CEDRE

Cycle des Évaluations Disciplinaires Réalisées sur Échantillons

Rapport technique

Sciences expérimentales 2018

École

Auteurs :

Vanessa AUGÉ
Yann ETEVE
Marion LE CAM
Louis-Marie NINNIN
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, de la jeunesse et des sports

Septembre 2020

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 des évaluations internationales telles que PIRLS ou PISA. Ces programmes d'évaluations sont des outils d'observation des acquis des élèves pour le pilotage d'ensemble du système éducatif (Trosseille & Rocher, 2015). Les évaluations du CEDRE (Cycle d'Évaluations Disciplinaires Réalisées sur Échantillons) révèlent ainsi, 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 l'organisation des apprentissages, 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 sur des modélisations 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 du système éducatif.

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 n'ont pas valeur de délivrance de diplômes, ni d'examen de passage ou d'attestation de niveau ; elles donnent une photographie instantanée de ce que savent et savent faire les élèves à la fin d'un cursus scolaire. En ce sens, il s'agit bien d'un bilan. Destinées à être renouvelées périodiquement, ces évaluations-bilans permettent également de disposer d'un suivi de l'évolution des acquis des élèves dans le temps. Pour cette raison, les épreuves ne peuvent pas être totalement rendues publiques car, devant être en grande partie reprises lors des prochains cycles d'évaluation, elles ne doivent pas servir d'exercices dans les classes.

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 d'enseignants : elles informent sur les compétences et les connaissances des élèves à la fin d'un cursus ; elles éclairent sur l'attitude et la représentation des élèves à l'égard de la discipline ; elles interrogent les pratiques d'enseignement au regard des programmes ; 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 Connaissances et compétences visées

1.2.1 Les programmes de référence

Un nouveau programme d'enseignement entre en vigueur en septembre 2016 (B.O. spécial n°11 du 26 novembre 2015), avec notamment une refonte des cycles, puisque le cycle 3 couvre désormais les classes de CM1, CM2 et 6ème. Le programme de sciences est découpé en sept grandes compétences et quatre thèmes. Chaque thème est décliné en attendus. (tableau 2)

Chaque situation d'évaluation est indexée sur la grille de compétences définie en fonction du programme d'enseignement des sciences, mais aussi selon deux types de contextes (personnel ou scolaire) et trois domaines cognitifs tels que définis dans l'étude internationale TIMSS. Le domaine "Connaitre" aborde les faits, les concepts et les procédures que les élèves doivent connaitre. Il est un préalable pour s'engager dans des activités culturelles plus complexes. "Appliquer" se centre sur l'aptitude des élèves à appliquer les connaissances et la compréhension des concepts, pour résoudre des problèmes ou répondre à des questions dans des contextes d'enseignement et d'apprentissage relativement familiers. "Raisonner", le domaine le plus exigeant, impose aux élèves de prendre en compte des situations nouvelles, des contextes peu courants ou plus complexes, ou encore de mettre en jeu plusieurs approches, plusieurs étapes ou plusieurs stratégies.

Les finalités des sciences sont rappelées dans l'introduction des programmes : "Au cours du cycle 2, l'élève a exploré, observé, expérimenté, questionné le monde qui l'entoure. Au cycle 3, les notions déjà abordées sont revisitées pour progresser vers plus de généralisation et d'abstraction, en prenant toujours soin de partir du concret et des représentations de l'élève. La construction de savoirs et de compétences, par la mise en oeuvre de démarches scientifiques et technologiques variées et la découverte de l'histoire des sciences et des technologies, introduit la distinction entre ce qui relève de la science et de la technologie, et

ce qui relève d'une opinion ou d'une croyance. La diversité des démarches et des approches (observation, manipulation, expérimentation, simulation, documentation ...) développe simultanément la curiosité, la créativité, la rigueur, l'esprit critique, l'habileté manuelle et expérimentale, la mémorisation, la collaboration pour mieux vivre ensemble et le goût d'apprendre.

En sciences, les élèves découvrent de nouveaux mondes de raisonnement en mobilisant leurs savoirs et leurs savoir-faire pour répondre à des questions. Accompagnés par leurs professeurs, ils émettent des hypothèses et comprennent qu'ils peuvent les mettre à l'épreuve, qualitativement ou quantitativement. Dans leur découverte du monde technique, les élèves sont initiés à la conduite d'un projet technique répondant à des besoins dans un contexte de contraintes identifiées. Enfin, l'accent est mis sur la communication individuelle ou collective, à l'oral comme à l'écrit en recherchant la précision dans l'usage de la langue française que requiert la science. D'une façon plus spécifique, les élèves acquièrent les bases de langages scientifiques et technologiques qui leur apprennent la concision, la précision et leur permettent d'exprimer une hypothèse, de formuler une problématique, de répondre à une question ou à un besoin, et d'exploiter des informations ou des résultats. Les travaux menés donnent lieu à des réalisations : ils font l'objet d'écrits divers retraçant l'ensemble de la démarche, de l'investigation à la fabrication. Une évaluation dans ce domaine a pour objet de mesurer le degré d'atteinte de ces finalités. L'évaluation vise à donner des informations sur les compétences acquises par les élèves. Les unités proposées se composeront donc d'items visant à mesurer les connaissances mémorisées d'une part, et, d'autre part la capacité des élèves à traiter les documents mis à leur disposition."

1.2.2 Tableau de compétences

1.3 Construction du test

Le bureau de l'évaluation des élèves de la DEPP élabore des évaluations par disciplines et niveaux scolaires. La préparation des unités et de leurs constituants fait intervenir des concepteurs, généralement des enseignants. La coordination est assurée par un chef de projet, membre de l'équipe du bureau de l'évaluation des élèves. Une application dédiée leur permet de créer, modifier ou éditer leur unité ; en outre cette application permet au chargé d'étude de gérer l'ensemble de l'évaluation (cf. plus loin l'encadré « GEODE »).

1.3.1 Elaboration des items

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

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, 2012 et 2015. Durant la période 2007-2013, un nouveau programme a été mis en place. Et au cours du cycle suivant (2013-2018), cela a été de nouveau le cas. Les évaluations de 2013 et 2018 ont respectivement pris en compte les nouvelles orientations successives. Pour chaque nouveau cycle, les items ont dû être réindexés sur une grille de compétences actualisée. L'évaluation 2018 permet d'observer, pour cette troisième prise d'informations le positionnement dans l'échelle des acquis des élèves.

Un équilibre de proportion entre les items considérés comme étant "faciles", "moyennement faciles" ou "difficiles" est recherché. Afin d'assurer une comparabilité, 73 items sur les 243 proposés aux élèves sont des items d'ancrage.

Deux types de formats de questions sont utilisés : les questions fermées (QCM, QCM-images, série, série-images) et les questions ouvertes appelant une réponse écrite (réponse courte - un chiffre, un nombre - ou réponse longue - production en autonomie de l'élève).

Un entraînement est prévu au début de chaque cahier afin de familiariser les élèves avec le type de question rencontré.

Les réponses des formats QCM, QCM-images sont saisies de manière automatisée à la fin de la passation. Les réponses des formats série et série-images sont

Tableau 2 – Grille de compétences du projet CEDRE

Contexte
personnel
scolaire
Domaine cognitif
Connaitre
Appliquer
Raisonner
Compétence
Pratiquer des démarches scientifiques et technologiques
Concevoir, créer, réaliser
S'approprier des outils et des méthodes
Pratiquer des langages
Mobiliser des outils numériques
Adopter un comportement éthique et responsable
Se situer dans l'espace et le temps
Thème
1. Matière, mouvement, énergie, information
2. Le vivant, sa diversité et les fonctions qui le caractérisent
3. Matériaux et objets techniques
4. La planète Terre. Les êtres vivants dans leur environnement
Attendu
1.1 Décrire les états et la constitution de la matière à l'échelle macroscopique.
1.2 Observer et décrire différents types de mouvements.
1.3 Identifier différentes sources d'énergie.
1.4 Identifier un signal et une information.
2.1Classer les organismes, exploiter les liens de parenté pour comprendre et expliquer l'évolution des organismes.
2.2 Expliquer les besoins variables en aliments de l'être humain ; l'origine et les techniques mises en oeuvre pour transformer et conserver les aliments.
2.3 Décrire comment les êtres vivants se développent et deviennent aptes à se reproduire.
2.4 Expliquer l'origine de la matière organique des êtres vivants et son devenir.
3.1 Identifier les principales évolutions du besoin et des objets.
3.2 Décrire le fonctionnement d'objets techniques, leurs fonctions et leurs constitutions.
3.3 Identifier les principales familles de matériaux.
3.4 Concevoir et produire tout ou partie d'un objet technique en équipe pour traduire une solution technologique répondant à un besoin.
3.5 Repérer et comprendre la communication et la gestion de l'information.
4.1 Situer la Terre dans le système solaire et caractériser les conditions de la vie terrestre.
4.2 Identifier des enjeux liés à l'environnement.

saisies de manière automatisée et donnent lieu à un regroupement ultérieur de leurs propositions. Dans le cas de ces séries, le seuil proposé correspond à une réponse correcte pour l'ensemble des propositions.

Les réponses des formats " réponse libre de l'élève " sont corrigées par des experts. Cela suppose la mise en place d'un dispositif de corrections, nécessitant la formation technique des correcteurs et l'élaboration d'un cahier de correction précis, déclinant les attendus pour éviter toute subjectivité ou la validation de réponses trop imprécises ou succinctes. Ce dispositif de correction à distance s'appuie sur le logiciel AGATE (cf. encadré "AGATE" p.29).

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 principale-

ment pour le format « papier », l'utilisation de la même technologie permet de générer du HTML pour la partie évaluation électronique (outil de maquettage).

1.3.2 Constitution des cahiers

Afin de pouvoir évaluer un nombre important d'items sans allonger le temps de passation pour l'élève, CEDRE utilise la méthodologie des cahiers tournants. Les items sont ainsi répartis dans des blocs d'une durée de 20 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 à passer l'ensemble de ceux-ci. Au final, pour l'évaluation CEDRE 2018, chaque cahier comprend deux séquences de 45 minutes. Les séances se terminent par un questionnaire de contexte, d'une durée de 20 minutes environ identique dans tous les cahiers, dans lequel l'élève doit répondre à des questions concernant notamment l'environnement scolaire, son intérêt et sa motivation pour les sciences expérimentales.

L'évaluation CEDRE 2018 est constituée de 13 cahiers tournants intégrant un ensemble de 13 blocs d'évaluations contenant des items de 2007 et 2013 repris à l'identique pour assurer une comparaison diachronique et de nouveaux items qui ont fait l'objet d'une expérimentation en 2017. Pour garantir la qualité de la comparaison avec 2007 et 2013, notamment en termes de passation des épreuves, l'évaluation de 2018 s'appuie sur 243 items dont 73 d'ancrage soit 30%.

1.4 Passation des évaluations

La passation de l'évaluation finale a eu lieu en mai 2018. Comme en 2013, cette évaluation a été précédée d'une expérimentation l'année N- 1 de façon à tester un grand nombre d'items auprès d'un échantillon réduit d'établissements. Dans chaque établissement, une personne a été désignée comme étant l'administrateur du test, son rôle étant de veiller au strict respect de la procédure à suivre pour que l'évaluation soit passée dans les meilleures conditions, quel que soit l'établissement ; la collecte de l'information s'est faite par questionnaires "papier-crayon".

L'anonymat des élèves et des personnels est respecté, chaque cahier étant repéré par un numéro. Une fois l'évaluation terminée, les cahiers et questionnaires sont renvoyés dans des conditionnements prévus à cet effet, préaffranchis

Tableau 3 – Exemple de répartition des blocs dans les cahiers

Cahier	Bloc 1	Bloc 2	Bloc 3	Bloc 4
E01	B5	B6	B12	B7
E02	B4	B13	B3	B8
E03	B6	B3	B2	B9
E04	B12	B2	B1	B13
E05	B3	B1	B7	B11
E06	B2	B7	B8	B10
E07	B1	B8	B9	B5
E08	B7	B9	B13	B4
E09	B8	B13	B11	B6
E10	B9	B11	B10	B12
E11	B13	B10	B5	B3
E12	B11	B5	B4	B2
E13	B10	B4	B6	B1

et pré-étiquetés. Aucun travail de correction n'a été demandé aux établissements.

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 sélectionnées 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 sélectionnées 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 : REP+), 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

Champ
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.

Sont donc exclus du champ :

— Les TOM.
— Les écoles hors contrat.
— Les écoles à l'étranger.
— Les écoles spécialisées.
— Les écoles de moins de 6 élèves de CM2.
— Les DOM.

Stratification
La stratification prend en compte le secteur d'enseignement de l'école :

1. écoles publiques hors éducation Prioritaire
2. écoles publiques en éducation Prioritaire
3. écoles privées

Modalités de sélection
Le tirage est à deux degrés. Le premier degré est composé d'écoles tirées dans chaque strate avec allocation proportionelle. Le deuxième degré consiste à interroger tous les élèves de l'école sélectionnée (tirage par grappe). La macro CUBE de l'INSEE est utilisée pour garantir des échantillons équilibrés sur la base de sondage selon certaines variables.

Dans chacune des strates, le tirage est équilibré sur la variable suivante :

— Le nombre total d'élèves de CM2

Echantillon 2018

L'échantillon vise 6 000 élèves répartis proportionnellement selon les trois strates.

Base de sondage
Le tableau 4 présente les exclusions dans la population ciblée.

Tableau 4 – Exclusions pour la base de sondage - CEDRE 2018 Sciences expérimentales École

	Établissements	**Elèves**
Ecoles accueillant des élèves de CM2	32 473	836 160
On retire les COM	32 473	836 160
On retire les écoles hors contrat	31 992	832 244
On retire les écoles spécialisées	31 971	831 676
On retire Mayotte	31 854	824 560
On retire les petites écoles (<6 CM2)	29 664	816 764
Base CEDRE CM2	29 664	816 764
On retire TIMSS CM1, CP12, Etape CP et Pilote MATHS	29 016	794 297
On retire TIMSS CM1, CP12, Etape CP et Pilote MATHS	29 016	794 297

Le tableau 5 présente la répartition de la population ciblée selon le secteur d'enseignement.

Tableau 5 – Répartition dans la base de sondage - CEDRE 2018 Sciences expérimentales École

Strate	**Établissements**	**Élèves**
1. Public hors EP	21 659	558 354
2. EP	3 161	115 025
3. Privé	4 196	120 918
Total	**29 016**	**794 297**

Échantillon

Le tableau 6 présente la répartition de l'échantillon selon le secteur d'enseignement. Au total, 220 écoles ont été sélectionnées.

Tableau 6 – Répartition dans l'échantillon - CEDRE 2018 Sciences expérimentales École

Strate	**Établissements**	**Élèves**
1. Public hors EP	162	4 156
2. EP	26	956
3. Privé	32	919
Total	**220**	**6 031**

2.3 État des lieux de la non-réponse

2.3.1 Non-réponse totale

Parmi la non-réponse totale, nous distinguons la non-réponse des établissements de la non-réponse des élèves des établissements participants. Les chiffres suivants ont été observés pour 2018.

88.2 % des établissements de l'échantillon ont répondu à l'évaluation (tableau 7). 83.5 % des effectifs attendus ont participé (tableau 8).

Tableau 7 – Non-réponse des établissements - CEDRE 2018 Sciences expérimentales École

Strate	Nb établissements attendus	Nb établissements répondants	% d'établissements répondants
1. Public hors EP	162	147	90.7 %
2. EP	26	21	80.8 %
3. Privé	32	26	81.2 %
Total	**220**	**194**	**88.2 %**

Tableau 8 – Non-réponse des élèves - CEDRE 2018 Sciences expérimentales École

Strate	Nb élèves attendus	Nb élèves répondants	% d'élèves répondants
1. Public hors EP	4 156	3 612	86.9 %
2. EP	956	768	80.3 %
3. Privé	919	657	71.5 %
Total	**6 031**	**5 037**	**83.5 %**

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"). La non réponse terminale a été étudiée par séquence et par cahier. Si un élève a passé moins de 50 % d'une séquence, on considère qu'il n'a pas vu la séquence (code "s").

Parmi les élèves concernés, la non-réponse terminale représente en moyenne :

— 1.5 items pour la séquence 1
— 2.7 items pour la séquence 2

On considère que :

— 56 élèves n'ont pas vu la séquence 1, dont :
  — 38 n'ont répondu à aucun item de la séquence
  — 18 ont répondu à moins de 50 % de la séquence
— 129 élèves n'ont pas vu la séquence 2, dont :
  — 94 n'ont répondu à aucun item de la séquence
  — 35 ont répondu à moins de 50 % de la séquence

Les élèves dont toutes les séquences sont codées en "s" sont classés en non réponse totale. C'est le cas pour 5 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.

Tableau 9 – Comparaison entre les marges de l'échantillon et les marges dans la population - CEDRE 2018 Sciences expérimentales École

Modalité	Variable	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population
Retard	1	56 275	62 891	6.89	7.7
	2	760 489	753 873	93.11	92.3
Sexe	1	411 812	415 488	50.42	50.87
	2	404 952	401 276	49.58	49.13
Strate	1	567 406	567 406	69.47	69.47
	2	127 334	127 334	15.59	15.59
	3	122 025	122 025	14.94	14.94

2.5 Précision

L'erreur standard (*se*) peut être calculée sur le score moyen de chaque année (tableau 10).

Tableau 10 – Scores moyens et erreurs standard associées - CEDRE 2018 Sciences expérimentales École

Année	Score moyen	Erreur standard
2007	250	1.59
2013	249.3	1.28
2018	249.1	1.04

Pour savoir par exemple si l'évolution entre 2013 et 2018 est significative , il faut calculer la valeur suivante :

$$\frac{\left|\hat{\hat{Y}}_{2018} - \hat{\hat{Y}}_{2013}\right|}{\sqrt{se^2_{\hat{\hat{Y}}_{2018}} + se^2_{\hat{\hat{Y}}_{2013}}}} \tag{2}$$

Entre 2013 et 2018, on obtient une valeur de 0.16 (inférieure à 1.96). Cela signifie que l'évolution du score moyen n'est pas statistiquement significative.

Les erreurs standards sont également calculées pour les répartitions dans les différents groupes de niveaux (tableaux 11 et 12).

Tableau 11 – Répartitions en % dans les groupes de niveaux - CEDRE 2018 Sciences expérimentales École

Année	Groupe <1	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5
2007	1.6	13.5	29.5	28.2	17.2	10
2013	2.7	13	29	28.4	16.9	10
2018	1.6	9.7	29.8	34.7	17.7	6.6

Tableau 12 – Erreurs standards des répartitions en % dans les groupes de niveaux - CEDRE 2018 Sciences expérimentales École

Année	Groupe <1	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5
2007	0.2	0.7	1	0.8	0.7	0.9
2013	0.2	0.6	0.7	0.7	0.6	0.6
2018	0.3	0.6	0.9	0.8	0.7	0.4

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 :

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

Tableau 13 – Effet du plan de sondage - CEDRE 2018 Sciences expérimentales École

Année	Erreur Standard	Erreur SAS	*Design Effect*
2007	1.59	0.97	1.64
2013	1.28	0.69	1.85
2018	1.04	0.6	1.72

Dans le cas d'un sondage en grappes, la précision est dégradée en comparaison d'un sondage aléatoire simple. Cela signifie qu'en 2018, un sondage aléatoire simple avec un effectif 1.72 fois moins important aurait conduit au même niveau de précision.

3 Analyse des items

3.1 Méthodologie

Pour une description générale de la méthodologie psychométrique employée dans les évaluations standardisées de compétences des élèves, le lecteur est invité à consulter Rocher (2015).

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.

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.

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.

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.

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)

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 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.

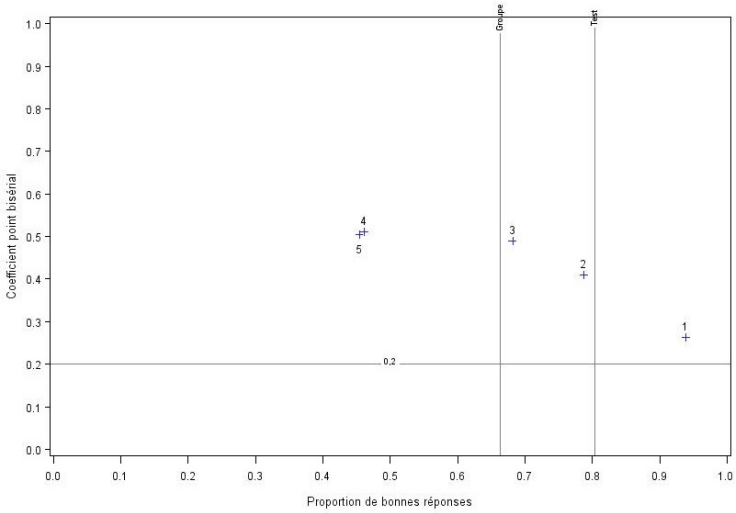
— Non-réponse terminale : l'élève s'est arrêté au cours de l'épreuve, potentiellement en raison d'un manque de temps. Des choix sont effectués pour déterminer le traitement de ces valeurs. Nous considérons que si un élève a passé moins de 50 % d'une séquence, il n'a pas vu la séquence, les valeurs manquantes 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, 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



Note de lecture : L'item présenté ici est une série de cinq questions de type « Vrai/Faux ». Chaque croix représente l'item correspondant au seuil de réussite retenu. Par exemple, si la réussite à l'ensemble est attribuée dès lors qu'une seule question est réussie, l'item obtenu a un taux de réussite d'environ 95 % et un coefficient bisérial d'environ 0,26. Si le seuil de réussite est fixé à 3 questions réussies sur 5, alors le taux de réussite baisse mécaniquement (autour de 65 % qui est le taux de réussite obtenu à l'ensemble des questions de cet item).

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 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

Le calcul des indices de discrimination conduit à éliminer 18 items dont les indices *rbis-point* sont trop faibles :

- 2 items communs à 2007 et 2013
- 1 item commun à 2007, 2013 et 2018
- 15 items de 2018

4 Modélisation

4.1 Méthodologie

4.1.1 Modèle de réponse à l'item

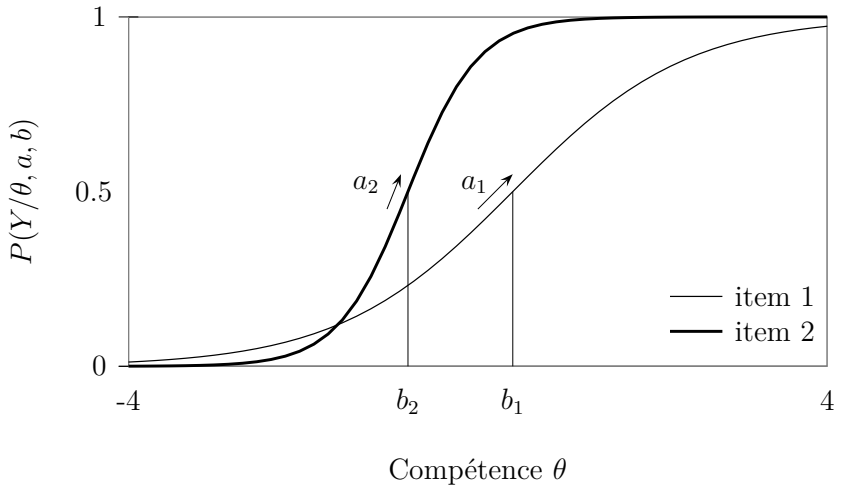
Le modèle de mesure utilisé est un modèle de réponse à l'item à deux paramètres avec une fonction de lien logistique (MRI 2PL) :

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

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 2 représente les courbes caractéristiques de deux items selon cette modélisation.

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



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

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. Pour une présentation générale, le lecteur est invité à consulter Rocher (2015).

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 2).

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 pour l'année 2007).

Sous l'hypothèse d'indépendance locale des items [4], la fonction de vraisemblance s'écrit :

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

4. Cette hypothèse signifie que les indicatrices de réussite des items sont indépendantes, conditionnellement au niveau de compétence θ. A niveau de compétence égal, deux items donnés ne sont pas corrélés : seule la compétence θ explique la corrélation entre deux items. Cette hypothèse est ainsi liée à l'hypothèse d'unidimensionnalité de θ (cf, Rocher, 2013).

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 utilisée. 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 :

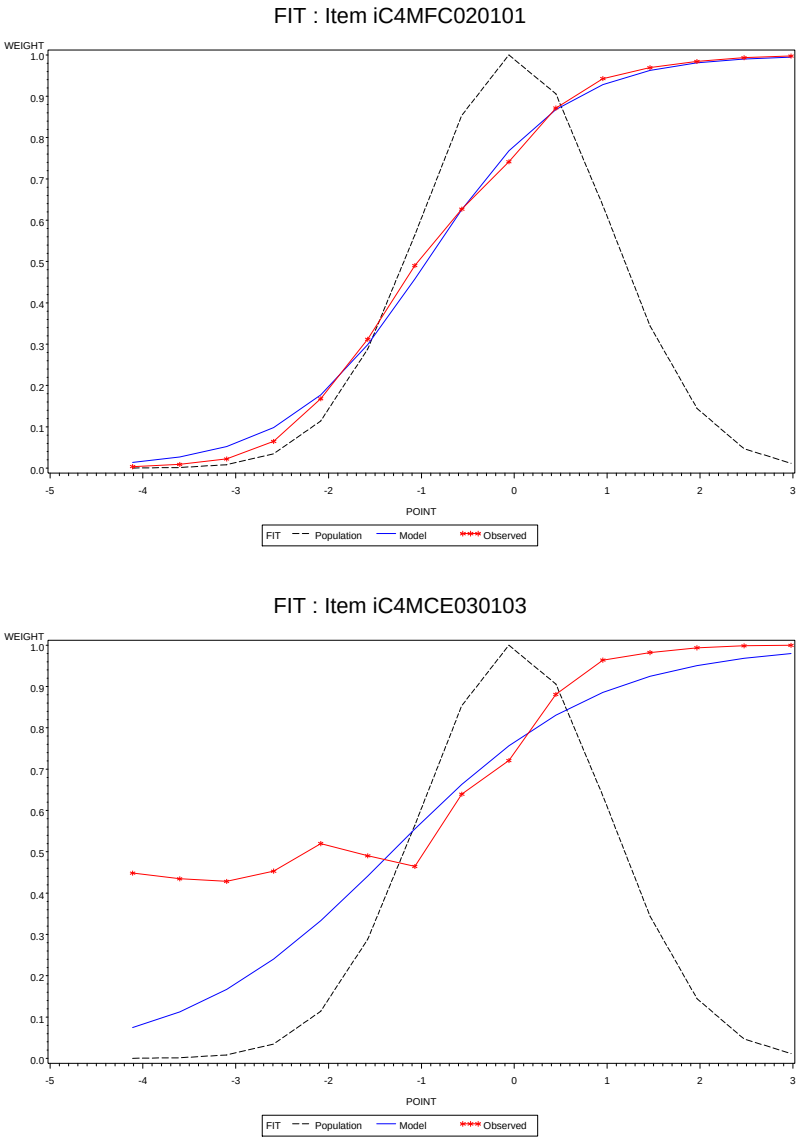
$$\frac{\partial lnL}{\partial \theta_i} + \frac{J}{2I} = 0 \tag{15}$$

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 3). Certaines procédures proposent de comparer directement les probabilités théoriques avec les proportions de réussite de groupes d'élèves. Plus généralement, nous pouvons écrire les résidus de la manière suivante :

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

Figure 3 – Exemples d'ajustements (FIT)



Note de lecture : La courbe bleue représente la courbe caractéristique de l'item telle qu'estimée par le modèle. La courbe en rouge relie des points qui correspondent aux taux de réussite observés 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 en haut. Il est très mauvais pour celui du bas.

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 et empiriquement (le lien théorique n'est pas établi) une loi normale (Smith, Schumaker, & Bush, 1998).

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.

Si la probabilité de réussite, conditionnellement au niveau mesuré, est différente selon les groupes d'élèves, alors il existe un fonctionnement différentiel.

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 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.

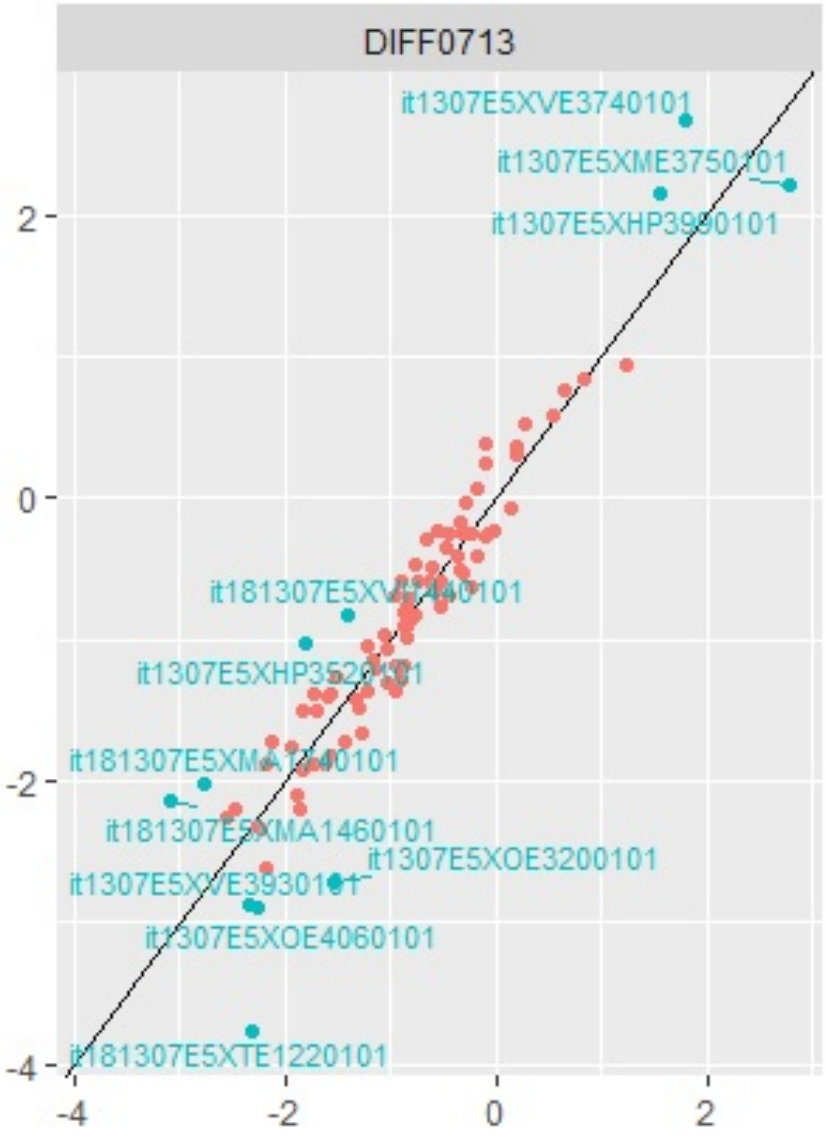
4.2 Résultats

4.2.1 Identification des fonctionnements différentiels d'items (FDI)

17 items ont été éliminés des calculs :

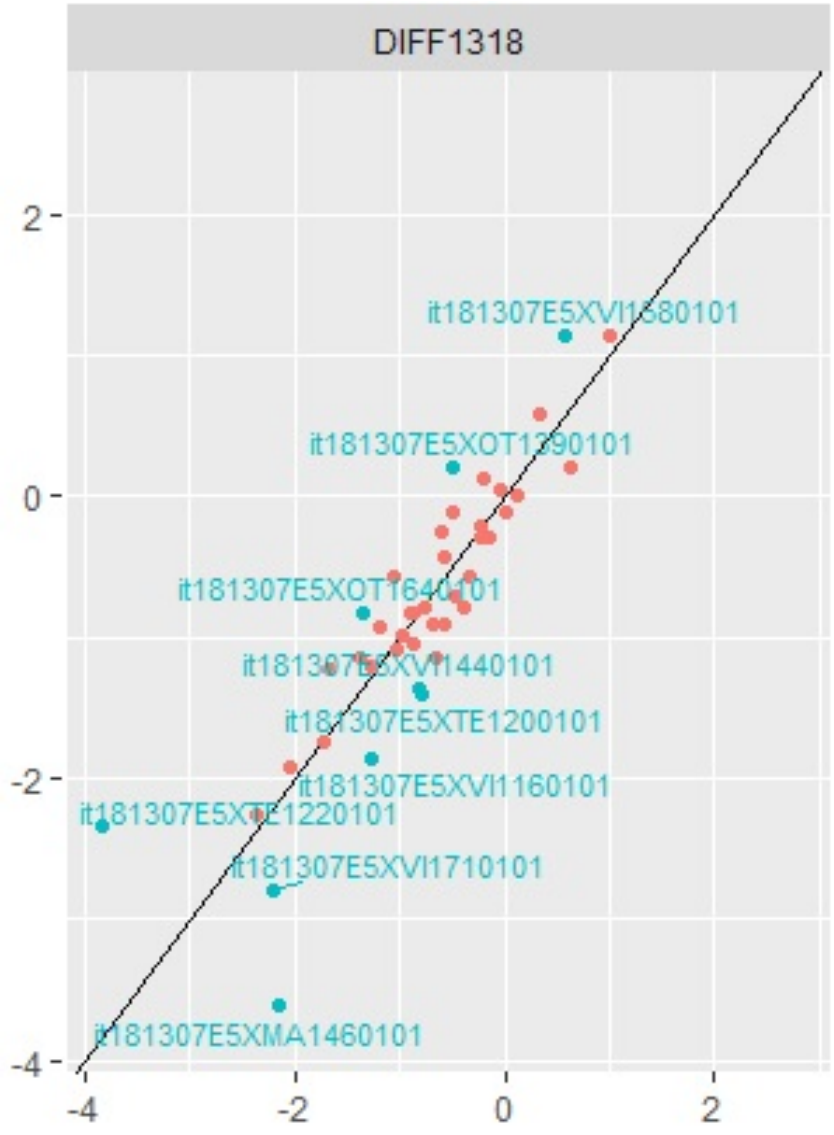
— 7 items pour 2007-2013
— 10 items pour 2007-2013-2018

Figure 4 – Comparaison des paramètres de difficulté 2007-2013 - (CEDRE Sciences expérimentales 2018 École)



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. Les items présentant un FDI apparaissent en bleu.

Figure 5 – Comparaison des paramètres de difficulté 2013-2018 - (CEDRE Sciences expérimentales 2018 École)



Note de lecture : Les points sont les items. En abscisse figure la valeur des paramètres de difficulté estimés en 2013, et en ordonnée la la valeur des paramètres de difficulté estimés et ajustés pour l'année 2018. Les items présentant un FDI apparaissent en bleu.

4.2.2 Identification des items présentant un mauvais ajustement (FIT)

Aucun item présentant un mauvais ajustement n'a été détecté.

4.2.3 Bilan de l'analyse des items

En considérant l'ensemble des items sur les 3 années, il y avait au départ :
- 145 items de 2018
- 55 items d'ancrage 2007-2013
- 43 items d'ancrage 2007-2013-2018

Cela représente 243 items passés par les élèves en tout, dont 188 en 2018.

Après suppression des items présentant un mauvais Rbis, un fonctionnement différentiel ou un mauvais ajustement, il reste :
- 130 items de 2018
- 46 items d'ancrage 2007-2013
- 32 items d'ancrage 2007-2013-2018

208 items sont donc conservés dans l'analyse, dont 162 utilisés dans l'évaluation 2018.

4.3 Calcul des scores

Comme indiqué précédemment, une analyse conjointe des données des 3 années 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 14 présente les résultats obtenus.

Tableau 14 – Niveaux de compétences (moyennes des scores et écarts-types) - CEDRE 2018 Sciences expérimentales École

Année	Score moyen	Écart-type
2007	250	50
2013	249.3	51.9
2018	249.1	42.5

5 Construction de l'échelle

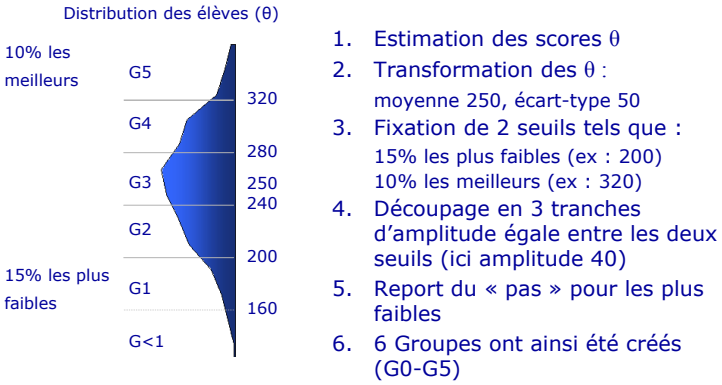
5.1 Méthode

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 expérimentales 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 < 1 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.

Figure 6 – Principes de construction de l'échelle



5.2 Caractérisation des groupes de niveaux

À partir de cette correspondance entre les items et les groupes, une description qualitative et synthétique des compétences maitrisées par les élèves des différents groupes est proposée.

Groupe < 1 (1,6 % des élèves)

Les élèves du groupe < 1 représentent 1,6 % des élèves interrogés. Ils sont en capacité de répondre correctement à très peu de questions (2,9 % de l'ensemble des questions). La majeure partie de leurs réussites concerne des connaissances simples en relation avec le vécu. Ce sont essentiellement des questions portant sur le domaine du vivant. La plupart des questions auxquelles ils répondent présentent un support iconographique (photographie).

Groupe 1 (9,7 % des élèves)

Les élèves du groupe 1 représentent 9,7 % des élèves interrogés. Ils sont capables de répondre correctement à très peu de questions (6,7 % de l'ensemble des questions). La majeure partie des bonnes réponses concerne le domaine du vivant. Ils commencent à être en capacité d'expliquer un phénomène simple (respiration, stades de développement). Ils sont capables de comprendre des modes de représentation formalisés (schéma, pictogramme) et de rendre compte d'observations.

Groupe 2 (29,8 % des élèves)

Les élèves du groupe 2 représentent 29,8 % des élèves interrogés. Ils réussissent en moyenne 38 % des items de l'évaluation. C'est le niveau à partir duquel les élèves peuvent réussir dans tous les thèmes étudiés à l'école élémentaire.

Ils réussissent des items descriptifs liés à des observations directes sans l'aide du support iconographique. Ils commencent à choisir des conclusions à des expériences scientifiques en rapport avec la vie quotidienne. Ils sont capables de traiter des données. Ils savent relier des connaissances acquises en sciences et technologie à des questions de sécurité et d'environnement. Ils sont en capacité de mettre en jeu leurs connaissances pour catégoriser les êtres vivants, les matériaux et les objets techniques et caractériser un phénomène naturel (séisme, éruption volcanique, tsunami). Ils peuvent reconnaitre les fonctions et identifier les évolutions des besoins et des objets techniques dans leur contexte. Ils peuvent mobiliser des compétences pour donner des repères sur les saisons, prélever des informations chiffrées sur un schéma d'astronomie, identifier les sources et formes d'énergie et les différentes formes de signaux, comprendre la communication de l'information. Ils savent reconnaitre la Terre dans le système solaire.

Groupe 3 (34,7 % des élèves)

Les élèves du groupe 3 représentent 34,7 % des élèves interrogés. C'est le groupe le plus important. Ils réussissent en moyenne 78,4 % des items de l'évaluation. Ils peuvent expliquer un phénomène dans tous les domaines étudiés. Ils sont capables de choisir le matériel adapté pour réaliser une expérience, réaliser une production ou effectuer une mesure. Ils savent interpréter un résultat, en tirer une conclusion. Ils mettent en oeuvre des compétences pour exploiter un document complexe (schéma, carte, tableau, texte, document composite) de façon à y rechercher des informations et les mettre en relation pour répondre à une question. Ils ont des capacités d'analyse et d'inférence. Ils sont en mesure de choisir des formulations de réponses, dont la compréhension nécessite un raisonnement plus complexe. Ils sont capables de décrire les mouvements de la Terre, de reconnaitre les planètes du système solaire et les situer. Ils savent identifier quelques impacts humains dans leur environnement (réchauffement climatique, disparition des animaux), relier certains phénomènes à des risques pour la population (séisme, cyclone). Ils peuvent identifier les différents mouvements (circulaire, rectiligne). Ils utilisent un vocabulaire scientifique précis pour décrire les fonctions de nutrition, digestion, respiration.

Groupe 4 (17,7 % des élèves)

Les élèves du groupe 4 représentent 17,7 % des élèves interrogés. Ils peuvent proposer une expérience pour tester une hypothèse. Leurs capacités à raisonner et inférer leur permettent de confirmer ou infirmer une hypothèse en prenant appui sur un document composite. Ils savent formaliser une partie de leur recherche en ordonnant les étapes d'une démarche scientifique. Ils sont capables de concevoir et schématiser un protocole pour répondre à une question initiale donnée en s'appuyant sur une liste de matériel suggérée. Ces élèves ont des compétences pour compléter un texte en s'appuyant sur un schéma et sa légende,

mais aussi pour se représenter mentalement une situation et d'en rendre compte. Ils sont en mesure d'identifier des caractères communs entre des organismes à partir d'une classification par emboitements.

Groupe 5 (6,6 % des élèves)

Les élèves du groupe 5 représentent 6,6 % des élèves interrogés. Ils ont de solides habiletés pour raisonner et exercer une analyse experte sur des documents scientifiques variés. Ils savent interpréter un graphique en inférant le complément des données directement représentées. Ils ont les compétences pour prélever dans un schéma des informations données indirectement afin de résoudre un problème.

5.3 Exemples d'items

5.3.1 Item caractéristique des groupes < 1 et 1

Cet exercice fait partie du thème " Matière, mouvement, énergie, information ". Il évalue la compétence " Pratiquer des langages ". Le domaine cognitif en jeu est " Appliquer ". Le contexte est personnel.

Figure 7 – Exemple groupe < à 1

Voici un pictogramme :

Questions

Quelle est est sa signification ? Coche la bonne réponse.

1	☐	interdiction de courir	
2	☐	défense de sortir par cette issue	
3	☐	défense d'entrer par cette porte	
4	☒	issue de secours	E5XMA720 101 23

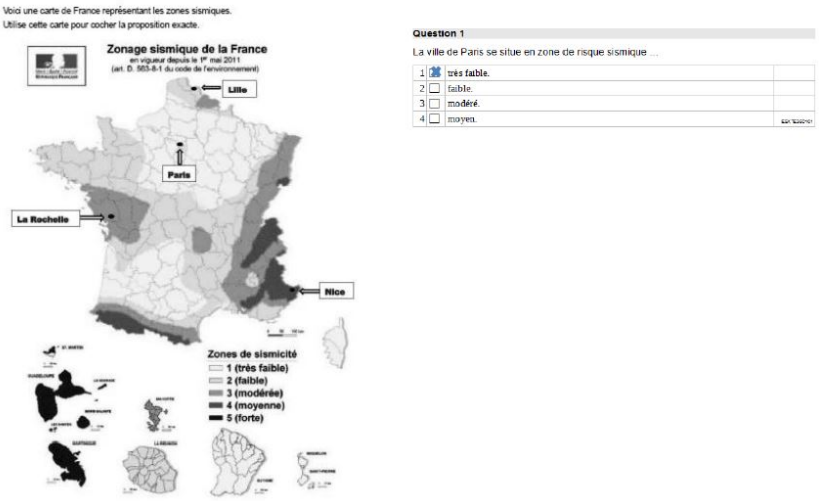
Ici, l'élève doit observer le pictogramme qui lui est proposé et choisir parmi quatre propositions celle qui correspond à sa signification. L'item est réussi par 92 % des élèves et la non réponse est faible. Il s'agit vraisemblablement d'un pictogramme connu, rencontré dans la vie quotidienne.

5.3.2 Item caractéristique du groupe 2

Cet exercice fait partie du thème " La planète Terre. Les êtres vivants dans leur environnement ". Il évalue la compétence " Pratiquer des langages ". Le domaine cognitif en jeu est " Appliquer ". Le contexte est scolaire.

L'exercice est composé d'une carte de France avec les légendes des zones sismiques allant de blanc à noir, et d'une question à choix multiples. Il présente un taux de réussite de près de 83 %. Pour répondre, il fallait associer la zone géographique où se trouve la ville de Paris au risque de sismicité indiqué en légende. Cet item présentait pour difficulté notoire la nécessité d'utiliser une légende avec un vocabulaire différent entre la question et la légende.

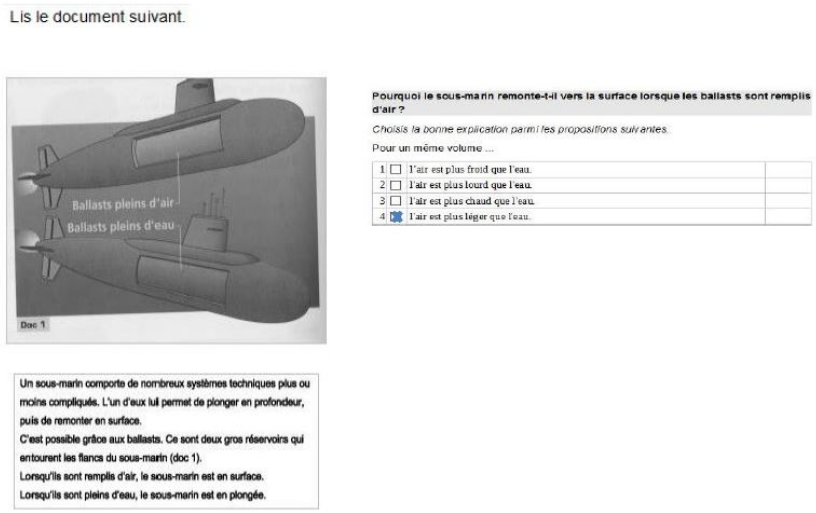
Figure 8 – Exemple groupe 2



5.3.3 Item caractéristique du groupe 3

Cet exercice fait partie du thème " Matière, mouvement, énergie, information ". Il évalue la compétence " Pratiquer des démarches scientifiques et technologiques". Le domaine cognitif en jeu est " Raisonner ". Le contexte est scolaire.

Figure 9 – Exemple groupe 3

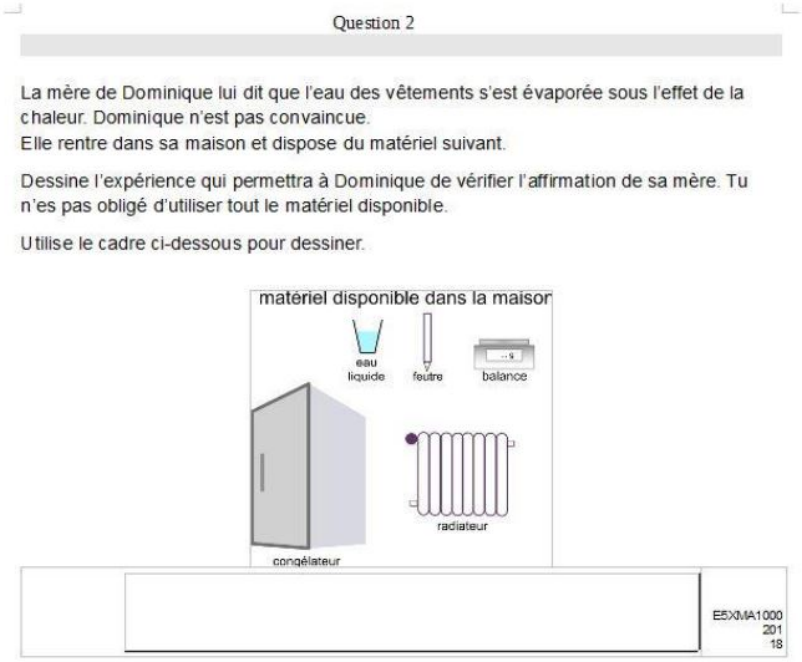


Le document est composé d'une illustration, d'un texte et d'une question à choix multiples. L'élève doit prélever des informations et mettre en relation le texte et l'illustration pour répondre à la question parmi 4 réponses proposées. Pour réussir cet item, il faut donc maitriser la lecture et la compréhension d'un document scientifique mais aussi passer d'une forme de langage scientifique à une autre (illustration et texte). L'élève doit aussi interpréter ses observations pour répondre à une question de nature scientifique. Le taux de réussite de 2018 (67,5 %) est moins important qu'en 2013 (72,9 %) . Si l'on regarde plus précisément les programmes alors que l'air et son caractère pesant étaient abordés aux cycles des approfondissements dans les programmes 2008, on ne parle de l'air qu'à travers des exemples de situations pour traiter des mélanges gazeux dans le programme de 2016. On peut donc supposer que le changement de programme de 2016 impacte directement le taux de réussite de cet exercice.

5.3.4 Item caractéristique du groupe 4

Cet exercice fait partie du thème " Matière, mouvement, énergie, information ". Il évalue la compétence " Pratiquer des démarches scientifiques et technologiques". Le domaine cognitif en jeu est " Appliquer ". Le contexte est scolaire.

Figure 10 – Exemple groupe 4

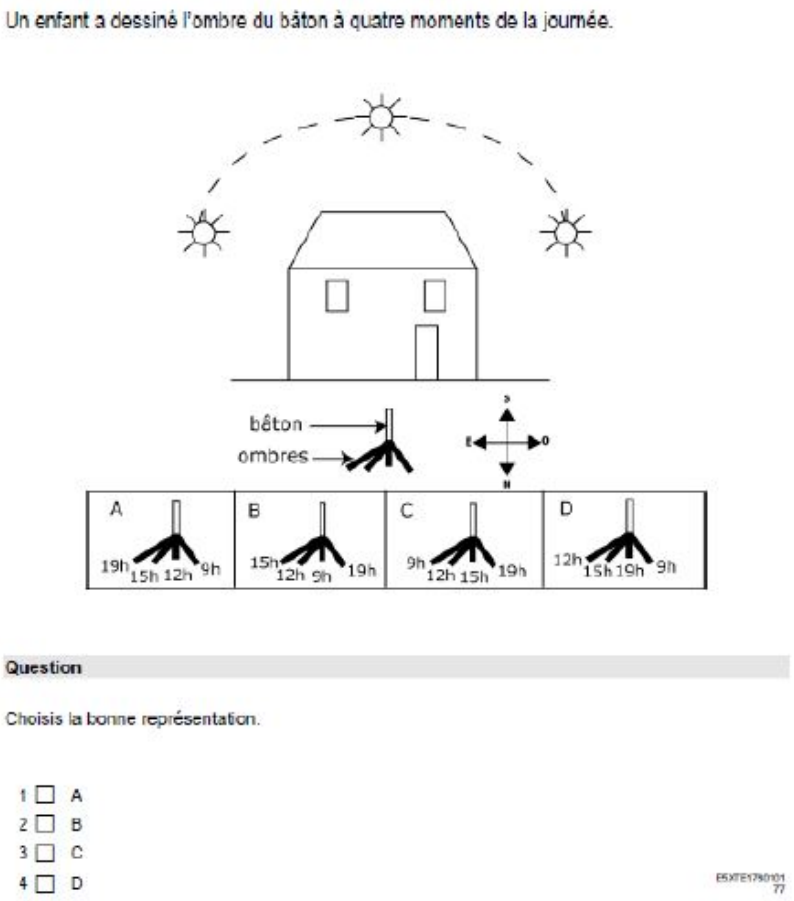


Ici, le taux de réussite est faible 39 % et le taux de non réponse 17 %. L'exercice est composé d'un texte présentant une situation problème et d'une liste de matériel disponible. L'élève doit saisir la situation problème, choisir le matériel et proposer un dispositif expérimental sous forme de dessin. Il doit lire et comprendre la situation qui lui est proposée. Il doit ensuite trier parmi le matériel qui lui est proposé celui qui est adapté pour réaliser l'expérience qui lui permettra de vérifier l'affirmation de la mère de Dominique : " l'eau des vêtements s'est évaporée ". L'élève est potentiellement confronté à trois types de difficultés : ses connaissances scientifiques, une compréhension fine en maitrise de la langue, puis la projection d'un résultat expérimental. Dans ce dernier item le fait de devoir élaborer un dispositif expérimental semble une tâche difficile à mettre en oeuvre pour l'élève.

5.3.5 Item caractéristique du groupe 5

Cet exercice fait partie du thème " La planète Terre. Les êtres vivants dans leur environnement ". Il évalue la compétence " Pratiquer des lagages". Le domaine cognitif en jeu est " Raisonner ". Le contexte est scolaire.

Figure 11 – Exemple groupe 5



L'item est composé d'une phrase présentant l'exercice, d'un schéma et d'une question à choix multiples. L'élève doit associer chacune des positions du soleil à un horaire. Il doit savoir que plus la trajectoire du soleil est basse à l'horizon, plus le soleil " se lève tard " et " se couche tôt ", et inversement. Il faut aussi voir qu'ici la rose des vents est inversée. Les élèves ont bien compris que la bonne réponse se situait parmi les réponses dont les positions horaires des ombres étaient ordonnées. La réponse qui a été majoritairement choisie est celle qui indique ces positions dans l'ordre chronologique, ce qui indique que ces élèves n'ont pas pris en compte l'orientation inversée de la rose des vents.

6 Variables contextuelles et non cognitives

6.1 Variables sociodémographiques et indice de position sociale

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. Le lecteur est invité à consulter la Note d'Information pour plus de détails (Augé, Etève, & Ninnin, 2019).

L'indice de position sociale mesure la proximité au système scolaire du milieu familial de l'enfant. Cet indice peut se substituer à la profession des parents pour mieux expliquer les parcours et la réussite scolaire de leurs enfants. Il consiste en une transformation des PCS en valeur numérique (Rocher, 2016).

Il n'a été possible d'établir des comparaisons qu'en termes de niveau social des écoles, et non au niveau individuel. En effet, en 2017, la PCS des parents est disponible pour chaque élève, mais elle ne l'était pas dans les cycles antérieurs. Pour chaque établissement des échantillons de 2007, 2013 et 2018, la moyenne de l'indice de position socio-scolaire a été calculée et la population a ensuite été découpée en quatre groupes selon les quartiles (tableau 15).

Tableau 15 – Score moyen selon l'indice de position sociale moyen de l'établissement (CEDRE 2007-2013-2018)

Indice moyen école	Année	Répartition (%)	Score moyen	Écart type
1er quart	2007	24.2	240	47
1er quart	2013	24.9	236	49
1er quart	2018	24.1	241	**43**
2e quart	2007	25.5	246	49
2e quart	2013	25.0	247	51
2e quart	2018	25.9	245	**41**
3e quart	2007	24.3	251	50
3e quart	2013	25.1	251	51
3e quart	2018	24.7	250	**41**
4e quart	2007	25.9	262	52
4e quart	2013	25.1	263	52
4e quart	2018	25.3	260	**43**

Note de lecture : en 2018, le score moyen des élèves appartenant au quart des écoles les plus défavorisées (1er quart) augmente de 5 points par rapport à 2013. Les évolutions significatives sont indiquées en gras.

6.2 Élaboration des questionnaires de contexte

Pour pouvoir davantage enrichir l'analyse des résultats, deux questionnaires de contexte ont été élaborés. Un questionnaire élève a été ajouté à la fin du cahier d'évaluation et un questionnaire enseignant était adressé aux enseignants des classes participantes à l'évaluation. Ces questionnaires ont été élaborés en collaboration avec des chercheurs et des spécialistes en sciences de l'éducation.

Le questionnaire enseignant interroge les enseignants sur leur niveau de formation et leur ancienneté. Ce questionnaire inclut aussi des questions sur les pratiques pédagogiques, les stratégies d'enseignement, le sentiment d'efficacité personnelle etc.

Le questionnaire élève interroge des dimensions dites conatives intéressantes à mettre en lien avec le score obtenu à l'épreuve - les stratégies de lecture, la motivation, la perception de soi et l'anxiété scolaire. De plus, il est demandé aux élèves d'évaluer la difficulté de l'épreuve et leur degré d'implication à faire le test.

Le questionnaire élève contient aussi un certain nombre de questions à renseigner par l'enseignant(e), il s'agit des questions concernant la catégorie socioprofes-

sionnelle des parents mais aussi le parcours de l'élève (raccourcissement de cycle ou maintien dans un cycle, orientation retenue etc.).

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 (Keskpaik. & Rocher, 2015). Cet instrument (cf. figure 12) a été introduit dans plusieurs évaluations conduites au niveau national par la DEPP, y compris dans CEDRE maîtrise de la langue. 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.

Le tableau 16 présente les grands résultats de cet instrument.

Tableau 16 – Résultats de l'instrument de mesure de la motivation au test (CEDRE 2018)

		%
Comment as-tu trouvé les exercices de cette évaluation ?	Très faciles	7.5
	Faciles	59.6
	Difficiles	31.0
	Très difficiles	1.9
Je me suis bien appliqué(e) pour faire cette évaluation	Pas du tout d'accord	1.9
	Pas d'accord	6.7
	D'accord	62.5
	Tout à fait d'accord	28.9
Je me suis autant appliqué(e) pour faire cette évaluation que le travail quotidien de classe	Pas du tout d'accord	6.7
	Pas d'accord	20.4
	D'accord	42.0
	Tout à fait d'accord	31.0

Figure 12 – 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. Elle a obtenu en mars 2015 la certification pour les évaluations CEDRE.

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

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

Augé, V., Etève, Y., & Ninnin, L.-M. (2019). CEDRE 2007-2013-2018, histoire, sciences en fin d'école : des résultats stables depuis 11 ans et un niveau plus homogène. *Note d'information*, *32*.

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., & Rocher, T. (2015). 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*, 119-139.

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.

Rocher, T. (2015). Mesure des compétences : méthodes psychométriques utlisées dans le cadre des évaluations des élèves. *Éducation et Formations*, *86-87*, 37-60.

Rocher, T. (2016). Construction d'un indice de position sociale des élèves. *Éducation et Formations*, *90*, 5-27.

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.

Trosseille, B., & Rocher, T. (2015). Les évaluations standardisées des élèves. perspective historique. *Éducation et Formations*, *85-86*, 15-35.

Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54 nº3*, 427-450.

Liste des tableaux

Table des figures