CEDRE

Cycle des Évaluations Disciplinaires Réalisées sur Échantillons

Rapport technique

Mathématiques 2019

École

Auteurs :

Louis PHILBERT
Yann ETEVE
Louis-Marie NINNIN
Reinaldo DOS SANTOS
Thierry ROCHER
Marion LE CAM

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

Decembre 2022

Table des matières

Introduction

La Direction de l'Evaluation, de la Prospective et de la Performance (DEPP) met en place des dispositifs d'évaluation des acquis des élèves reposant sur des épreuves standardisées. Elle est également maître d'oeuvre 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'Evaluations Disciplinaires Réalisées sur Echantillons) 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 maitrise pratiquement complète de ces compétences à une maitrise 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 : 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 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.

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

Les connaissances et compétences permettant de cerner les acquis des élèves ont été retenues selon les finalités assignées à l'enseignement des mathématiques. Une évaluation en mathématiques a pour objet de confronter les résultats du fonctionnement pédagogique du système éducatif aux objectifs qui lui sont assignés.

L'évaluation CEDRE en fin d'école en mathématiques vise à faire le point sur les acquis des élèves et à mesurer l'évolution de ces connaissances et compétences sur trois temps de mesure (2008, 2014 et 2019).

Les connaissances et compétences telles qu'elles sont définies dans le programme constituent le cadre de cette évaluation (programmes officiels en vigueur à partir de la rentrée scolaire 2015-2016 parus au BO spécial n° 11 du 26 novembre 2015 ainsi que les ressources de référence du site Eduscol).

Le programme de mathématiques de 2015 intègre trois changements majeurs :

- **La définition des cycles** (« Le cycle 3 relie désormais les deux dernières années de l'école primaire et la première année du collège, dans un souci renforcé de continuité pédagogique et de cohérence des apprentissages au service de l'acquisition du socle commun de connaissances, de compétences et de culture. Ce cycle a une double responsabilité : consolider les apprentissages fondamentaux qui ont été engagés au cycle 2 et qui conditionnent les apprentissages ultérieurs ; permettre une meilleure transition entre l'école primaire et le collège en assurant une continuité et une progressivité entre les trois années du cycle. »
- Les mathématiques sont découpées en trois champs (Nombres et calculs, Grandeurs et mesures, Espace et géométrie).
- La prise en compte du socle commun intégrant les **nomenclatures issues du second degré** (Chercher, Modéliser, Représenter, Raisonner, Calculer

et Communiquer).

L'évaluation CEDRE 2019 prend en compte ces changements :

Le cycle 3 se terminant en fin de sixième, les concepteurs se réfèrent aux repères de progressivité pour identifier un attendu de fin de CM2. Les items précédemment codés en six champs (Connaissance des nombres entiers naturels, Fractions et nombres décimaux mathématiques, Calcul, Espace et géométrie, Grandeurs et mesures, Organisation et gestion de données) ont été ventilés dans les trois champs du programme en vigueur (Nombres et calcul, Grandeurs et mesures, Espace et géométrie). La prise en compte de la nouvelle nomenclature du socle commun conduit à un nouveau codage des items de l'évaluation afin de proposer une analyse complémentaire dans un dossier à paraître.

Afin de cerner au mieux les objectifs de l'évaluation un tableau de pilotage définit pour chaque champ mathématique, les thèmes abordés :

Nombres et calculs :

- (NA) Utiliser et représenter les grands nombres entiers
- (NB) Utiliser et représenter des fractions simples
- (NC) Utiliser et représenter les nombres décimaux
- (ND) Calculer avec des nombres entiers
- (NE) Calculer avec des nombres décimaux
- (NF) Résoudre des problèmes en utilisant des fractions simples
- (NG) Résoudre des problèmes en utilisant les nombres décimaux
- (NH) Résoudre des problèmes en utilisant le calcul

Grandeurs et mesures :

- (MI) Comparer, estimer, mesurer des grandeurs géométriques avec des nombres entiers
- (MJ) Comparer, estimer, mesurer des grandeurs géométriques avec des nombres décimaux
- (MK) Utiliser le lexique, les unités, les instruments de mesures spécifiques de ces grandeurs
- (ML) Résoudre des problèmes impliquant des grandeurs
- (MM) Proportionnalité

Espace et géométrie :

- (GN) Se repérer et se déplacer dans l'espace en utilisant ou en élaborant des représentations
- (GO)Reconnaître, nommer, décrire, reproduire, représenter, construire quelques solides et figures géométriques
- (GP) Reconnaître et utiliser quelques relations géométriques
- (I) Initiation à la programmation

1.2.1 Particularité de l'évaluation 2019

L'évaluation CEDRE 2019 est entièrement effectuée au format numérique à l'aide de tablettes tactiles. Le passage au numérique est un enjeu majeur. Le passage d'une évaluation sur support papier à une évaluation sur support numérique doit être contrôlé afin de garantir la comparabilité entre deux points de mesure d'une part, et d'autre part permettre la mise œuvre d'items tenant compte des nouvelles possibilités apportées par ce l'outil numérique (ajouter un renvoi vers la partie bridge).

L'évaluation se déroule donc selon 3 séquences comme suit :

- 1ère séquence d'au plus 45 minutes : Présentation de l'évaluation, Calcul mental, et un premier ensemble de situations mathématiques des champs Nombres et calculs, Espace et géométrie, Grandeurs et mesures.
- 2ème séquence d'au plus 45 minutes : deux ensembles de situations mathématiques des champs Nombres et calculs, Espace et géométrie, Grandeurs et mesures.
- 3ème séquence d'au plus 45 minutes : un ensemble de situations mathématiques des champs Nombres et calculs, Espace et géométrie, Grandeurs et mesures suivi par un questionnaire de contexte interrogeant les élèves sur leur intérêt vis-à-vis des mathématiques et leurs habitudes à manier des outils numériques.

1.2.2 Les épreuves

Il s'agit pour l'élève de répondre à des items présentés sur tablettes selon différentes interactions. Pour chacune d'entre elles, un tutoriel explicatif est systématiquement proposé avant la passation.

QCM (exemple 1), l'élève clique sur une réponse ce qui valide son choix.

Tableau-série (exemple 2), l'élève doit choisir pour chaque ligne une réponse.

Champ libre (exemple 3 et exemple 4), l'élève dispose du clavier alpha pour apporter sa réponse ou d'un clavier numérique.

Grille magnétique et distributeur d'étiquettes (exemple 5), l'élève dispose d'autant d'étiquettes qu'il veut pour poser et effectuer son opération.

Grille magnétique et outils de dessin (exemple 6), l'élève dispose d'une trousse à outils lui permettant de produire des tracés géométriques, de produire des figures répondant à des grandeurs données, et par remplissage de zone, de répondre à une panoplie d'items dont les fractions d'une surface.

Tâche complexe (exemple 7), l'élève dispose d'une zone de brouillon dans laquelle il peut effectuer ses calculs, noter des étapes intermédiaires et produire

une réponse. Cette interaction permet de récupérer toutes les actions des élèves afin d'analyser leur stratégie de réponse à un problème.

Construction de courbes et de graphiques à bâtons (exemple 8 et exemple 9), l'élève dispose d'une interface spécifique pour positionner les points remarquables d'une courbe (celle-ci se dessine automatiquement) ou pour maîtriser la hauteur de chaque bâton.

Calculatrice cassée (exemple 10), interaction spécifique dans laquelle l'élève doit retrouver un résultat avec les seules touches disponibles. Comme pour la tâche complexe, nous disposons de toutes les actions effectuées par les élèves.

Programmation (exemple 11), l'élève dispose d'une interface de programmation par blocs. Pour l'évaluation de 2019, nous n'avons convoqué que 2 types de blocs : avancer et tourner.

Au total, ce sont 259 interactions de 10 types différents (tableau 2).

Tableau 2 – Nombre d'items par interaction

Type d'interaction	**Nombre d'interactions**
Calculatrice cassée	12
Calcul posé	12
Champ libre	32
Graphique en bâtons	6
Graphique en courbes	6
Outils-géométrie	55
Programmation	6
QCM	110
Tableau-série	11
Tâche complexe	9

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 »).

FIGURE 1 – Exemple 1 : QCM

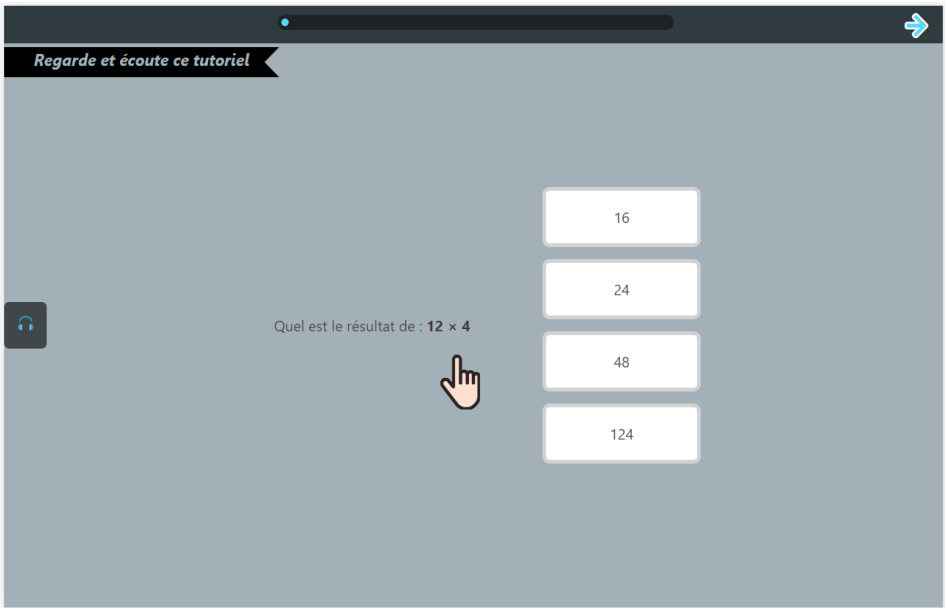


FIGURE 2 – Exemple 2 : Tableau-série

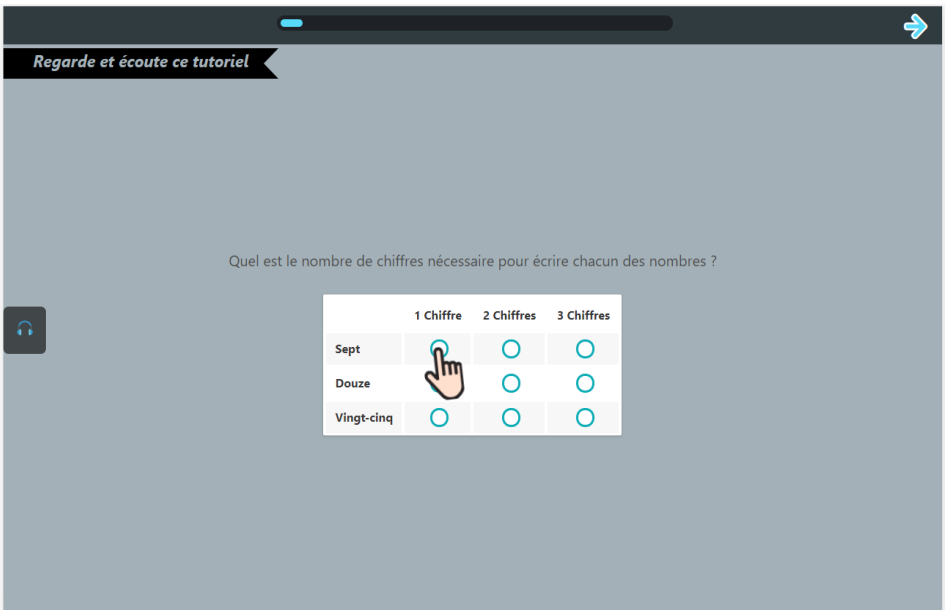


FIGURE 3 – Exemple 3 : Item avec clavier

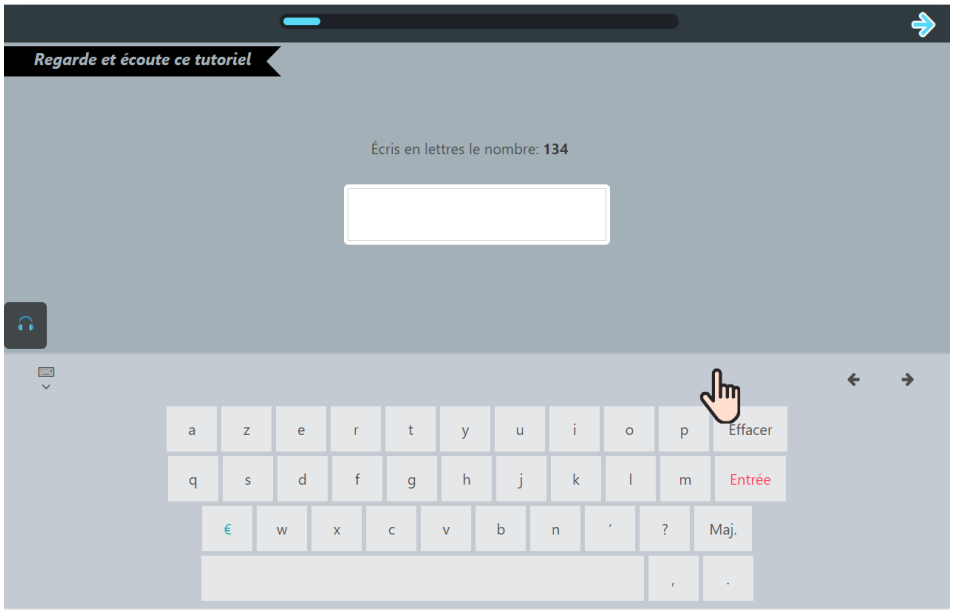


FIGURE 4 – Exemple 4 : Item avec clavier

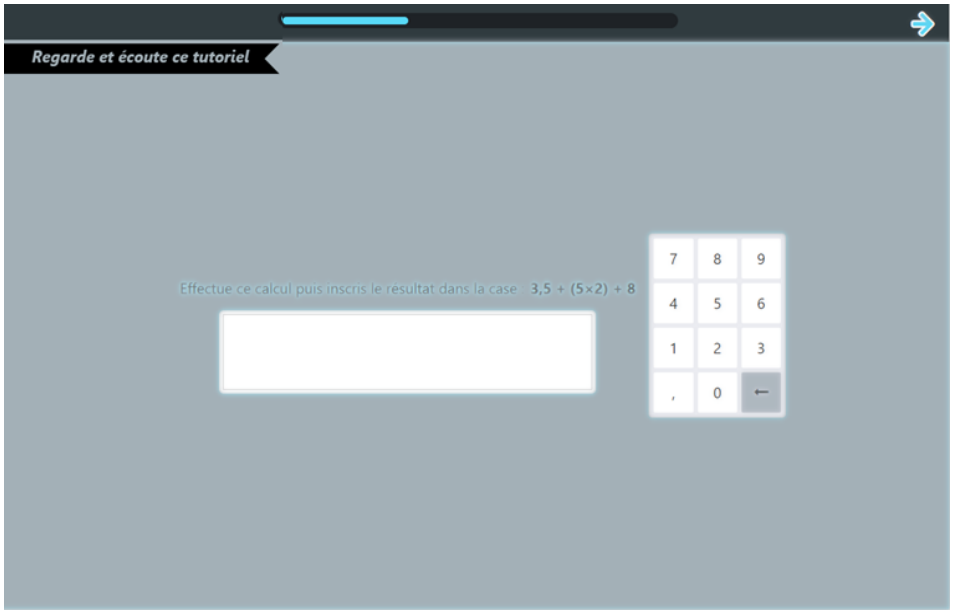


FIGURE 5 – Exemple 5 : Grilles magnétiques et distributeurs d'étiquette

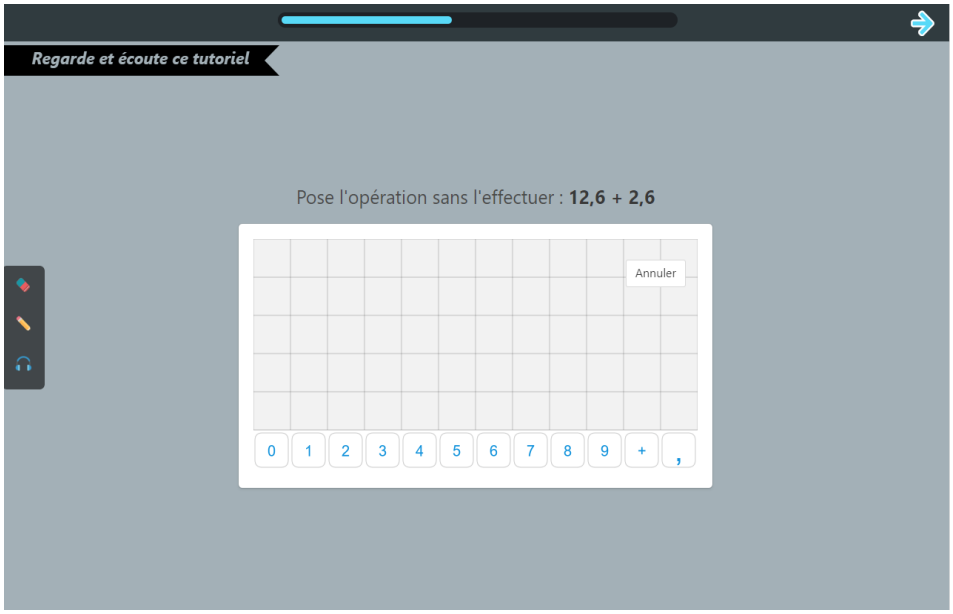


FIGURE 6 – Exemple 6 : Grilles magnétiques et outils de dessin

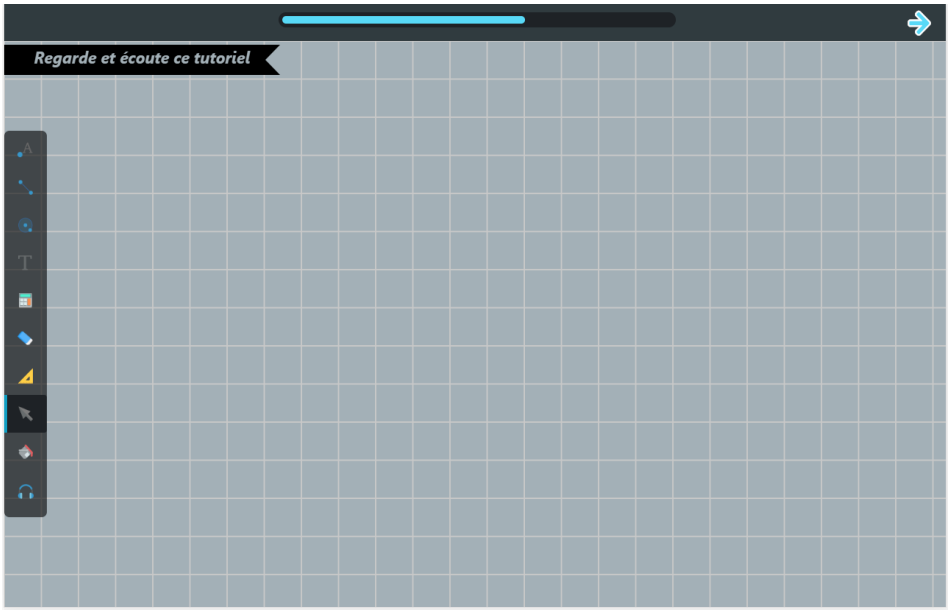


FIGURE 7 – Exemple 7 : Tâche complexe

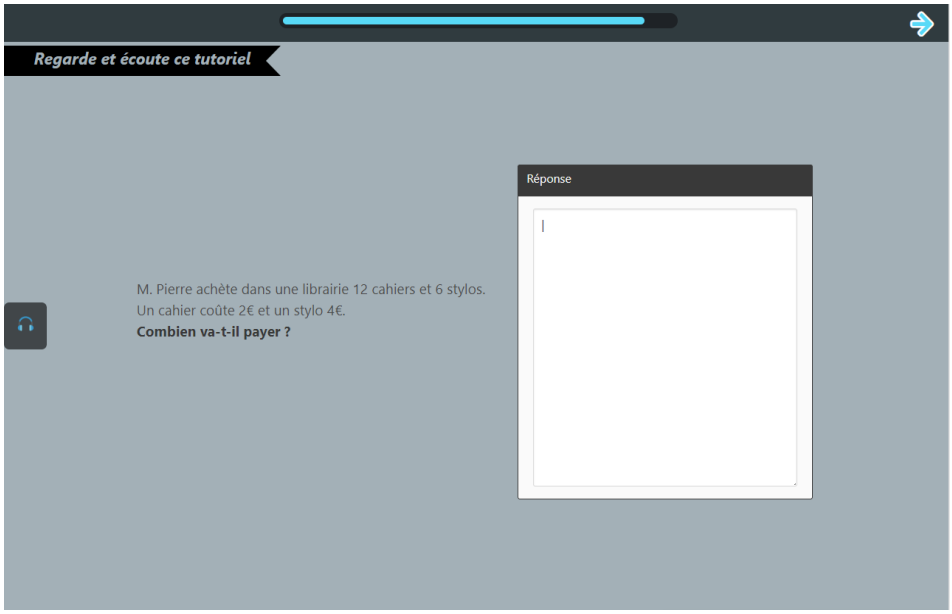


FIGURE 8 – Exemple 8 : Construction de courbes et de graphiques à bâtons

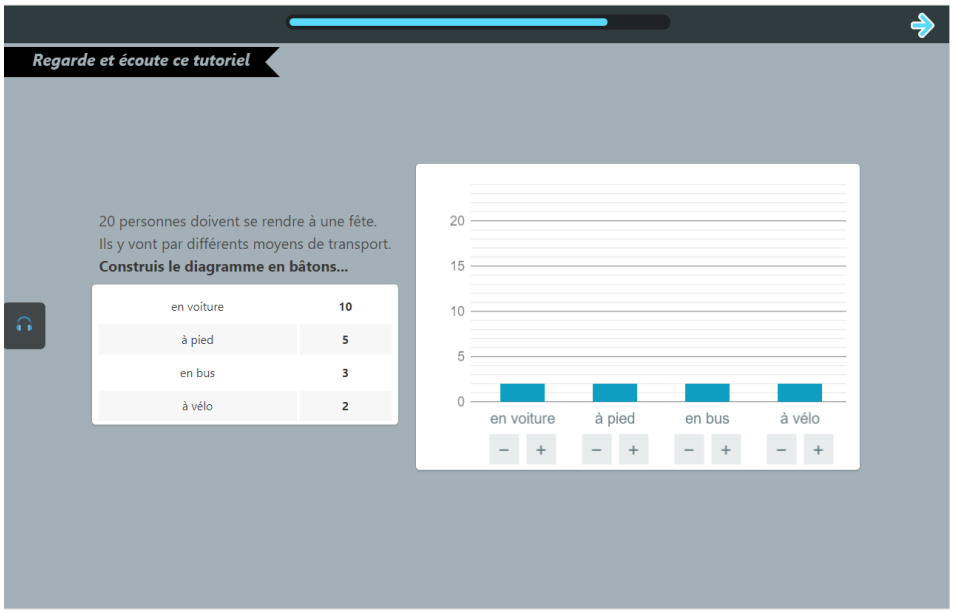


FIGURE 9 – Exemple 9 : Construction de courbes et de graphiques à bâtons

FIGURE 10 – Exemple 10 : Calculatrice cassée

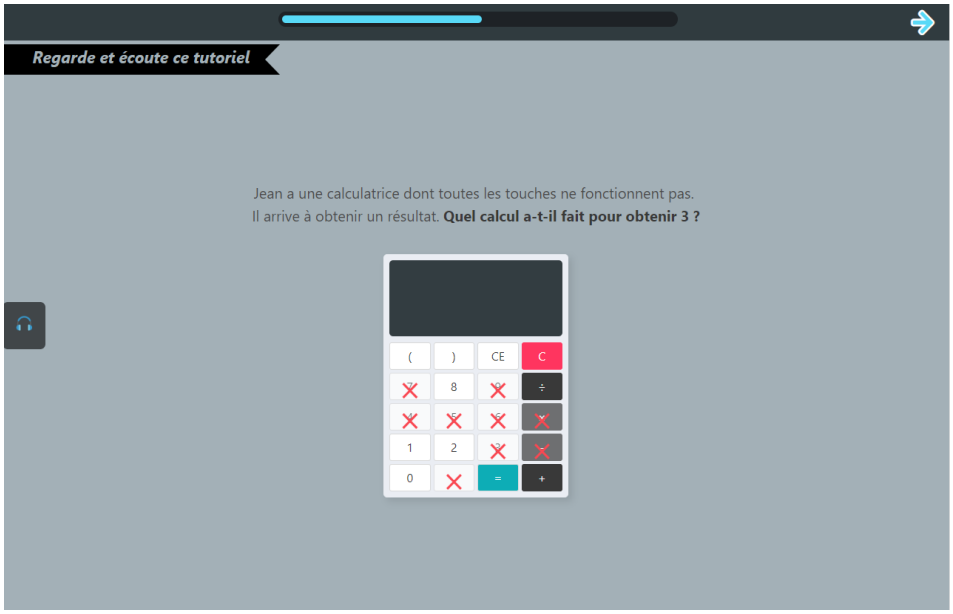
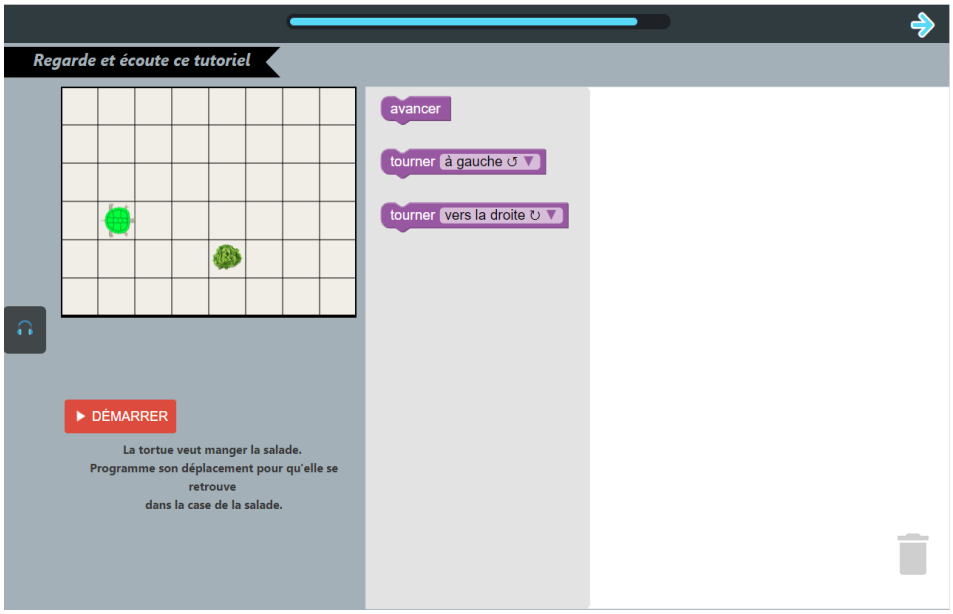


FIGURE 11 – Exemple 11 : Programmation



1.3.1 Élaboration des items

Les items sont le fruit d'un travail collectif des concepteurs, encadré par le chef de projet, 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. L'item est alors soumis à un « cobayage », c'est à dire une passation auprès d'une ou plusieurs classes pour estimer sa difficulté et recueillir les réactions des élèves. Un équilibre de proportion entre les items considérés comme étant de difficulté « facile », « moyenne » ou « difficile » est recherché. Les items des six domaines sont pour certains identiques à ceux proposés en 2008 afin d'assurer une comparabilité de qualité. Les réponses de tous les formats de question sont disponibles à la fin de la passation. Dans le cas de ces tableaux-séries, des seuils statistiques ont été établis pour valider les réponses des élèves.

1.3.2 Constitution des blocs numériques

L'évaluation numérique est également constituée de 13 modules reprenant la méthodologie des cahiers tournants (Tableau 3) intégrant trois temps de passation (séquence).

Tableau 3 – Design de l'évaluation

Modules	Séquence 1		Séquence 2		Séquence 3	
19BME01	Calcul mental	Bloc 4	Bloc 1	Bloc 3	Bloc 8	Questionnaire
19BME02	Calcul mental	Bloc 5	Bloc 2	Bloc 4	Bloc 9	Questionnaire
19BME03	Calcul mental	Bloc 6	Bloc 3	Bloc 5	Bloc 10	Questionnaire
19BME04	Calcul mental	Bloc 7	Bloc 4	Bloc 6	Bloc 11	Questionnaire
19BME05	Calcul mental	Bloc 8	Bloc 5	Bloc 7	Bloc 12	Questionnaire
19BME06	Calcul mental	Bloc 9	Bloc 6	Bloc 8	Bloc 13	Questionnaire
19BME07	Calcul mental	Bloc 10	Bloc 7	Bloc 9	Bloc 1	Questionnaire
19BME08	Calcul mental	Bloc 11	Bloc 8	Bloc 10	Bloc 2	Questionnaire
19BME09	Calcul mental	Bloc 12	Bloc 9	Bloc 11	Bloc 3	Questionnaire
19BME10	Calcul mental	Bloc 13	Bloc 10	Bloc 12	Bloc 4	Questionnaire
19BME11	Calcul mental	Bloc 1	Bloc 11	Bloc 13	Bloc 5	Questionnaire
19BME12	Calcul mental	Bloc 2	Bloc 12	Bloc 1	Bloc 6	Questionnaire
19BME13	Calcul mental	Bloc 3	Bloc 13	Bloc 2	Bloc 7	Questionnaire

1.4 Passation des évaluations

La passation de l'évaluation finale a eu lieu en mai 2019. Comme en 2008 et en 2014, 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 école le directeur ou la directrice est désigné comme responsable de l'évaluation. Son rôle consiste à :

— gérer l'évaluation au sein de l'école ;
— transmettre tous les documents relatifs à l'évaluation aux enseignants ;
— remonter les problèmes rencontrés ;
— récupérer après la passation les résultats de son école.

L'enseignant ou l'enseignante est responsable de la gestion de sa classe. Son rôle consiste à :

— prévenir les enfants et les parents d'élèves de l'évaluation CEDRE ;
— prévenir les parents d'élèves de l'évaluation selon les habitudes de l'école ;
— organiser sa classe en groupes afin que l'administrateur de test travaille dans les meilleures conditions possibles.

Un administrateur de test est désigné par l'académie. C'est la personne qui assure la passation des évaluations auprès des élèves. Son rôle consiste à :

— gérer le planning des passations ;
— effectuer les passations selon le guide ;
— assurer les remontées des tests des élèves.

Chaque séquence était passée en fonction d'un planning déterminé à l'avance par l'administrateur de test. Celui-ci prend en charge un groupe d'élèves (15 au maximum), et présente aux élèves, le matériel et l'évaluation. Il répond aux questions des élèves afin que ceux-ci comprennent bien ce que l'on attend d'eux. Il les met en garde sur le fait que l'évaluation se déroule en autonomie et qu'en aucune façon il ne pourra les aider. Enfin, il assure pendant la passation la résolution de problèmes techniques.

A l'issue de la passation, il transfère les données totalement anonymisées sur le serveur du ministère. Aucun travail de correction n'a été demandé aux écoles.

Les professeurs des écoles des classes concernées ont également dû renseigner un questionnaire de contexte en ligne deux mois avant le début de la passation des épreuves par les élèves. L'anonymat des élèves et des personnels a été respecté.

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. 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 (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.

Mayotte est également exclu

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 2019

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 2019 Mathématiques École

	Établissements	**Elèves**
Ecoles accueillant des élèves de CM2	32 464	847 063
On retire les COM	32 464	847 063
On retire les écoles hors contrat	31 873	842 519
On retire les écoles spécialisées	31 854	842 046
On retire Mayotte	31 737	834 817
On retire les petites écoles (<6 CM2)	29 653	827 280
Base CEDRE CM2	29 653	827 280
On retire TIMSS CM1, CP12, Etape CP et Pilote MATHS	27 780	765 266
Base de tirage Cedre MATHS CM2	27 780	765 266

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

TABLEAU 5 – Répartition dans la base de sondage - CEDRE 2019 Mathématiques 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, 199 écoles ont été sélectionnées.

Tableau 6 – Répartition dans l'échantillon - CEDRE 2019 Mathématiques École

Strate	**Établissements**	**Élèves**
1. Public hors EP	96	2 511
2. EP	68	2525
3. Privé	35	1084
Total	**199**	**6 120**

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 2019.

93.47 % des établissements de l'échantillon ont répondu à l'évaluation (tableau 7). 85.44 % des effectifs attendus ont participé (tableau 8).

Tableau 7 – Non-réponse des établissements - CEDRE 2019 Mathématiques École

Strate	Nb établissements attendus	Nb établissements répondants	% d'établissements répondants
1. Public hors EP	96	92	95.83 %
2. EP	68	62	91.18 %
3. Privé	35	32	91.43 %
Total	**199**	**186**	**93.47 %**

Tableau 8 – Non-réponse des élèves - CEDRE 2019 Mathématiques École

Strate	Nb élèves attendus	Nb élèves répondants	% d'élèves répondants
1. Public hors EP	2 511	2 293	91.32 %
2. EP	2 525	2 075	82.18 %
3. Privé	1 084	861	79.43 %
Total	**6 120**	**5 229**	**85.44 %**

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 les dernières réponses d'un élèves sont vides alors il n'a pas eu le temps de les voir. La non réponse terminale a été étudiée par séquence et par cahier. Toutes les observations en non-réponses terminales sont enlevées des données

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

— 0.964 items pour la séquence 1
— 1.37 items pour la séquence 2
— 0.027 items pour la séquence 3

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 2019 Mathématiques École

Modalité	Variable	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population
Retard	1	48 508	75 282	6.51	9,1
	2	696 104	751 997	93.49	90.9
Sexe	1	377 785	424 395	50.74	51.30
	2	366 828	402 885	49.26	48.70
Strate	1	532 609	574 910	71.53	69.49
	2	109 034	129 294	14.64	15.63
	3	102 970	123 076	13.83	14.88

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 2019 Mathématiques École

Année	Score moyen	Erreur standard
2008	250	2.1
2014	248.6	1.6
2019	231.8	2.3

Pour savoir par exemple si l'évolution entre 2014 et 2019 est significative , il faut calculer la valeur suivante :

$$\frac{\left|\hat{\bar{Y}}_{2014} - \hat{\bar{Y}}_{2019}\right|}{\sqrt{se^2_{\hat{\bar{Y}}_{2019}} + se^2_{\hat{\bar{Y}}_{2014}}}} \tag{2}$$

Entre 2014 et 2019, on obtient une valeur de 5.95 (supérieure à 1.96). Cela signifie que l'évolution du score moyen est 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 2019 Mathématiques École

Année	Groupe <1	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5
2008	2.6	12.4	25.5	31.3	18.2	10
2014	3.7	12.6	26.1	28.6	18.8	10.2
2019	8.8	17	28.6	25.4	14.6	5.6

Tableau 12 – Erreurs standards des répartitions en % dans les groupes de niveaux - CEDRE 2019 Mathématiques École

Année	Groupe <1	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5
2008	0.36	1.07	1.05	0.96	0.95	1.06
2014	0.34	0.62	0.82	0.60	0.67	0.74
2019	0.8	0.79	0.92	0.85	0.99	0.73

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 :

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

Tableau 13 – Effet du plan de sondage - CEDRE 2019 Mathématiques École

Année	Erreur Standard	Erreur SAS	*Design Effect*
2008	2.1	0.77	7.44
2014	1.6	0.56	8.16
2019	2.3	0.87	6.99

Dans le cas d'un sondage en grappes, la précision est dégradée en comparaison d'un sondage aléatoire simple. Cela signifie qu'en 2019, un sondage aléatoire

simple avec un effectif 7 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 & Grégoire, 2002)).

Fidélité

Dans le cadre de la théorie classique des tests, la fidélité (*reliability*) est définie comme la corrélation entre le score observé et le score vrai : le test est fidèle, lorsque l'erreur de mesure est réduite. Une manière d'estimer cette erreur de mesure consiste par exemple à calculer les corrélations entre les différents sous-scores possibles : plus ces corrélations sont élevées, plus le test est dit fidèle [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 & 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.

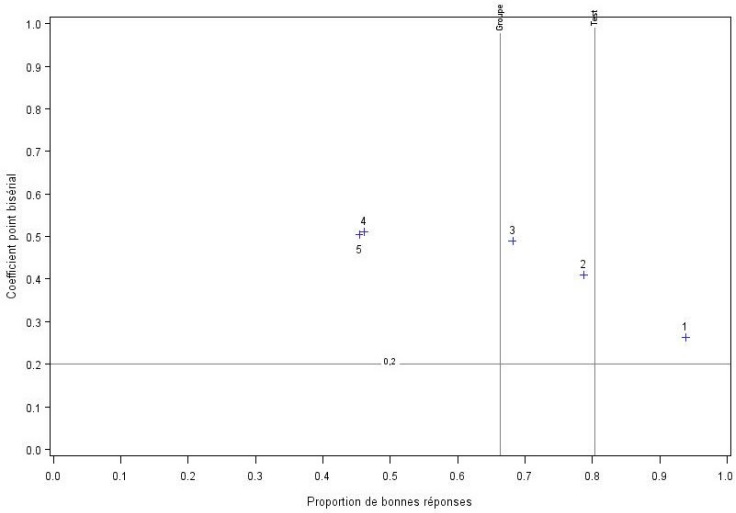
— Non-réponse terminale : l'élève s'est arrêté au cours de l'épreuve, potentiellement en raison d'un manque de temps. Des choix sont effectués pour déterminer le traitement de ces valeurs. Nous considérons que si un élève a passé moins de 50 % d'une séquence, il n'a pas vu la séquence, les valeurs manquantes 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 12). 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 12 – 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 1 items dont l'indice *rbis-point* est trop faible :

— 1 item commun à 2014 et 2019

4 Modélisation

4.1 Méthodologie

4.1.1 Modèle de réponse à l'item

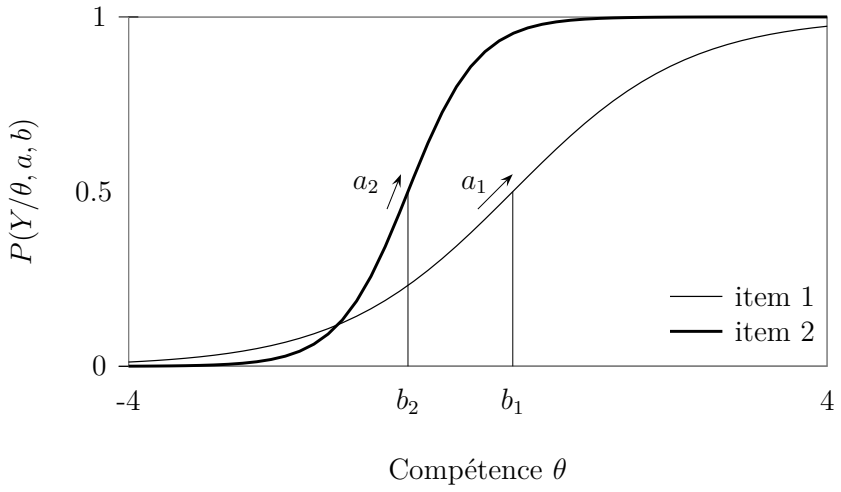
Le modèle de mesure utilisé est un modèle de réponse à l'item à deux paramètres avec une fonction de lien logistique (MRI 2PL) :

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

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

FIGURE 13 – 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 13).

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 2009).

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 :

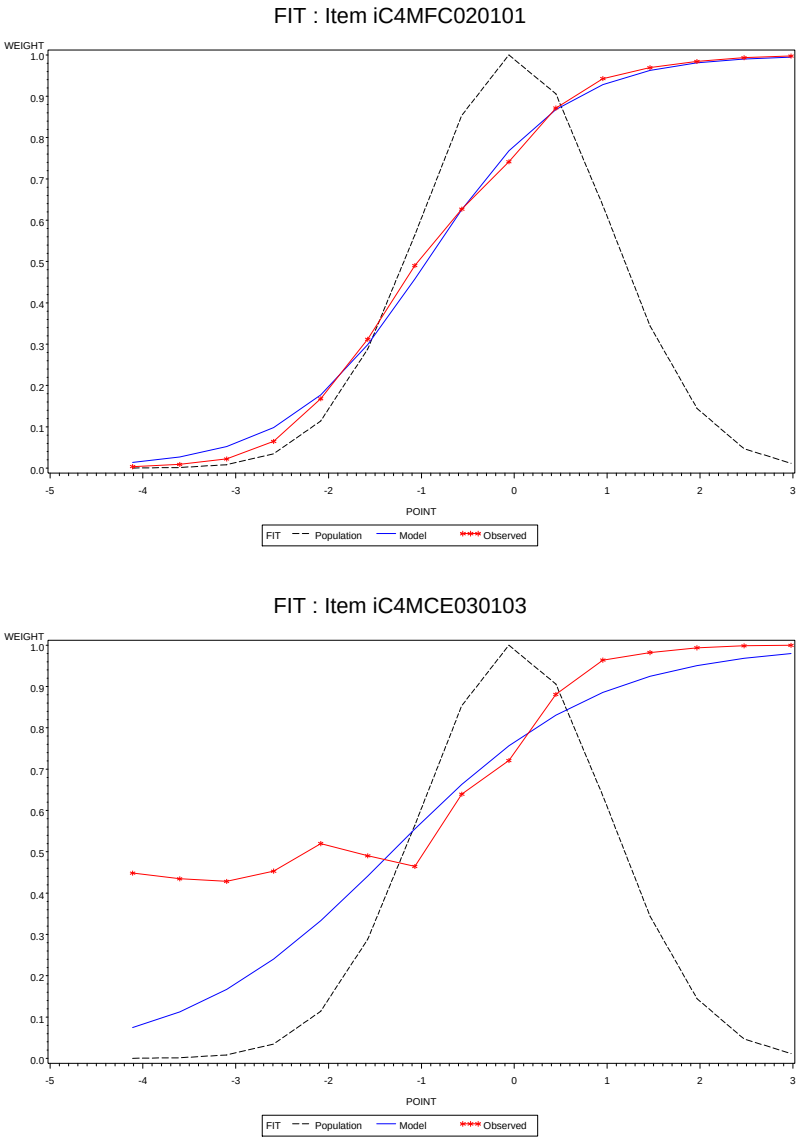
$$\frac{\partial lnL}{\partial \theta_i} + \frac{J}{2I} = 0 \tag{15}$$

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 14). 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 14 – 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.

4.1.6 Transition papier-numérique : étude de comparabilité

Contexte

L'enquête CEDRE est une série temporelle, c'est-à-dire qu'elle a pour objectif premier de pouvoir comparer les performances des élèves de cycle en cycle. Cette caractéristique implique que les différentes générations de l'enquête soient comparables et que le construit testé à chaque cycle soit donc identique.

La DEPP s'est engagée dans la transition d'enquêtes réalisées sur papier vers des enquêtes au format numérique. Cette transition offre de nombreux avantages, aussi bien sur le plan technique qu'en termes de potentialités d'études. Toutefois, la modification du mode d'administration des items aux élèves ne va pas sans poser certaines questions d'ordre méthodologique, qui peuvent mettre en péril la comparabilité des résultats entre les cycles.

Objectifs

Pour assurer cette comparabilité, la Théorie de Réponse à l'Item fournit un ensemble d'outils mé-thodologiques robustes. L'enquête CEDRE s'appuie notamment sur l'utilisation d'items dits d'ancrage, c'est-à-dire repris à l'identique d'un cycle sur l'autre. Ce sont ces items qui permettent de mettre sur la même échelle de performance les résultats des élèves des différents cycles.

Toutefois, la théorie psychométrique impose un certain nombre de contraintes pour que son usage soit pertinent. Une de ces contraintes, essentielle, est l'invariance locale des items. Autrement dit, chaque item doit mesurer le même trait latent, et avec la même précision, pour l'ensemble des sujets, quel que soit son cycle.

Comparer les élèves évalués en 2019 avec ceux des cohortes précédentes ne pouvait donc se faire que sous l'hypothèse que les items restaient parfaitement identiques (notamment en termes de difficulté) lors de leur changement de mode (passage du papier au numérique).

Il était donc nécessaire de construire une cohorte intermédiaire, soumise à une enquête au format mixte, à la fois papier et numérique, servant de "pont" entre les cycles au format papier et les cycles au format numérique.

Méthodologie

L'étude de comparabilité effectuée en 2018 était composée d'items de 2014, repris à l'identique, permettant une comparaison diachronique et leur transposition au format numérique. Ces items étaient répartis en deux cahiers papier et deux modules numériques.

L'échantillon a été construit selon la même méthodologie que pour toutes les enquêtes Cedre, à savoir un sondage par grappes stratifié. Ce tirage est stratifié selon la nature de l'établissement (public, privé, éducation prioritaire), et équilibré selon le sexe et le retard (étant considérés "en retard" les élèves ayant redoublé au moins une fois).

Cette enquête de comparabilité, ou "bridge study", était essentiellement définie par deux choses : d'une part, le design qui a présidé à sa construction, et d'autre part, les hypothèses statistiques qui sous-tendaient ce design.

Lorsqu'on parle de design expérimental, il s'agit à la fois de déterminer le choix des items qui constitueront l'enquête, mais aussi le plan de rotation, c'est-à-dire quels items seront vus par quels élèves.

En ce qui concerne les items, l'ensemble des items d'ancrage ont été repris sous leurs deux formats, papier et numérique. Comme un élève ne peut pas rencontrer deux fois le même item, sans quoi l'effet d'apprentissage serait incontestable, ceux-ci ont été répartis en deux groupes d'items A et B.

Les élèves ont également été répartis dans deux groupes 1 et 2. Les élèves du groupe 1 se sont vu soumettre les items du groupe A au format papier et les items du groupe B au format numérique, tandis que les élèves du groupe 2 se sont vu soumettre les items du groupe A au format numérique et les items du groupe B au format papier. Ainsi, les difficultés des items dans leurs deux modes (papier et numérique) ont pu être calculées distinctement. L'écart de difficulté entre les versions papier et numérique des items (aussi appelé "effet mode") a ensuite été reporté sur la passation CEDRE 2019, afin de la rendre comparable avec les cohortes précédentes.

Précisons que le choix de répartition des élèves au sein de chaque groupe s'est fait au niveau de la classe de manière aléatoire. Cette consigne stricte est régie par la théorie psychométrique. En effet, en construisant deux designs distincts comme nous l'avons fait, rien ne permet a priori de dire que les deux échelles de performance seront équivalentes. Pour cela, il faut que les deux échantillons d'élèves soient représentatifs de la même population.

Effets fixes et effets aléatoires

En théorie, le tirage aléatoire de deux sous-échantillons au sein d'une même population sont également représentatifs de la population. Malheureusement, la méthode de tirage (tirage équilibré par strate) ne garantit pas le côté totalement aléatoire. On peut distinguer les biais subis par le plan de sondage entre effets communs et effets distincts aux deux groupes. Les effets fixes correspondent à la variabilité interclasse, c'est-à-dire aux biais de sondage qui pèsent de manière identique sur les deux sous-échantillons. Les effets distincts corres-pondent à la variabilité intra-classe, c'est-à-dire au biais créé par la scission de chaque classe en deux sous-groupes distincts.

Variabilité interclasse

Les biais de sondage liés à la variabilité interclasse sont identiques à ceux constatés lors des cycles précédents. Ils peuvent être corrigés par une repondération adaptée (calage sur marges), et sont pris en compte lors des calculs de précision. De plus, ils n'impactent pas les deux sous-groupes de l'étude de comparabilité, puisque ceux-ci les portent de manière identique (après repondération). Ils s'apparentent à des effets fixes, et seront donc traités comme tel.

Principes fonctionnels

Cette méthode présente deux avantages. Le tirage aléatoire simple minimise le biais de sélection pour chacun des sous-groupes, qui ne dépend plus que de la taille de chaque sous-groupe. De plus, il limite les effets aléatoires aux individus, rendant ainsi fixes les effets portés par les variables de niveau supérieur (classe, établissement, ...).

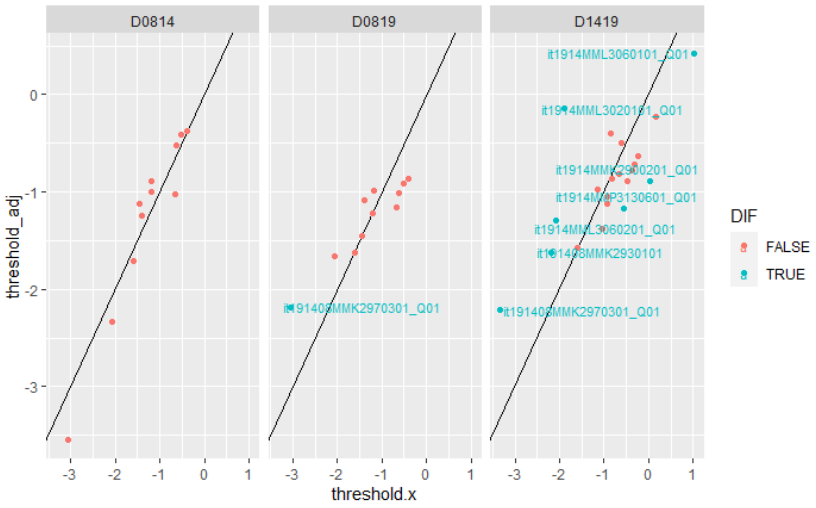
4.2 Résultats

4.2.1 Identification des fonctionnements différentiels d'items (FDI)

8 items ont été éliminés des calculs :

— 7 items pour 2014-2019
— 1 items pour 2008-2014-2019

FIGURE 15 – Comparaison des paramètres de difficulté 2008-2014 2008-2019 et 2014-2019 - (CEDRE Mathématiques 2019 École)



Note de lecture : Les points sont les items d'ancrage. En abscisse figure la valeur des paramètres de difficulté estimés au cycle précédant, et en ordonnée la la valeur des paramètres de difficulté estimés et ajustés pour le cycle. 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 :

- 270 items de 2019
- 18 items d'ancrage 2008-2014
- 12 items d'ancrage 2008-2014-2019

Cela représente 300 items passés par les élèves en tout, dont 270 en 2019.

Après suppression des items présentant un mauvais Rbis, un fonctionnement différentiel ou un mauvais ajustement, il reste :

- 270 items de 2019
- 10 items d'ancrage 2014-2019
- 11 items d'ancrage 2008-2014-2019

291 items sont donc conservés dans l'analyse, dont 270 utilisés dans l'évaluation 2019.

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 2008. Le tableau 14 présente les résultats obtenus.

Tableau 14 – Niveaux de compétences (moyennes des scores et écarts-types) - CEDRE 2019 Mathématiques École

Année	Score moyen	Écart-type
2008	250	50
2014	249	52
2019	232	54

5 Construction de l'échelle

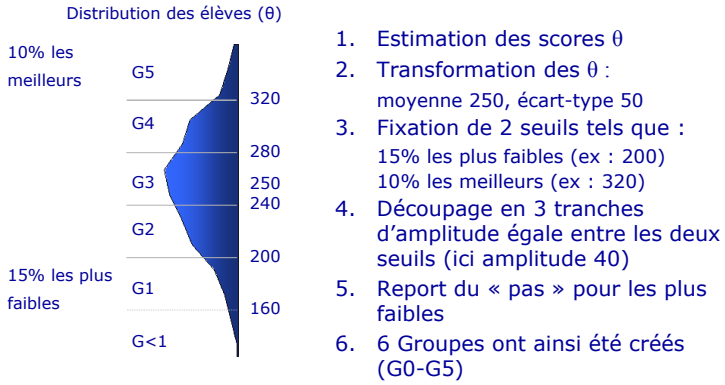
5.1 Méthode

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 Mathématiques 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 2009. Puis, comme le montre la figure 16, 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 16 – Principes de construction de l'échelle



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 la Note d'information n°20-33 – Septembre 2020 [5].

Groupe < 1 (8,8 % des élèves)

Ces élèves peuvent répondre ponctuellement à quelques items simples. Les réussites observées se fondent essentiellement sur des situations ayant trait à la vie courante – « estimer la taille d'objets usuels » –, à des pratiques scolaires ancrées – « repérer si une figure est symétrique par rapport à un axe vertical » –, donner une réponse par lecture directe – « lecture d'un nombre sur une règle graduée ». Ils maîtrisent très peu de compétences ou de connaissances exigibles en fin d'école primaire.

Groupe 1 (17 % des élèves)

Ces élèves ont des connaissances des nombres qui leur permettent la mise en œuvre d'opérations (additions et soustractions), néanmoins l'utilisation des retenues dans la soustraction n'est pas acquise. La construction du nombre en classes n'est pas solide. Les réussites observées s'appuient essentiellement sur des automatismes scolaires. Certains de ces mécanismes leur permettent de réussir des problèmes additifs directs qui ne nécessitent qu'une seule étape pour leur résolution.

5. Cedre 2008-2014-2019 Mathématiques en fin d'école : des résultats en baisse Note d'information - N°20.33 – septembre 2020, Louis-Marie Ninnin, Jean-Marc Pastor.

Groupe 2 (28,6 % des élèves)

Ces élèves ont des connaissances sur les nombres entiers qui leur permettent de réussir un certain nombre de problèmes de type additif, voire soustractif sans étape intermédiaire. La réussite à quelques items éloignés des pratiques scolaires montre les premiers signes de transfert de compétences et l'adoption d'une stratégie pour résoudre une situation nouvelle. Ils traitent l'information et sont capables de retrouver un résultat correct, mais ils échouent quand il s'agit de produire une réponse en autonomie.

Groupe 3 (25,4 % des élèves)

Ces élèves ont une connaissance solide des nombres entiers et une première connaissance stable des nombres décimaux. Ils ont une pratique du calcul avec les quatre opérations et manient des notions comme le double et la moitié d'un nombre, le tiers d'un entier et le multiple de trois. S'ils sont capables de résoudre des problèmes simples de proportionnalité, leurs acquis restent fragiles lorsqu'il s'agit de produire en autonomie une réponse. Ils font preuve d'une première culture mathématique et d'une bonne connaissance du vocabulaire spécifique en géométrie. Ces élèves maîtrisent une grande partie des connaissances et des compétences exigibles à la fin de l'école.

Groupe 4 (14,6 % des élèves)

Ces élèves sont capables de faire un traitement fin de l'information, de réussir des problèmes utilisant la proportionnalité lorsque les mesures de longueur sont explicites, et lorsque la relation additive est évidente. Ils sont capables de mettre en œuvre des stratégies évoluées, de résoudre des problèmes complexes et de produire des réponses en autonomie pour des situations peu fréquentes en classe. Ces élèves ont acquis la majeure partie des connaissances et des compétences exigibles en fin d'école.

Groupe 5 (5,6 % des élèves)

Ces élèves manient habilement les concepts mathématiques de fin d'école primaire. Cela leur permet de gérer une masse d'information plus grande, de sélectionner les éléments utiles de ceux accessoires, d'imaginer des solutions et de produire un travail en autonomie. Ces élèves font preuve d'expertise dans les compétences et connaissances de fin d'école primaire, ils maîtrisent tous les champs du programme et font preuve de capacité d'abstraction, de rigueur et de précision. Ces élèves ont acquis l'ensemble des connaissances et des compétences exigibles en fin d'école primaire.

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 (Eteve, Dos Santos, & 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 2019, 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 2008, 2014 et 2019, 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 2008-2014-2019)

Indice moyen école	Année	Score moyen	Écart type
1er quart	2008	237	48
1er quart	2014	229	48
1er quart	2019	**207**	51
2e quart	2008	247	51
2e quart	2014	242	50
2e quart	2019	**226**	51
3e quart	2008	254	49
3e quart	2014	258	53
3e quart	2019	**237**	51
4e quart	2008	262	48
4e quart	2014	265	52
4e quart	2019	257	50

Note de lecture : en 2019, le score moyen des élèves appartenant au quart des écoles les plus défavorisées (1er quart) dimunue de 22 points par rapport à 2014. 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-

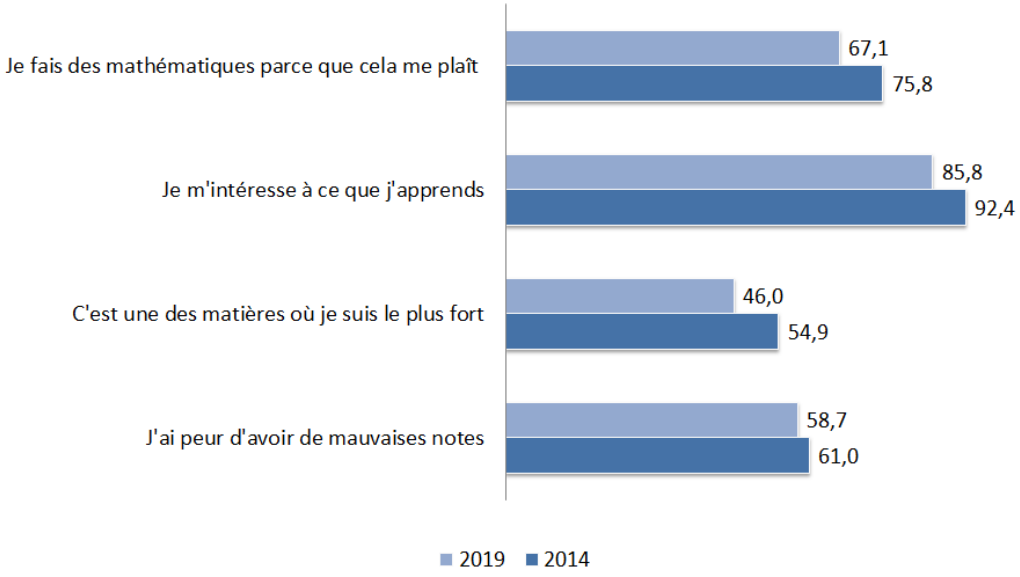
sionnelle des parents mais aussi le parcours de l'élève (raccourcissement de cycle ou maintien dans un cycle, orientation retenue etc.).

6.3 Motivation des élèves face à la situation d'évaluation

L'enquête Cedre a permis de collecter les opinions des élèves sur le contexte dans lequel ils étudient. Les questions concernent deux axes régulièrement investigués dans les études au sujet des mathématiques : l'intérêt et le plaisir pour la discipline et la perception de soi. L'analyse de l'évolution des réponses pointe une baisse de l'adhésion aux affirmations proposées Figure : 17.

D'une manière générale, les élèves sont moins nombreux à déclarer faire des mathématiques par plaisir (67,1 % en 2019 contre 75,8 % en 2014). En effet, ils s'intéressent moins aux apprentissages en mathématiques (85,8 % contre 92,4 %) et sont moins nombreux à attendre les séances avec impatience (54,7 % contre 77 %). Leur perception « positive » de leur niveau est aussi en baisse : 46 % estiment que c'est en mathématiques qu'ils sont les plus forts contre 54,9 % en 2014. Enfin, 58,7 % des élèves ont peur d'avoir des mauvaises notes, proportion comparable à celle observée en 2014 (61 %).

FIGURE 17 – Des élèves moins nombreux à faire des mathématiques par plaisir



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.

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.

Eteve, Y., Dos Santos, R., & Ninnin, L.-M. (2019). CEDRE 2008-2014-2019 - mathématiques en fin d'école : des résultats en baisse. *Note d'information*, *16*.

Laveault , D., & Grégoire, J. (2002). *Introduction aux théories des tests en psychologie et en sciences de l'éducation* (J.-M. De Ketele, J.-M. Van der Maren, & M. Duru-Bellat, Eds.). De Boeck.

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