'une part à communiquer aux télécorrecteurs une grille de correction très précises et d'autre part à corriger collectivement à blanc un nombre défini d'imagettes pour s'assurer de la compréhension et de la bonne mise en oeuvre des consignes. Puis, pendant la télécorrection, il arbitre les litiges lors des doubles-corrections. Le correcteur corrige les items en portant un codage de réussite/erreur sur chaque item. En cas de doute, il peut se référer à son superviseur de groupe. Une messagerie interne complète le dispositif et permet un échange de point de vue entre les différents acteurs.

3.3 Résultats

3.3.1 Pouvoir discriminant des items

Rappelons la répartition des items pour CEDRE Sciences 3e :
– 122 items communs aux évaluations 2007 et 2013 ;
– 109 items 2013 ;
– 43 items 2007 non repris en 2013.

Nous avons éliminé 37 items en raison d'un faible indice *rbis-point* :
– 9 items communs ;
– 28 items 2013 ;
Notons que lorsque l'on calcule les indices de discrimination sur l'ensemble des items, de nombreux items apparaissent faiblement discriminant. En effet, la dimensionnalité des items a évolué entre les deux années (cf. paragraphe suivant). La modélisation s'appuiera d'ailleurs sur les seuls items communs à 2007 et 2013.

3.3.2 Dimensionnalité

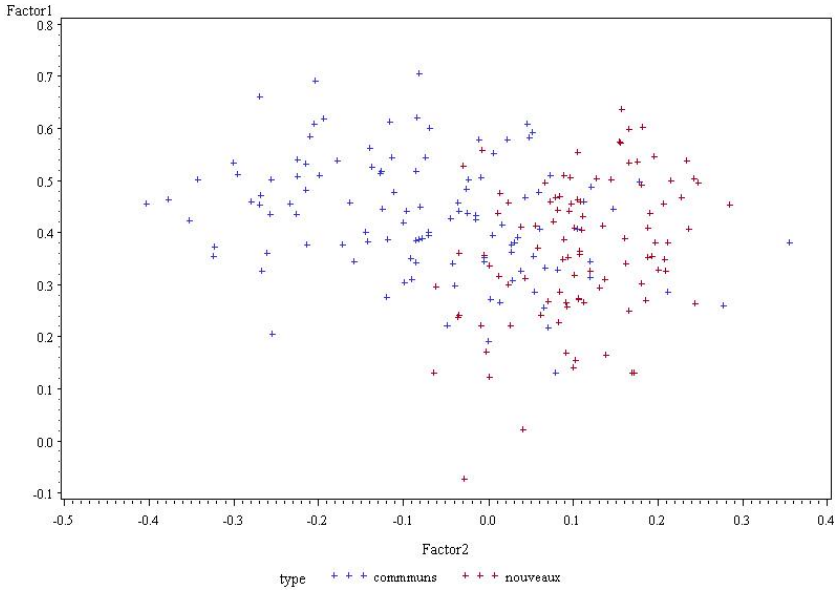
Le tableau 17 présente les résultats de l'analyse factorielle des items effectuée sur l'année 2013.

Tableau 17 – Analyse en composantes principales

	Valeur Propre	Difference	Proportion	Proportion cumulee
1	32.9	29.4	0.223	0.223
2	3.6	0.30	0.024	0.247
3	3.3	0.32	0.022	0.269

Malgré une structure fortement unidimensionnelle, il apparaît une différentiation entre les items nouveaux de 2013 et les items repris de 2007, comme le montre la figure 2 qui représente les items de 2013 sur le premier plan de l'analyse factorielle. Ce phénomène est à relier à l'évolution du cadre d'évaluation lié aux programmes de 2008 (cf. première partie du rapport).

Figure 2 – Premier plan factoriel des items de 2013



4 Modélisation

4.1 Méthodologie

4.1.1 Modèle de réponse à l'item

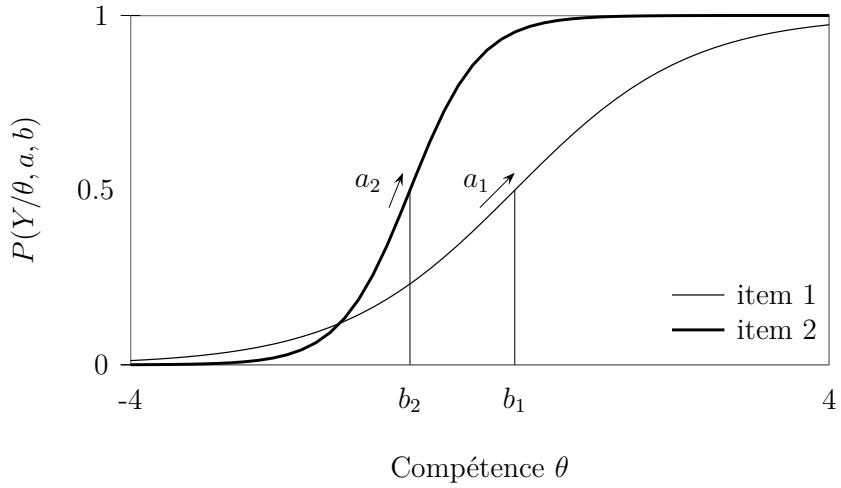
Le modèle de mesure utilisé est un modèle de réponse à l'item à deux paramètres avec une fonction de lien logistique (MRI 2PL) :

$$P_{ij} = P(Y_i^j = 1 \,|\, \theta_i, a_j, b_j) = \frac{e^{1,7a_j(\theta_i - b_j)}}{1 + e^{1,7a_j(\theta_i - b_j)}} \tag{10}$$

où la probabilité P_{ij} que l'élève i réussisse l'item j est fonction du niveau de compétence θ_i de l'élève i, du niveau de difficulté b_j de l'item j, ainsi que de la discrimination de l'item a_j ($a_j > 0$). La constante 1,7 est introduite pour rapprocher la fonction sigmoïde de la fonction de répartition de la loi normale.

La figure 3 représente les courbes caractéristiques de deux items selon cette modélisation.

Figure 3 – 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 ($a_1 < a_2$).

L'avantage de ce type de modélisation, c'est de séparer deux concepts-clé, à savoir la difficulté de l'item et le niveau de compétence de l'élève. Les MRI

ont un intérêt pratique pour la construction de tests et la comparaison entre différents groupes d'élèves : 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.

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 3).

4.1.2 Procédures d'estimation

L'estimation est 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 ici des éléments concernant ces procédures.

Estimation des paramètres des items

Nous reprenons les notations de l'équation (10) qui formule la probabilité P_{ij} d'un élève i de répondre correctement à un item j dans le cadre d'un modèle de réponse à l'item, avec les items sont dichotomiques.

Notons tout d'abord que les modèles présentés ne sont pas identifiables. En effet, 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. Dans CEDRE, nous levons l'indétermination en standardisant la distribution des θ pour les données du premier cycle (en l'occurrence, moyenne de 250 et écart-type de 50 poour l'année 2007).

Sous l'hypothèse d'indépendance locale des items, la fonction de vraisemblance s'écrit :

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

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

La procédure MML (*Marginal Maximum Likelihood*) est utilisées. Elle consiste à estimer les paramètres des items en supposant que les paramètres des individus

sont issus d'une distribution fixée *a prori* (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* $\mathbf{y}_i$ donné peut s'écrire :

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

avec g la densité de θ.

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

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

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. Pour plus de détails, le lecteur est invité à consulter Rocher (2013).

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.

Estimation des niveaux de compétence

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 (11).

Cependant, l'estimateur du maximum de vraisemblance, noté $\theta_i^{(ML)}$, est biaisé : les propriétés classiques de l'estimateur selon la méthode du maximum de vraisemblance ne sont pas vérifiées puisque le nombre de paramètres augmente avec le nombre d'observations. Ce biais vaut :

$$B(\theta_i^{(ML)}) = \frac{-J}{2I^2} \tag{14}$$

avec

$I = \sum_{j=1}^{J} \frac{P_{ij}^{'2}}{P_{ij}(1-P_{ij})} = \sum_{j=1}^{J} a_j^2 P_{ij}(1-P_{ij})$

et

$J = \sum_{j=1}^{J} \frac{P_{ij}^{'} P_{ij}^{''}}{P_{ij}(1-P_{ij})} = \sum_{j=1}^{J} a_j^3 P_{ij}(1-P_{ij})$

Pour obtenir un estimateur non biaisé, Warm (1989) a proposé de maximiser une vraisemblance pondérée $w(\theta)L(\mathbf{y}, \mathbf{a}, \mathbf{b}, \theta)$, en choisissant $w(\theta)$ de manière à ce que l'annulation de la dérivée du logarithme de la vraisemblance pondérée revienne à résoudre l'équation suivante :

$$\frac{\partial lnL}{\partial \theta_i} + \frac{J}{2I} = 0 \tag{15}$$

4.1.3 Indice d'ajustement (FIT)

L'ajustement des items au modèle est étudié. Graphiquement, cela revient à comparer les courbes caractéristiques estimées avec les résultats observés (cf. figure 4). Certaines procédures proposent de comparer directement les probabilités théorique avec les proportions de réussite de groupes d'élèves. Plus généralement, nous pouvons écrire les résidus de la manière suivante :

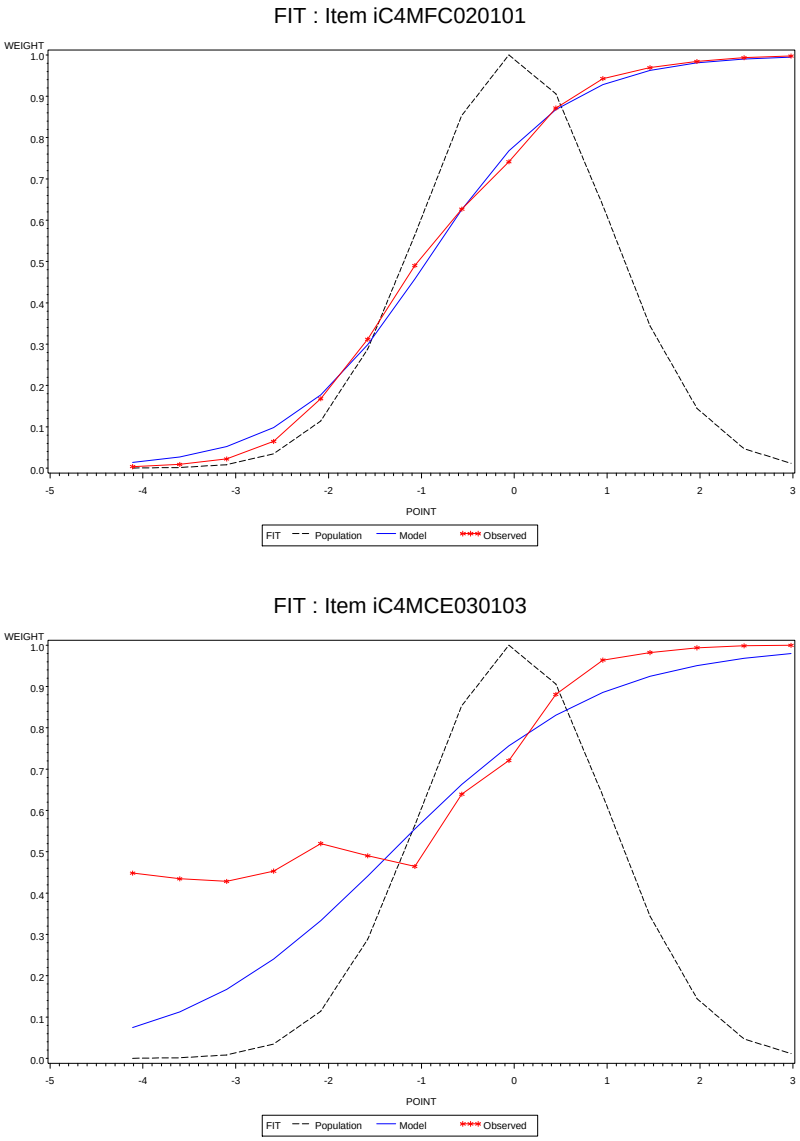
$$z_{ij} = \frac{Y_i^j - P_{ij}}{\sqrt{P_{ij}(1-P_{ij})}} \tag{16}$$

Les carrés des résidus suivent typiquement une loi du χ^2. L'indice *Infit* d'un item correspond à la moyenne pondérée des carrés des résidus, qui peut s'écrire :

$$Infit_j = \frac{1}{\sum_{i=1}^{n} w_{ij}} \sum_{i=1}^{n} w_{ij} z_{ij}^2 = \frac{1}{\sum_{i=1}^{n} w_{ij}} \sum_{i=1}^{n} (Y_i^j - P_{ij})^2 \tag{17}$$

avec le poids $w_{ij} = P_{ij}(1-P_{ij})$. Une transformation de cet indice est utilisé de manière à obtenir une statistique suivant approximativement une loi normale (Smith, Schumaker, & Bush, 1998).

Figure 4 – Exemples d'ajustements (FIT)



Note de lecture : La courbe bleue représente la courbre caractéristique de l'item telle qu'estimée par le modèle. La courbe en rouge relie des points qui correspondent aux taux de réussite observé à cet item pour 15 groupes d'élèves de niveaux de compétence croissants. Enfin, la courbe en pointillée représente la distribution des niveaux de compétence.
Clairement, l'ajustement du modèle est excellent pour l'item présenté à gauche. Il est très mauvais pour celui de droite.

4.1.4 Fonctionnement Différentiel d'Item (FDI)

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.

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 :

$$P(Y \mid Z, G) \neq P(Y \mid Z) \tag{18}$$

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 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. Plus précisément, les paramètres des items sont estimés séparément pour les deux années, puis ajustés en tenant compte de la différence moyenne entre les deux séries de paramètres. La règle retenue pour identifier un FDI est celle d'un écart de paramètres de difficulté β d'au moins 0,5 (cf. Rocher, 2013 pour plus de détails).

4.1.5 L'information du test

Dans le cadre d'un modèle de réponse à l'item à deux paramètres, l'information d'un item j est définie par :

$$I_j(\theta) = (1,7a_j)^2 P_j(\theta)(1 - P_j(\theta)) \tag{19}$$

avec $P_j(\theta)$, la probabilité de réussite à l'item pour individu de compétence θ.

L'information moyenne du test pour un élève de compétence θ est la somme de l'information apporté par chaque item pour θ. La courbe d'information d'information du test est tracée pour un ensemble de valeurs de θ.

L'erreur de mesure étant inversement proportionnelle à l'information, cette courbe d'information permet de visualiser la précision avec laquelle le niveau de compétence θ des élèves est estimé.

4.2 Résultats

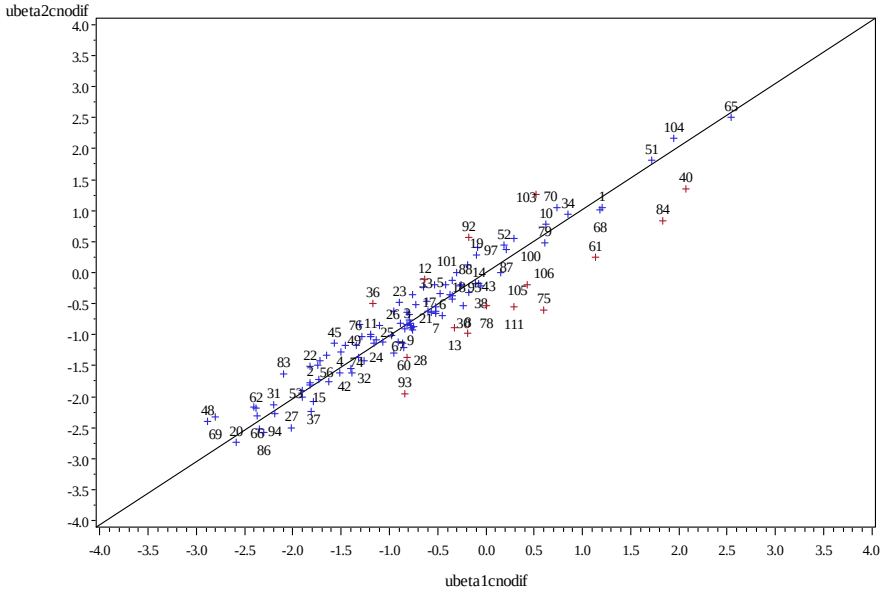
4.2.1 Identification des fonctionnements différentiels d'items (FDI)

L'analyse des FDI a permis de détecter 15 items : 11 items en faveur de 2013, 4 items en faveur de 2007 (figure 5). 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 3e.

4.2.2 Identification des items présentant un mauvais ajustement (FIT)

Aucun item n'a été supprimé pour cause de mauvais FIT.

Figure 5 – Comparaison des paramètres de difficulté 2007-2013



Note de lecture : Les points sont les items. En abscisse figure la valeur des paramètres de difficulté estimés en 2007, et en ordonnée la la valeur des paramètres de difficulté estimés et ajustés pour l'année 2013.

4.2.3 Bilan de l'analyse des items

Au départ, il y avait :
– 122 items communs
– 109 items de 2013
Nous n'incluons pas les items de 2007 non repris dans l'analyse.

Après suppression des items ayant un *rbis-point* inférieur à 0,2 et des items présentant un fonctionnement différentiel ou un mauvais ajustement, il reste :
– 98 items communs
– 81 items de 2013

4.3 Calcul des scores

Comme indiqué précédemment, une analyse conjointe des données (2007 et 2013), uniquement sur les items communs aux deux années, ce qui a permis d'estimer les paramètres des items, puis les niveaux de compétences θ des élèves. Afin de lever l'indétermination du modèle, la moyenne des θ a été fixé à 250

et leur écart-type à 50, pour l'échantillon de 2007. Le tableau 18 présente les résultats obtenus.

Tableau 18 – Niveaux de compétences (moyenne et écart-type)

annee	**N**	**Moyenne**	**Ecart-Type**
2007	4128	250.0	50.0
2013	5598	249.2	51.3

5 Construction de l'échelle

5.1 Méthode

Les modèles de réponse à l'item permettent de positionner sur une même échelle les paramètres de difficulté des items et les niveaux de compétences des élèves. Cette correspondance permet de caractériser les compétences maîtrisées pour différents groupes d'élèves.

Les scores en sciences estimés selon le modèle de réponse à l'item présenté dans la partie précédente ont été standardisés de manière à obtenir une moyenne de 250 et un écart-type de 50 pour l'année 2007. Puis, comme le montre la figure 6, la distribution des scores est « découpée » en six groupes de la manière suivante : nous déterminons le score-seuil en-deça 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.

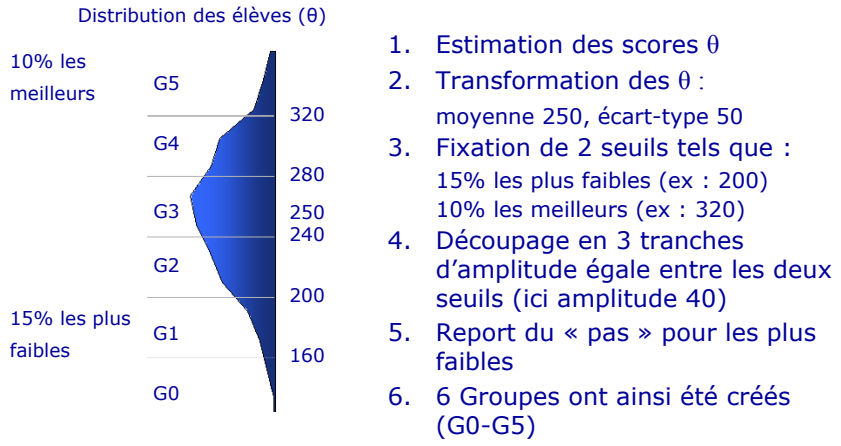
5.2 Caractérisation des groupes de niveaux

A 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 (Andreu, Etève, & Garcia, 2014).

Groupe < 1 (2,4 % des élèves)

Si les élèves de ce groupe sont en mesure de mobiliser des acquis, ces capacités ne s'exercent que sur une partie (2/3) des domaines abordés en sciences à l'école élémentaire. Ils n'apportent aucune réponse aux questions posées dans les domaines du fonctionnement du vivant, de l'unité et de la diversité du vivant ainsi que de l'énergie. La moitié des réussites de ce groupe concerne des notions relevant des programmes de cycle 2. Dans la majorité des cas, leurs

Figure 6 – Principes de construction de l'échelle



réussites concernent des situations proches de la vie quotidienne. Ces élèves savent faire des prélèvements directs. Très majoritairement leurs connaissances se manifestent dans le domaine du corps humain.

Groupe 1 (13,3 % des élèves)

Les élèves du groupe 1 réussissent en moyenne 38 % des items. Ces élèves réussissent 5 items de plus que les élèves du groupe précédent. Ces items concernent essentiellement des connaissances notionnelles et lexicales relatives au corps humain.

Groupe 2 (28,4 % des élèves)

Les élèves du groupe 2 maîtrisent 56 % des items. Les élèves du groupe 2 ont des connaissances et mobilisent des compétences dans tous les domaines des sciences à l'école élémentaire.
Ils sont capables de comprendre, lire et interpréter des phénomènes dynamiques (décrire un processus à partir d'une représentation figée) : traitement de l'eau, rythme cardiaque et chaînes alimentaires. Ils savent prélever une information donnée sur des supports différents : tableaux, graphique, dessins, schémas. Le groupe 2 est notamment caractérisé par une meilleure maîtrise du lexique dans des situations de questionnement direct (connaissance déclarative : squelette, articulations, expiration, inspiration, circulation etc.). ls savent choisir le connecteur logique adapté pour déterminer une cause ou une connaissance. Ils sont capables de mobiliser leurs compétences pour effectuer des opérations intellectuelles simples (premier niveau de raisonnement : associer, reconnaître, reconstituer, caractériser, distinguer) à partir de situations représentées (schéma, gra-

phique, dessin, photo, tableau). Ils sont capables de déterminer si une situation présente un danger électrique.

Groupe 3 (29,0 % des élèves)

Les élèves du groupe 3 maîtrisent les 2/3 des connaissances attendues en fin de cycle 3. Le lexique spécifique est bien connu dans tous les domaines évalués. Les principales notions acquises qui relèvent du vivant portent essentiellement sur les fonctions de nutrition : digestion, respiration, et sur le fonctionnement du vivant : mode de reproduction. En ce qui concerne les objets techniques, ils maîtrisent les notions de circuit, d'interrupteur, et connaissent les concepts d'« isolant » et de « conducteur ». Par ailleurs, les élèves de ce groupe sont capables de schématiser et de modéliser des circuits (ouverts et fermés).
Ces élèves réussissent presque les trois quarts (72,9 %) des situations nécessitant un raisonnement. Ils sont capables de lire et d'interpréter des supports variés : schémas, tableaux à double entrée, graphiques, radiographies, dessins, texte documentaire illustré, photographies. Ils parviennent également à comparer, déduire et à établir des relations de causes à effets. Ils ont acquis une première connaissance des étapes de la démarche d'investigation. Sur des situations complexes, à partir de documents, ils sont capables de répondre à des questions de prélèvement et de raisonnement. Ils savent identifier le protocole expérimental adapté et choisir la conclusion d'une expérience donnée.

Groupe 4 (17,1 % des élèves)

Les élèves du groupe 4 réussissent près de 80 % des items, ils manifestent des compétences solides en sciences expérimentales. Les situations nécessitant un raisonnement sont réussies à 82,2 %. Les compétences de lecture, d'interprétation et de compréhension sont bien activées sur des supports complexes. Ils parviennent à lire et à comprendre un document, mais également à le légender. Les élèves prélèvent aisément des informations implicites pour les mettre en relation avec leurs connaissances et les réinvestir à bon escient. Ils parviennent à faire des inférences. Ils font preuve de capacités d'abstraction pour se construire des représentations mentales. D'autre part, ils peuvent prendre appui sur leurs acquis pour rédiger des réponses construites sur des situations complexes.

Groupe 5 (9,9 % des élèves)

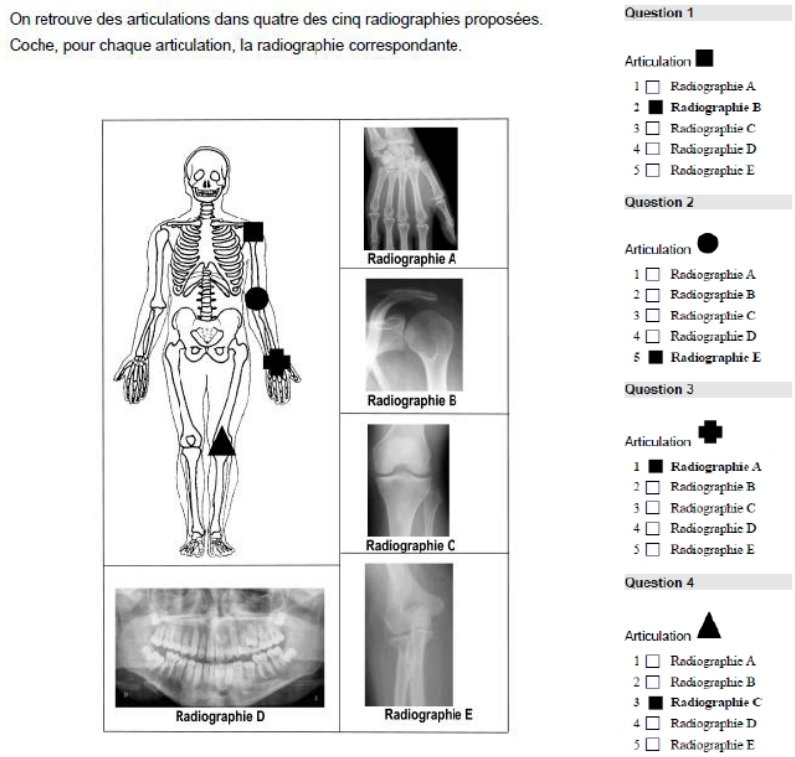
Ces élèves ont des connaissances approfondies (91 %) dans l'ensemble des domaines. Leurs performances par rapport à la démarche d'investigation les distinguent des élèves des autres groupes. Ils savent inférer les informations contenues dans un texte documentaire (présentée sous diverses formes : texte et iconographie) pour identifier un dispositif expérimental équivalent représentant la situation (modélisation). Ils savent associer une modélisation à une situation réelle décrite sous forme d'un texte et d'une représentation géographique. Ils sont

capables de concevoir et schématiser un protocole pour répondre à une question initiale donnée en s'appuyant sur une liste de matériel suggéré. Ils savent tracer un graphique à partir de données, interpréter un graphique en inférant le complément des données qui sont directement représentées. Ils savent faire le lien entre une connaissance générale donnée dans un texte avec son application particulière dans des exemples (lien connaissances générales - connaissances spécifiques). Ils sont capables d'avoir une compréhension fine (lexicale et syntaxique) du texte pour recoder des informations non données directement (inférences). Ils savent associer un comportement décrit par un texte à l'action correspondante dans la gestion des déchets (réutilisation - recyclage - réutilisation). Ils savent prélever dans un schéma des informations de mesures données indirectement pour résoudre un problème. En résumé, les élèves maîtrisent toutes les compétences nécessaires pour mettre en œuvre une démarche d'investigation. À noter le bon niveau d'acquisition des habiletés en lecture compréhension : capacité à mettre en relation des informations données sous différentes formes.

5.3 Exemples d'items

5.3.1 Item caractéristique du groupe < 1

Cet exercice fait partie du domaine fonctionnement du corps humain et de la santé. Les emplacements de certaines articulations, l'épaule, le coude, le poignet et le genou sont repérés sur un dessin de squelette humain. Les élèves doivent retrouver la radiographie correspondant à la bonne articulation.



Cet exercice est réussi à près de 96 % dans son ensemble. Deux propositions présentent un score supérieur à 90 % (épaule et poignet), alors que les deux autres se situent autour de 58 % (coude, genou). Plusieurs items sont nécessaires à valider la bonne réponse. Le seuil défini ici est de 2 sur 4, c'est-à-dire qu'il faut indiquer 2 bonnes réponses sur 4 pour valider l'item dans son entier. L'identification de la main pour l'articulation du poignet, ainsi que de la clavicule pour l'articulation de l'épaule sont lisibles. En revanche la baisse des scores pour les articulations du genou et du coude est vraisemblablement liée à la difficulté d'identification sur la radio.

5.3.2 Items caractéristiques du groupe 1

5.3.2.a Exemple 1

Dans cet item sur le fonctionnement du corps humain, on évalue les connaissances lexicales sur les articulations du bras. À la différence de l'exercice présenté pour le groupe inférieur à 1, ici il n'y a pas de document iconographique. Des noms d'articulations sont proposés aux élèves dans un tableau vrai / faux à double entrée. Il leur faut indiquer s'il s'agit d'une articulation du bras ou non. Cet exercice est réussi à 77 % dans son ensemble alors que les réussites, item par item, varient de 88 (cou) à 95 % (coude). Le seuil défini pour ce multi-items est exhaustif, il convient d'indiquer 6 bonnes réponses sur 6 pour valider la question dans son entier. Ce qui explique la différence significative de score entre les réussites item par item et la réussite globale à l'exercice. On constate sur l'ensemble des exercices liés à l'observation des mouvements corporels et du squelette des taux de réussite très satisfaisants. Cet objet d'étude, lié aux activités physiques et sportives vécues par les élèves, semble abordé dans une majorité de classes.

Les noms suivants correspondent-ils à des articulations du bras ?

	Vrai	Faux
cheville	☐ 1	■ 2
cou	☐ 1	■ 2
coude	■ 1	☐ 2
épaule	■ 1	☐ 2
genou	☐ 1	■ 2
poignet	■ 1	☐ 2

5.3.2.b Exemple 2

Le deuxième exercice présenté pour le groupe 1 relève du domaine le fonctionnement du vivant et il évalue des connaissances lexicales sur les changements d'un être vivant au cours du temps. Les élèves doivent choisir entre 4 termes pour définir l'augmentation irréversible de la taille et du poids. Cet item est réussi à presque 82 %, ce qui tend à indiquer que le lexique qui définit l'augmentation irréversible du poids et de la taille est bien connu.

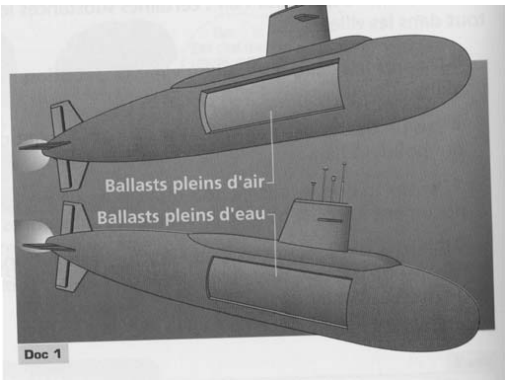
Chez un être vivant, l'augmentation irréversible de la taille et du poids s'appelle ...

1 ☐ la métamorphose.
2 ■ **la croissance.**
3 ☐ la naissance.
4 ☐ le vieillissement.

5.3.3 Item caractéristique du groupe 2

L'item concerne le domaine de l'air. La situation permet d'évaluer la connaissance du caractère pesant de l'air. À partir d'un dessin légendé qui décrit le fonctionnement des ballasts, les élèves doivent choisir parmi les explications proposées, celle qui leur parait la plus pertinente pour justifier le fait que le sous-marin remonte à la surface lorsque les ballasts sont pleins d'air.

Lis le document suivant.



Un sous-marin comporte de nombreux systèmes techniques plus ou moins compliqués. L'un d'eux lui permet de plonger en profondeur, puis de remonter en surface.
C'est possible grâce aux ballasts. Ce sont deux gros réservoirs qui entourent les flancs du sous-marin (doc 1).
Lorsqu'ils sont remplis d'air, le sous-marin est en surface.
Lorsqu'ils sont pleins d'eau, le sous-marin est en plongée.

Pourquoi le sous-marin remonte-t-il vers la surface lorsque les ballasts sont remplis d'air ? **Groupe 2**

Choisis la bonne explication parmi les propositions suivantes.

Pour un même volume ...

1 ☐ l'air est plus froid que l'eau.
2 ☐ l'air est plus lourd que l'eau.
3 ☐ l'air est plus chaud que l'eau.
4 ■ l'air est plus léger que l'eau.

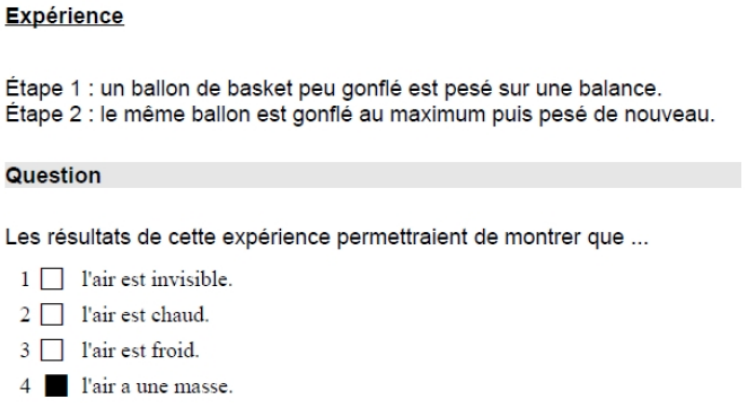
E5XME1460301

5.3.4 Item caractéristique du groupe 3

Il s'agit ici d'un item qui relève du domaine de l'air. L'évaluation porte sur la connaissance du caractère pesant de l'air, sa capacité à être manipulé et transporté, et sur la compréhension d'une expérimentation qui permet de le démon-

trer. Dans cet item, les étapes d'une expérience sont exposées aux élèves : 1-On pèse un ballon peu gonflé sur une balance. 2-On pèse de nouveau ce ballon qui, cette fois, est gonflé au maximum. Quatre hypothèses de ce que cette expérience permet de démontrer sont proposées aux élèves :
- L'air est invisible.
- L'air est chaud.
- L'air est froid.
- L'air a une masse.

Plus d'un élève sur 2 est capable de faire le lien entre le protocole expérimental et le but recherché. Environ 1 élève sur 4 (distracteur le plus fort) choisit une explication qui renvoie à une connaissance de la vie courante (l'air est invisible) qui n'a rien à voir avec le protocole.


Expérience

Étape 1 : un ballon de basket peu gonflé est pesé sur une balance.
Étape 2 : le même ballon est gonflé au maximum puis pesé de nouveau.

Question

Les résultats de cette expérience permettraient de montrer que ...

1 ☐ l'air est invisible.
2 ☐ l'air est chaud.
3 ☐ l'air est froid.
4 ■ l'air a une masse.

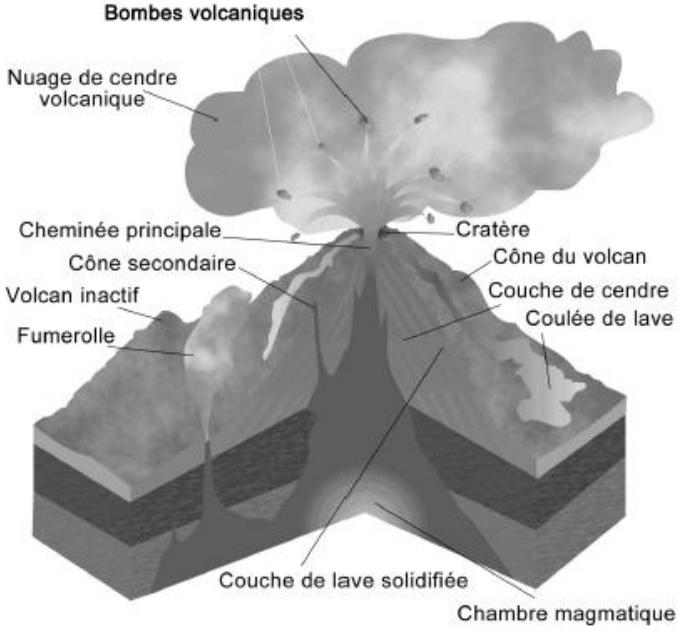
5.3.5 Item caractéristique du groupe 4

Cet item représentatif du groupe 4 porte sur le domaine le ciel et la terre (volcans et séismes). Dans cette situation, les élèves doivent lire un document scientifique : la représentation en coupe partielle et légendée d'un volcan en activité. Ils doivent ensuite compléter un texte de closure (texte « à trous ») en utilisant les mots clés de la légende.

Le taux global de réussite est de 51 % ; il correspond à 6 propositions correctes sur 9 (des tolérances ont été intégrées lors de la correction : cône du volcan / volcan, cratère / nuage de cendres / cendres volcaniques, lave solidifiée / cendres). Quand on observe les taux de réussite item par item, on repère globalement 2 catégories :
-un ensemble compris entre 50 et 60 % de taux de réussite (6 items) ce sont des items nécessitant un prélèvement direct ;
-et un deuxième groupe de taux de bonnes réponses entre 25 et 40 % (3 items),

pour ces items, il est nécessaire de prendre en compte les indices grammaticaux pour s'assurer du choix du bon mot ou groupe de mots. Le taux de non réponse augmente dans ce cas (jusqu'à 1 élève sur 4).

Lis le document suivant.



Les [Nuage de cendre volcanique] et le [bombes volcaniques] s'envolent dans le ciel au dessus du [Cratère].

Une [Coulée de lave] s'écoule sur le [Cône du volcan].

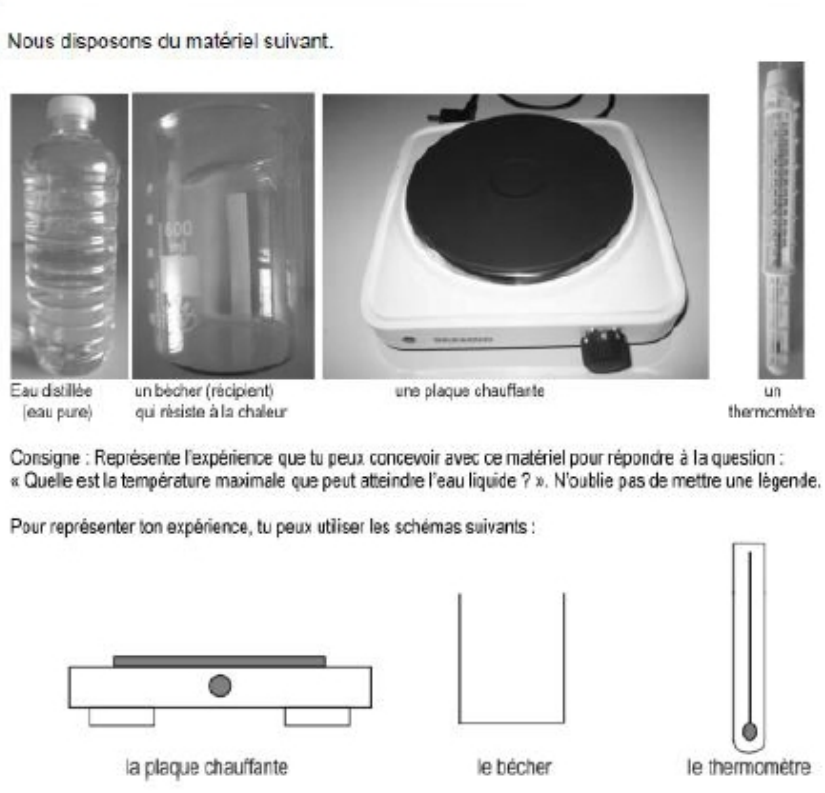
[Cratère] est situé au sommet de la [cheminée principale].

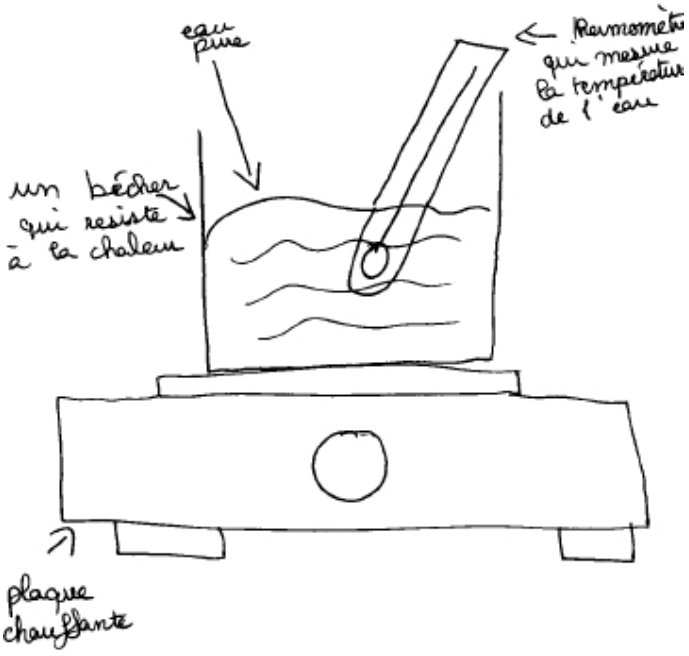
Le cône volcanique est composé de couches de [Cendre]. L'ensemble du cône volcanique est recouvert par de la [De lave solidifiée].

5.3.6 Items caractéristiques du groupe 5

5.3.6.a Exemple 1

Il s'agit là d'un item issu de la situation complexe « Les états de l'eau ». Cet item constitue la deuxième question de la situation, il s'agit de concevoir et schématiser un protocole pour répondre à la question : « Quelle est la température maximale que peut atteindre l'eau liquide ? ». La légende est exigée. Une liste de matériel est proposée. On note 31 % de non réponse pour 34 % de bonnes réponses. Parmi ces réponses correctes, des nuances sont à apporter ; seuls 15 % des élèves parviennent à légender complètement le schéma et 19 % proposent un schéma sans légende ou avec une légende partielle. Dans les 35 % d'élèves qui ont produit une réponse incorrecte, 19 % réalisent un schéma dans lequel il manque un élément (par exemple, pas de thermomètre) ou sur lequel apparaît un élément non pertinent.





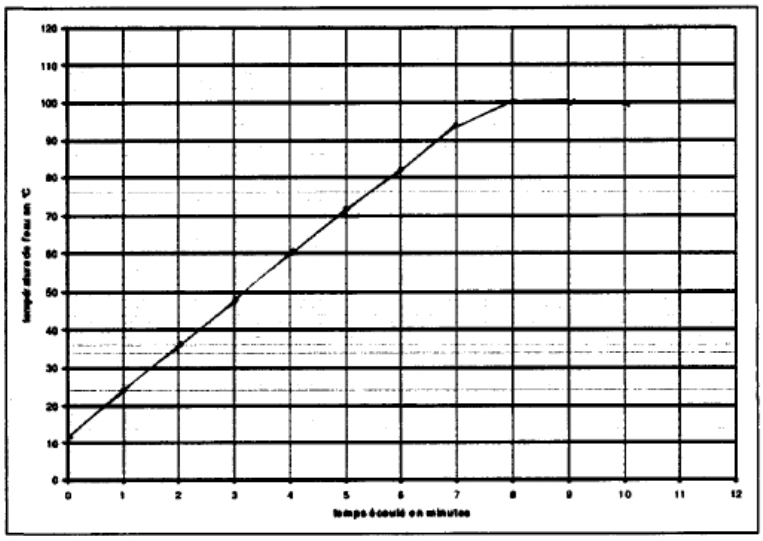
Dans cet item (situation « les états de l'eau »), à l'aide d'un tableau de mesures de températures, il faut construire un graphique pour exprimer les résultats de l'expérience ci-avant. Cet item met donc en jeu des compétences mathématiques. Les bonnes réponses s'élèvent à 20 % alors que la non-réponse atteint près de -18 %. Les réponses correctes sont ventilées de la manière suivante :
-20 % des élèves réalisent une représentation (courbe ou points) qui respecte les 11 graduations (de 0 à 10) pour les temps indiqués.
-1,5 % des élèves proposent une représentation qui respecte toutes les graduations jusqu'à la douzième minute (11min : 100°C, 12 min : 100°C). Ils savent que la température de l'eau se stabilise dès que commence l'ébullition.
-Pour 5 % des élèves, la représentation respecte 9 graduations (1 erreur ou omission).
-4 % des élèves respectent 9 graduations (2 erreurs ou omissions).
-2 % représentent toutes les graduations sous forme d'histogramme.

On réalise l'expérience et on obtient le tableau de mesures ci-dessous (document 1).

Minutes	0	1	2	3	4	5	6	7	8	9	10
Températures En °C (degrés Celsius)	12	24	36	48	60	72	82	94	100	100	100

Document 1

À l'aide du support ci-dessous, trace le graphique qui correspond au document 1.



5.3.6.b Exemple 2

L'exercice porte ici sur le mouvement de la terre autour du soleil dans le domaine le ciel et la terre. L'activité consiste à apparier les ombres d'un gnomon aux différentes heures de la journée. (Le gnomon est un instrument astronomique pour prendre la hauteur du soleil déterminée par la longueur de son ombre projetée sur une table généralement plane, la connaissance de ce lexique n'était pas nécessaire pour répondre à la question). Pour cela, ils doivent choisir la bonne proposition parmi 4 schémas fournis. Une rose des vents (avec le sud en haut et le nord en bas) est fournie. Une proposition est incohérente (les différentes heures se suivent dans un ordre impossible) et peut, de ce fait, être écartée sans mobiliser les connaissances nécessaires. Ce distracteur a quand même été choisi par près de 15 % des élèves.

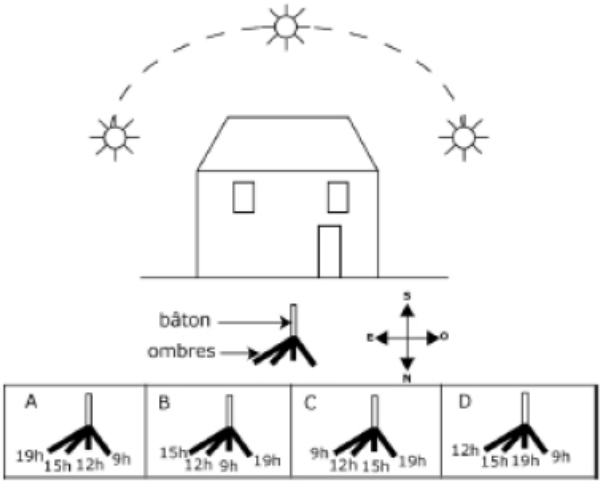
Les connaissances à mobiliser pour résoudre cette situation sont triples :
- le soleil se lève à l'est (et se couche à l'ouest) ;
- la lumière suit un trajet rectiligne ;

- l'ombre portée est située à l'opposé de la source de lumière (ici, le soleil).

La proposition qui obtient le plus de choix des élèves (37 %) est celle pour laquelle les heures sont organisées de gauche à droite (sens de lecture habituel).

Les élèves qui ont choisi la bonne réponse ont des connaissances robustes sur « ombres et lumières », « points cardinaux et boussoles » qu'ils sont capables de mobiliser conjointement. La connaissance indispensable pour résoudre la situation est que le soleil se lève à l'est et elle est à mobiliser sur une rose des vents avec le Sud en haut). Il faut également savoir que le trajet de la lumière est rectiligne.

Un enfant a dessiné l'ombre du bâton à quatre moments de la journée.



Question

Choisis la bonne représentation.

1 ■ A
2 ☐ B
3 ☐ C
4 ☐ D

6 Variables contextuelles et non cognitives

6.1 Variables sociodémographiques

Un certain nombre de variables sociodémographiques permettent d'enrichir l'analyse des résultats. Le score moyen des élèves est ainsi analysé en fonction du genre, du retard scolaire et quand les effectifs le permettent en fonction du secteur d'enseignement (tableau 19). Le lecteur est invité à consulter la Note d'Information pour plus de détails (Andreu et al., 2014). Malheureusement, nous ne disposons pas de la PCS des parents, comme c'est le cas pour la 3e.

Tableau 19 – Répartition (en %) et score moyen en sciences et répartition selon les groupes de niveaux en 2007 et en 2013

	annee	Répartition (en %)	Score Moyen	Ecart-Type
Ensemble	2007	100.0	250	50
Ensemble	2013	100.0	249	51
Garçons	2007	51.4	252	52
Garçons	2013	50.2	250	53
Filles	2007	48.6	248	48
Filles	2013	49.8	248	50
Elèves en retard	2007	15.8	215	37
Elèves en retard	2013	12.0	215	42
Elèves à l'heure	2007	84.2	257	49
Elèves à l'heure	2013	88.0	254	51

6.2 Variables conatives

Le questionnaire interroge également certaines dimensions dites « conatives », relevant d'aspects non cognitifs. Les items correspondants font d'abord l'objet d'une analyse factorielle exploratoire en facteurs corrélés permettant d'explorer la structure des items (Keskpaik, 2011). Les différentes dimensions sont validées puis un indice est calculé pour chacune d'entre elle, en considérant le premier axe d'une Analyse en Composantes Principales (ACP).

Le tableau 20 présente en guise d'illustration les items d'une de ces dimensions, en l'occurrence le sentiment d'efficacité en sciences.

Tableau 20 – Exemple de variable conative - sentiment d'efficacité en sciences

Question	**1er Axe ACP**
Je pense que j'ai un bon niveau en sciences	0.71
Je comprends bien ce que nous faisons en sciences	0.70
Je pense que je peux bien réussir en sciences	0.68
Les sciences, c'est trop difficile pour moi	0.65

Note de lecture : Les élèves devaient répondre à ces questions sur échelle dite de Lickert, de « Pas du tout d'accord » à « Tout à fait d'accord »

6.3 Motivation des élèves face à la situation d'évaluation

Les évaluations standardisées des élèves, telles que CEDRE ou PISA, renvoient à des enjeux politiques croissants, alors qu'elles restent à faible enjeu 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.

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 (figure 7). Les données recueillies 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 (Keskpaik. & Rocher, à paraître).

Figure 7 – Instrument de mesure de la motivation au test

[Q1]

Comment as-tu trouvé les exercices de cette évaluation ?

1☐ Très faciles
2☐ Faciles
3☐ Difficiles
4☐ Très difficiles

[Q2]

Es-tu d'accord avec ces affirmations ?

(Coche une case par ligne)

	Pas du tout d'accord	Pas d'accord	D'accord	Tout à fait d'accord
Je me suis bien appliqué(e) pour faire cette évaluation	☐ 1	☐ 2	☐ 3	☐ 4
Je me suis autant appliqué(e) à faire cette évaluation que le travail quotidien de classe	☐ 1	☐ 2	☐ 3	☐ 4

7 Annexe

Certification AFNOR pour les évaluations CEDRE

La DEPP est engagée dans un processus de certification. L'audit aura lieu en mars 2015.

Les finalités de la certification

Les finalités sont les suivantes :
- inscrire les processus d'évaluation dans une dynamique pérenne d'amélioration continue ;
- renforcer la prise en compte des attentes des usagers dans la formalisation des objectifs des évaluations et la restitution de leurs résultats ;
- faire reconnaître par une certification de service la qualité du service rendu et la continuité du respect des engagements pris.

Les enjeux pour la DEPP

Il y a deux enjeux forts pour la DEPP, l'un interne, l'autre externe :
- améliorer les processus de construction des instruments d'évaluation des acquis des élèves, fiabiliser ces processus par une démarche de contrôle-qualité ;
- valoriser l'enquête CEDRE comme un standard de qualité procédurale dans le domaine de l'évaluation.

Plus spécifiquement, le projet de certification des évaluations CEDRE est porteur d'enjeux pour la DEPP en termes de communication sur la validité scientifique, la sincérité, l'objectivité et la fiabilité des évaluations, ainsi que sur l'éthique et le professionnalisme des équipes.

La démarche qualité

Elle est fondée sur un référentiel élaboré sur mesure, selon une démarche officielle reconnue par les services publics et en lien avec les représentants des utilisateurs du service et les professionnels. La transparence vis-à-vis des usagers est assurée par la communication des résultats des enquêtes de satisfaction annuelles.

Les engagements de service

Le référentiel d'engagements comporte 18 engagements (cf. encadré page suivante).

Les engagements de service de la DEPP

Des objectifs clairs et partagés

Nous associons les parties intéressées à la définition de notre programme d'évaluation.

Nous formalisons dans un « cadre d'évaluation » les résultats attendus et les paramètres techniques de l'évaluation, ses délais et les limites associées aux moyens mis en œuvre.

Des évaluations fondées sur l'expertise pédagogique

Nous définissons avec les parties intéressées les acquis à évaluer et les mesurons en intégralité.

Nous mobilisons, tout au long de l'évaluation, un groupe expérimenté composé d'enseignants de terrain, de formateurs, d'inspecteurs et de chercheurs.

Tous nos items sont testés, analysés et validés avec le groupe expert avant d'être utilisés dans le cadre d'une évaluation.

Les meilleures pratiques méthodologiques et statistiques au service de l'objectivité

Afin de garantir l'application des meilleures méthodes statistiques, nous prenons en compte avec exigence les principes du « Code de bonnes pratiques de la statistique européenne ».

Nous tirons un échantillon représentatif garantissant le maximum de précision de mesure, à partir du plan de sondage défini dans le respect du « cadre d'évaluation ».

Nous garantissons l'objectivité et la qualité des données recueillies par la standardisation des processus d'administration et de correction des tests.

Une mesure fiable et des comparaisons temporelles pertinentes

Afin de garantir l'application des meilleures méthodes psychométriques, nous prenons en compte avec exigence les recommandations internationales sur l'utilisation des tests.

Nous analysons les réponses apportées par les élèves aux items afin d'en garantir la validité psychométrique.

Nous modélisons une échelle de compétences servant de référence et offrons des comparaisons temporelles fiables et lisibles.

Nous caractérisons les niveaux de cette échelle et déterminons avec le groupe expert les seuils de maîtrise des compétences évaluées, permettant de vous décrire en détail les performances des élèves.

Des analyses enrichies par des données de contexte

Nous systématisons le recueil d'informations standardisées relatives aux élèves et à leur environnement scolaire et social, dans le respect le plus strict des règles de confidentialité.

Nous éclairons les résultats de nos évaluations par la mise en relation des scores avec ces données.

Transparence des méthodes et partage des résultats

Nous publions et présentons les résultats de chacune de nos évaluations.

Nous mettons à disposition un rapport technique précisant les méthodes utilisées dans le cadre de l'évaluation.

Nous participons, dans le cadre de conventions collaboratives, à des analyses complémentaires des données que nous produisons.

Références

Andreu, S., Etève, Y., & Garcia, E. (2014). CEDRE 2013 - grande stabilité des acquis en sciences en fin d'école depuis 2007. *Note d'information*, *14.27*.

Ardilly, P. (2006). *Les techniques de sondage*. Technip.

Christine, M., & Rocher, T. (2012, janvier). 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. In *Journées de méthodologie statistique.* Paris.

Keskpaik, S. (2011). L'analyse factorielle exploratoire. *Document de travail - série Méthodes*, *M03*.

Keskpaik., S., & Rocher, T. (à paraître). La motivation des élèves français face à des évaluations à faibles enjeux. comment la mesurer ? son impact sur les réponses. *Education et formations*, *85-86*.

Rocher, T. (1999). *Psychométrie et théorie des sondages*. Mémoire de Master non publié, Université Paris VI.

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 non publiée, Université Paris-Ouest.

Sautory, O. (1993). La macro calmar. redressement d'un échantillon par calage sur marges. *Série des documents de travail de l'INSEE*, *Document F9310*.

Smith, R., Schumaker, R., & Bush, J. (1998). Using item mean squares to evaluate fit to the rasch model. *Journal of Outcome Measurement*, *2 n°1*, 66-78.

Tillé, Y. (2001). *Théorie des sondages. echantillonnage et estimation en populations finies. cours et exercices avec solution*. Paris : Dunod.

Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54 n°3*, 427-450.

Liste des tableaux

Table des figures