# CEDRE

Cycle des Évaluations Disciplinaires Réalisées sur Échantillons

# Rapport technique

Histoire-géographie et enseignement moral et civique 2017

École

Auteurs :

Marion LE CAM
Louis-Marie NINNIN
Jean-Marc PASTOR
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 et de la jeunesse

Juin 2019

# 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

Les documents de référence pour la construction des items sont les programmes officiels en vigueur à partir de la rentrée scolaire 2015-2016. Ceux-ci introduisent un découpage des cycles différents des périodes antérieures. Le cycle 3 se terminant désormais en fin de sixième ; nous nous sommes basés sur les répartitions et les progressions en fin de CM2 pour conforter l'évaluation en fin d'école.

Les finalités de l'histoire et de la géographie sont rappelées dans l'introduction des programmes :

" L'histoire et la géographie aident l'élève à construire une première intelligence du temps historique et de la diversité des espaces transformés par l'activité humaine. Elles lui donnent les références culturelles nécessaires pour que le monde des hommes commence à prendre sens pour lui. "

L'évaluation CEDRE en fin d'école en histoire, géographie et enseignement moral et civique a pour objectif de faire le point des connaissances et des compétences des élèves tant sur le plan des savoirs que des savoir-faire d'une part, et de mesurer l'évolution de ces connaissances et compétences entre 2006, 2012 et 2017 d'autre part. Les connaissances et compétences telles qu'elles sont définies dans les programmes officiels constituent le cadre de cette évaluation.

### 1.2.2 Tableau de compétences

En 2017 comme en 2012 et 2006, une même grille de " compétences " est utilisée pour les trois champs de cette évaluation : histoire, géographie et enseignement moral et civique. La plupart des intitulés sont restés identiques entre les trois points de mesure ce qui permet d'assurer la comparaison entre les prises d'informations.

Tableau 2 – Tableau des compétences

| | **Histoire** | **Géographie** | **Enseignement moral et civique** |
|---|---|---|---|
| Connaître la notion | NH | NG | NE |
| Savoir dater | DH | | DE |
| Localiser | | SG | |
| Utiliser les outils de la discipline | UH | UG | UE |
| Produire | PH | PG | PE |
| Lexique | LH | LG | LE |
| Accéder et s'approprier des informations | XH | XG | XE |
| Comprendre le sens général d'un document | CH | CG | CE |
| Apprécier la pertinence de l'information | IH | IG | IE |
| Justifier ses choix | JH | JG | JE |
| Mettre en relation deux documents | RH | RG | RE |
| Mobiliser des connaissances | MH | MG | ME |

*Note de lecture :* Chaque compétence est définie par une lettre (N pour notion, D pour dater,..) chaque matière est définie de façon identique (H pour histoire, G pour géographie et E pour enseignement moral et civique). Le croisement de la matière et de la compétence est matérialisé par un couple de lettres, associé à l'item produit.

Toutefois des aménagements sont effectués pour répondre au programme en cours par la prise en compte :

- De l'enseignement moral et civique qui se substitue à l'éducation civique ;
- De nouvelles habiletés à développer sur des supports numériques :
  - " Accéder et s'approprier des informations " qui correspond à la navigation nécessaire pour atteindre l'information dans un site Internet ;
  - " Apprécier la pertinence de l'information " qui correspond au tri et aux choix à faire devant des listes de résultat de recherche.
- Du socle commun de connaissances, de compétences et de culture pour lequel avons apposé une nouvelle grille de lecture associant chaque item de l'évaluation à son ou ses domaines du socle (tableau 3). A noter qu'il ne s'agit pas de quantifier la proportion des élèves qui valident le socle mais bien d'éclairer sur les compétences du socle mises en jeu dans l'évaluation CEDRE.

Tableau 3 – Domaines du socle

| Domaine 1 | Domaine 2 | Domaine 3 | Domaine 5 |
|---|---|---|---|
| Comprendre le sens général d'un document | Trouver, sélectionner et exploiter des informations dans une ressource numérique | Formation de la personne et du citoyen | Localiser |
| Extraire les informations pertinentes pour répondre à une question | Identifier un document et savoir pourquoi il doit être identifié. | | Situer des lieux les uns par rapport aux autres |
| S'approprier le lexique spécifique de la discipline | Méthode d'investigation (Se poser des questions, poser des questions, formuler des hypothèses, justifier, vérifier...) | | Notion d'échelle |
| Réaliser des graphiques | Identifier une ressource numérique | | Caractériser des espaces |
| Ecrire en histoire et géographie | Connaître différents systèmes d'information | | Utiliser des cartes, des photos aériennes, des graphiques ... |

L'évaluation CEDRE se décline en deux volets : le premier concerne des items présentés dans un cahier d'évaluation ; le second correspond à celui passé sur ordinateur.

— 1er volet : évaluation " papier " : il s'agit pour l'élève de répondre à des unités présentées sur un support " papier ". Une unité se compose d'un ou de plusieurs documents que l'élève devra utiliser pour répondre aux questions. Les unités sont regroupées dans des blocs ; les blocs dans des cahiers. C'est l'évaluation de référence pour le descriptif des acquis des élèves ainsi que l'instrument permettant la comparaison entre les différents points de mesure.

— 2nd volet : évaluation " numérique " : il s'agit pour l'élève de répondre à des unités présentées sur un support " électronique ". Plusieurs cas se présentent. L'unité propose un document à l'élève et des questions afférentes à ce dernier. Le document peut prendre la forme d'une image ou d'un texte ; d'un ensemble de textes et d'images regroupés dans une liseuse, d'un ensemble de documents regroupés dans un site, d'une vidéo ou de cartes interactives spécialement programmées pour l'évaluation. Les unités sont " ventilées " dans des situations ; les situations dans treize modules. A noter que cette partie n'est pas prise en compte pour la comparaison temporelle mais fait l'objet d'une analyse spécifique.

### 1.2.3 Calendrier de l'évaluation

La passation de l'évaluation CEDRE histoire, géographie et enseignement moral et civique en fin d'école a eu lieu en mai 2017. La méthodologie observée par la DEPP en matière d'évaluation des élèves s'étale sur trois années selon le schéma suivant :

Tableau 4 – Les étapes de la réalisation de l'évaluation

| 1ère année | 2ème année | 3ème année |
|---|---|---|
| 2015-2016 | 2016-2017 | 2017-2018 |
| Étape 1 | Étape 2 | Étape 3 |
| Expérimentation | Évaluation | Analyse et publication des résultats |
| Passation mai 2016 | Passation mai 2017 | |
| Objectif :<br>Prise en compte du cahier des charges.<br><br>Création d'items<br>- Items sur support papier<br>- Items sur support numérique<br><br>Expérimentation effectuée auprès d'un échantillon représentatif d'élèves. | Objectif :<br>Analyse de l'expérimentation.<br><br>Construction de l'évaluation à partir d'une sélection d'items<br><br>- Items validés après l'expérimentation ;<br>- Items d'ancrage (repris des évaluations précédentes).<br><br>Evaluation effectuée auprès d'un échantillon représentatif d'élèves. | Objectif :<br>Analyse des premiers résultats<br><br>Publications :<br>- Note d'information<br>- Repères et références statistiques (RERS)<br>- Etat de l'école |

## 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.

Un équilibre de proportion entre les items considérés comme étant ”faciles”, ”moyennement faciles” ou ”difficiles” est recherché. Afin d'assurer une comparabilité, 132 items sur les 211 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 saisies de manière automatisée et donne lieu à un regroupement ultérieur de leurs proposition. 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. partie ” Analyse des items ”).

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

**Objectifs**

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

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

électronique.

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

**Principes fonctionnels**

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

### 1.3.2 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 histoire-géographie 2017, 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 l'histoire et la géographie.

L'évaluation CEDRE 2017 est constituée de 13 cahiers tournants intégrant un ensemble de 13 blocs d'évaluations contenant des items de 2006 et 2012 repris à l'identique pour assurer une comparaison diachronique et de nouveaux items qui ont fait l'objet d'une expérimentation en 2016. Pour garantir la qualité de la comparaison avec 2006 et 2012, notamment en termes de passation des épreuves, l'évaluation de 2017 s'appuie sur 211 items dont 132 d'ancrage soit 62%.

Tableau 5 – 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 |

## 1.4 Passation des évaluations

La passation de l'évaluation finale a eu lieu en mai 2017. Comme en 2012, 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 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 sélectionnés participent. Le choix de sondages par grappe est motivé par la facilité de gestion. En effet, le fait de sélectionner tous les élèves d'une école permet d'éviter de mettre en place des procédures de tirage au sort d'élèves une fois les écoles tirées.

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

### 2.1.2 Redressement de la non réponse : calage sur marges

Comme toute enquête réalisée par sondage, les évaluations des élèves sont exposées à la non-réponse. Bien que les taux de retour soient élevés, il est nécessaire de tenir compte de la non-réponse dans les estimations car celle-ci n'est pas purement aléatoire (par exemple, la non-réponse est plus élevée chez les élèves en retard). Afin de la prendre en compte, un calage sur marges est effectué à l'aide de la macro CALMAR, également disponible sur le site Internet de l'INSEE. La méthode de calage sur marges consiste à modifier les poids de sondage $d_i$ des répondants de manière à ce que l'échantillon ainsi repondéré soit représentatif de certaines variables auxiliaires dont on connait les totaux sur la population (Sautory, 1993). C'est une méthode qui permet de corriger la non-réponse mais également d'améliorer la précision des estimateurs. En outre, elle a pour avantage de rendre cohérents les résultats observés sur l'échantillon pour ce qui concerne des informations connues sur l'ensemble de la population.

Les nouveaux poids $w_i$, calculés sur l'échantillon des répondants $S'$, vérifient l'équation suivante pour les $K$ variables auxiliaires sur lesquelles porte le calage :

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

Ils sont obtenus par minimisation de l'expression $\sum_{i \in S'} d_i G(\frac{w_i}{d_i})$ où G désigne une fonction de distance, sous les contraintes définies dans l'équation 1.

**Tirage après élimination de la base des échantillons précédemment tirés**

La situation est la suivante : un échantillon d'établissements a été sélectionné pour participer à une évaluation ; un deuxième échantillon doit être tiré pour une autre évaluation. Nous souhaitons éviter que des établissements soient interrogés deux fois. Il s'agit donc de gérer le non-recouvrement entre les échantillons et d'assurer également un tirage du deuxième échantillon. Nous nous concentrons ici sur le non-recouvrement des échantillons mais notons qu'une approche plus générale incluant un taux de recouvrement non nul (pour permettre des analyses croisées entre enquêtes) dans un cadre de tirage équilibré est en cours de développement avec une application à des données issues d'évaluations standardisées (Christine & Rocher, 2012).

**Notations**

Un échantillon $S_1$ a été tiré. Il est connu et les probabilités d'inclusion des établissements $\pi_j^1$ sont également connues. On souhaite alors tirer un échantillon $S_2$ dans la population $U$ avec les probabilités $\pi_j^2$, 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 $\pi_j^{2/S_1}$ 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 2017**

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

**Base de sondage**
Le tableau 6 présente les exclusions dans la population ciblée.

Tableau 6 – Exclusions pour la base de sondage - CEDRE 2017 Histoire-géographie et enseignement moral et civique École

| | **Établissements** | **Elèves** |
|---|---|---|
| Ecoles accueillant des élèves de CM2 | 32 494 | 838 137 |
| On retire les COM | 32 494 | 838 137 |
| On retire les écoles hors contrat | 32 096 | 834 707 |
| On retire les écoles spécialisées | 32 073 | 834 129 |
| On retire les DOM | 31 129 | 795 336 |
| On retire les petites écoles (<6 CM2) | 28 977 | 787 587 |
| Base CEDRE CM2 | 28 977 | 787 587 |
| On retire les échantillons précédents | 28 329 | 765 480 |
| Base de tirage CEDRE CM2 HG | 28 329 | 765 480 |

Le tableau 7 présente la répartition de la population ciblée selon le secteur d'enseignement.

Tableau 7 – Répartition dans la base de sondage - CEDRE 2017 Histoire-géographie et enseignement moral et civique École

| **Strate** | **Établissements** | **Élèves** |
|---|---|---|
| 1. Public hors EP | 21 725 | 558 699 |
| 2. EP | 3 075 | 109 453 |
| 3. Privé | 4 177 | 119 435 |
| **Total** | **28 977** | **787 587** |

**Échantillon**

Le tableau 8 présente la répartition de l'échantillon selon le secteur d'enseignement. Au total, 222 écoles ont été sélectionnées.

Tableau 8 – Répartition dans l'échantillon - CEDRE 2017 Histoire-géographie et enseignement moral et civique École

| **Strate** | **Établissements** | **Élèves** |
|---|---|---|
| 1. Public hors EP | 166 | 4 236 |
| 2. EP | 24 | 846 |
| 3. Privé | 32 | 960 |
| **Total** | **222** | **6 042** |

## 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 2017.

93.7 % des établissements de l'échantillon ont répondu à l'évaluation (tableau 9). 90.1 % des éffectifs attendus ont participé (tableau 10).

Tableau 9 – Non-réponse des établissements - CEDRE 2017 Histoire-géographie et enseignement moral et civique École

| Strate | Nb établissements attendus | Nb établissements répondants | % d'établissements répondants |
|---|---|---|---|
| 1. Public hors EP | 166 | 156 | 94 % |
| 2. EP | 24 | 20 | 83.3 % |
| 3. Privé | 32 | 32 | 100 % |
| **Total** | **222** | **208** | **93.7 %** |

Tableau 10 – Non-réponse des élèves - CEDRE 2017 Histoire-géographie et enseignement moral et civique École

| Strate | Nb élèves attendus | Nb élèves répondants | % d'élèves répondants |
|---|---|---|---|
| 1. Public hors EP | 4 236 | 3 857 | 91.1 % |
| 2. EP | 846 | 638 | 75.4 % |
| 3. Privé | 960 | 947 | 98.6 % |
| **Total** | **6 042** | **5 442** | **90.1 %** |

### 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 :
— 5.1 items pour la séquence 1
— 5.4 items pour la séquence 2

On considère que :
— 101 élèves n'ont pas vu la séquence 1, dont :
  — 90 n'ont répondu à aucun items de la séquence
  — 11 ont répondu à moins de 50 % de la séquence
— 129 élèves n'ont pas vu la séquence 2, dont :
  — 110 n'ont répondu à aucun items de la séquence
  — 19 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 56 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 11 – Comparaison entre les marges de l'échantillon et les marges dans la population - CEDRE 2017 Histoire-géographie et enseignement moral et civique École

| Modalité | Variable | Marge échantillon | Marge population | Pourcentage échantillon | Pourcentage population |
|---|---|---|---|---|---|
| Retard | 1 | 62 233.79 | 71 578 | 7.90 | 9.09 |
| | 2 | 725 353.31 | 716 009 | 92.10 | 90.91 |
| Sexe | 1 | 400 221.39 | 401 131 | 50.82 | 50.93 |
| | 2 | 387 365.71 | 386 456 | 49.18 | 49.07 |
| Strate | 1 | 558 699.05 | 558 699 | 70.94 | 70.94 |
| | 2 | 109 453 | 109 453 | 13.90 | 13.90 |

## 2.5 Précision

L'erreur standard (*se*) peut être calculée sur le score moyen de chaque année (tableau 12).

Tableau 12 – Scores moyens et erreurs standard associées - CEDRE 2017 Histoire-géographie et enseignement moral et civique École

| Année | Score moyen | Erreur standard |
|---|---|---|
| 2006 | 250 | 1.54 |
| 2012 | 250.7 | 1.43 |
| 2017 | 252.2 | 1.74 |

Pour savoir par exemple si l'évolution entre 2012 et 2017 est significative , il faut calculer la valeur suivante :

$$\frac{\left|\hat{\bar{Y}}_{2017} - \hat{\bar{Y}}_{2012}\right|}{\sqrt{se^2_{\hat{\bar{Y}}_{2017}} + se^2_{\hat{\bar{Y}}_{2012}}}} \tag{2}$$

Entre 2012 et 2017, on obtient une valeur de 0.63 (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 13 et 14).

Tableau 13 – Répartitions en % dans les groupes de niveaux - CEDRE 2017 Histoire-géographie et enseignement moral et civique École

| Année | Groupe <1 | Groupe 1 | Groupe 2 | Groupe 3 | Groupe 4 | Groupe 5 |
|---|---|---|---|---|---|---|
| 2006 | 2.1 | 12.9 | 27.7 | 29.1 | 18.2 | 10 |
| 2012 | 2.6 | 12.3 | 27.8 | 29.4 | 16.5 | 11.5 |
| 2017 | 2.6 | 12.8 | 26.4 | 27.3 | 18.8 | 12.1 |

Tableau 14 – Erreurs standards des répartitions en % dans les groupes de niveaux - CEDRE 2017 Histoire-géographie et enseignement moral et civique École

| Année | Groupe <1 | Groupe 1 | Groupe 2 | Groupe 3 | Groupe 4 | Groupe 5 |
|---|---|---|---|---|---|---|
| 2006 | 0.2 | 0.7 | 0.8 | 0.6 | 0.7 | 0.7 |
| 2012 | 0.3 | 0.7 | 0.7 | 0.7 | 0.6 | 0.6 |
| 2017 | 0.3 | 0.7 | 0.8 | 0.7 | 0.7 | 0.9 |

### *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 15 – Effet du plan de sondage - CEDRE 2017 Histoire-géographie et enseignement moral et civique École

| Année | Erreur Standard | Erreur SAS | *Design Effect* |
|---|---|---|---|
| 2006 | 1.54 | 0.57 | 2.69 |
| 2012 | 1.43 | 0.69 | 2.07 |
| 2017 | 1.74 | 0.71 | 2.44 |

Dans le cas d'un sondage en grappes, la précision est dégradée en comparaison d'un sondage aléatoire simple. Cela signifie qu'en 2017, un sondage aléatoire simple avec un effectif 2.44 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 $\alpha$ 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 $\alpha$ 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 $\sigma_S$ 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 $\phi$, 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 $\tau_j$, 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.

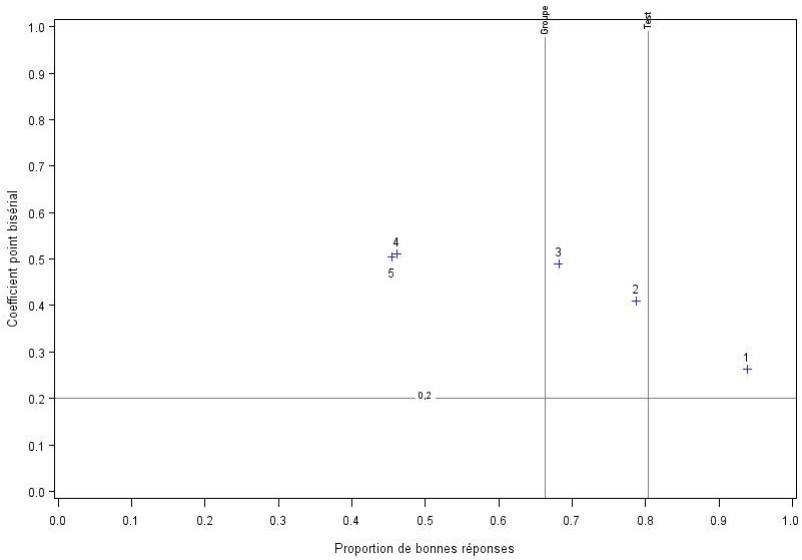
— Non-réponse terminale : l'élève s'est arrêté au cours de l'épreuve, potentiellement en raison d'un manque de temps. Des choix sont effectués pour déterminer le traitement de ces valeurs. Nous considérons que si un élève a passé moins de 50 % d'une séquence, il n'a pas vu la séquence, les valeurs manquantes 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 56 items dont les indices *rbis-point* sont trop faibles :

- 29 items de 2006
- 2 items communs à 2006 et 2012
- 4 items communs à 2006, 2012 et 2017
- 15 items de 2012
- 6 items de 2017

# 4 Modélisation

## 4.1 Méthodologie

### 4.1.1 Modèle de réponse à l'item

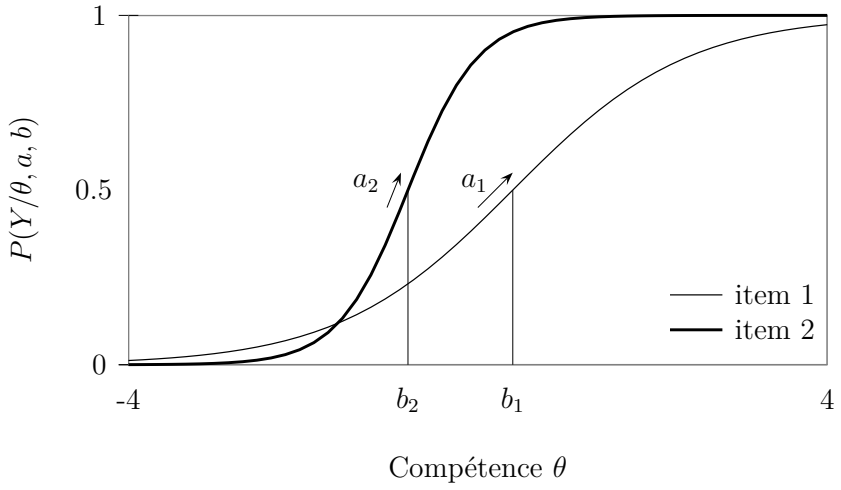
Le modèle de mesure utilisé est un modèle de réponse à l'item à deux paramètres avec une fonction de lien logistique (MRI 2PL) :

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

où la probabilité $P_{ij}$ que l'élève $i$ réussisse l'item $j$ est fonction du niveau de compétence $\theta_i$ 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 $\theta$ 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 $\theta$ pour les données du premier cycle (en l'occurrence, moyenne de 250 et écart-type de 50 pour l'année 2006).

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 $\theta$. A niveau de compétence égal, deux items donnés ne sont pas corrélés : seule la compétence $\theta$ explique la corrélation entre deux items. Cette hypothèse est ainsi liée à l'hypothèse d'unidimensionnalité de $\theta$ (cf, Rocher, 2013).

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

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 $\theta$ 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 $\theta_i$, 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 :

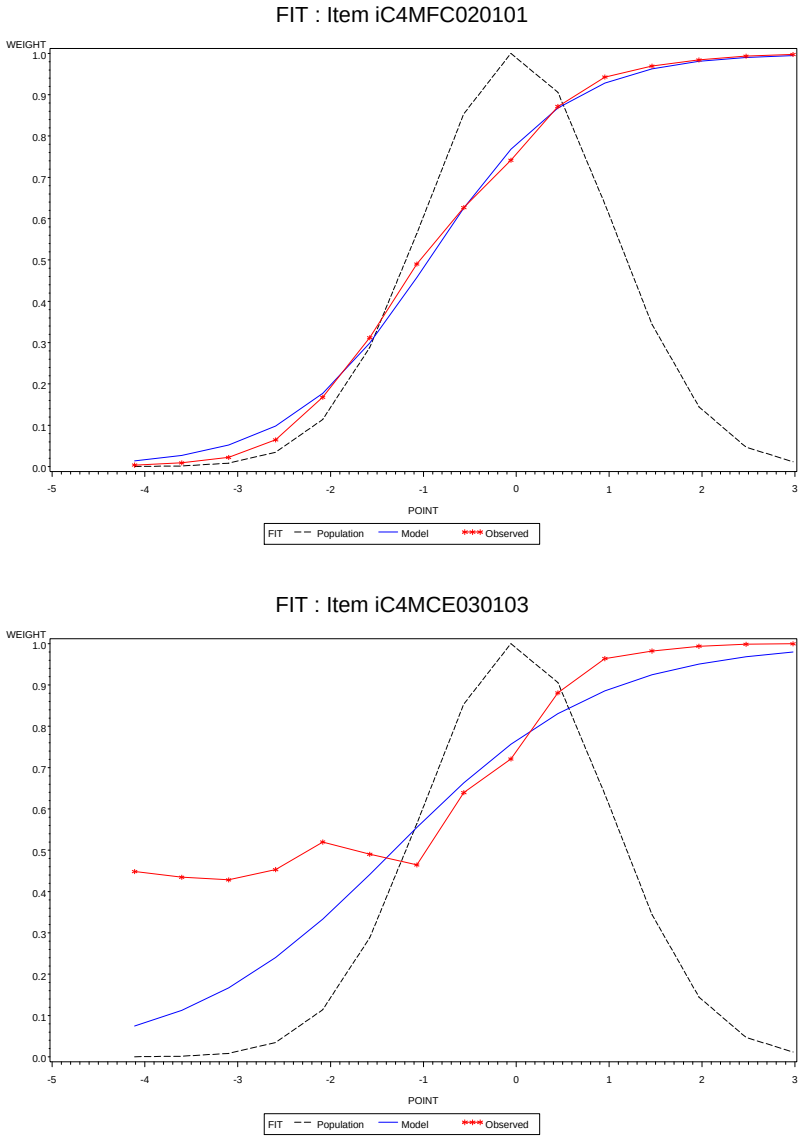
$$\frac{\partial lnL}{\partial \theta_i} + \frac{J}{2I} = 0 \tag{15}$$

### 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 théorique avec les proportions de réussite de groupes d'élèves. Plus généralement, nous pouvons écrire les résidus de la manière suivante :

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

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 observé à cet item pour 15 groupes d'élèves de niveaux de compétence croissants. Enfin, la courbe en pointillée représente la distribution des niveaux de compétence.
Clairement, l'ajustement du modèle est excellent pour l'item présenté en haut. Il est très mauvais pour celui du bas.

Les carrés des résidus suivent typiquement une loi du $\chi^2$. 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 $\beta$ 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 $\theta$.

L'information moyenne du test pour un élève de compétence $\theta$ est la somme de l'information apporté par chaque item pour $\theta$. La courbe d'information du test est tracée pour un ensemble de valeurs de $\theta$. 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 $\theta$ des élèves est estimé.

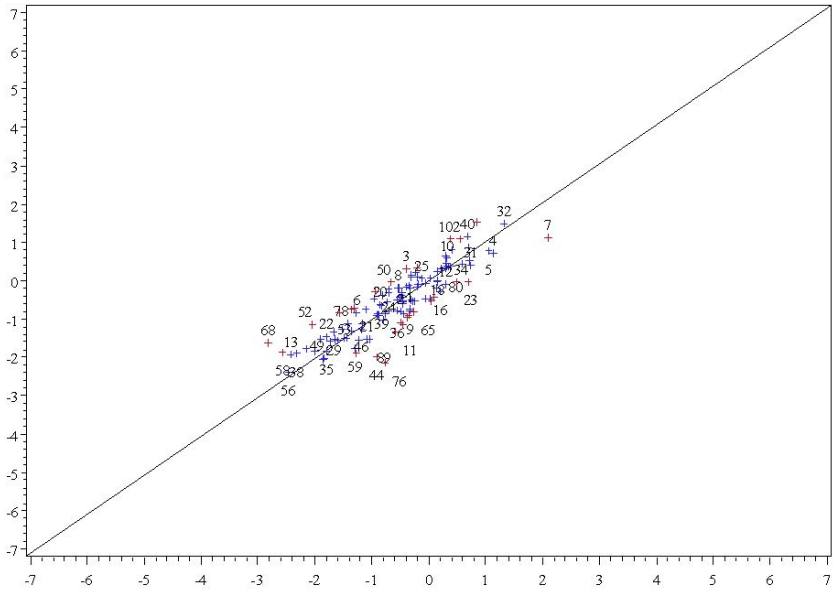
## 4.2 Résultats

### 4.2.1 Identification des fonctionnements différentiels d'items (FDI)

32 items ont été éliminés des calculs :

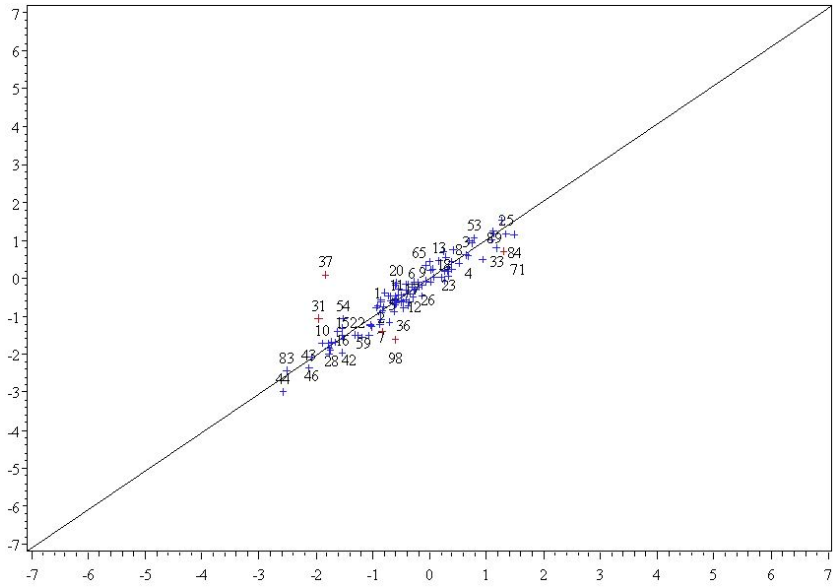
— 27 items pour 2006-2012
— 5 items pour 2012-2017

Figure 4 – Comparaison des paramètres de difficulté 2006-2012 - (CEDRE Histoire-géographie et enseignement moral et civique 2017 École)



Note de lecture : Les points sont les items. En abscisse figure la valeur des paramètres de difficulté estimés en 2006, et en ordonnée la la valeur des paramètres de difficulté estimés et ajustés pour l'année 2012. Les items présentant un FDI apparaissent en bleu.

Figure 5 – Comparaison des paramètres de difficulté 2012-2017 - (CEDRE Histoire-géographie et enseignement moral et civique 2017 École)



Note de lecture : Les points sont les items. En abscisse figure la valeur des paramètres de difficulté estimés en 2012, et en ordonnée la la valeur des paramètres de difficulté estimés et ajustés pour l'année 2017. Les items présentant un FDI apparaissent en bleu.

### 4.2.2 Identification des items présentant un mauvais ajustement (FIT)

3 items ont été éliminés des calculs :

— 3 items pour 2006

### 4.2.3 Bilan de l'analyse des items

En considérant l'ensemble des items sur les 3 années, il y avait au départ :

— 185 items de 2006
— 58 items de 2012
— 79 items de 2017
— 39 items d'ancrage 2006-2012
— 48 items d'ancrage 2012-2017
— 93 items d'ancrage 2006-2012-2017

Cela représente 502 items passés par les élèves en tout, dont 220 en 2017.

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

— 153 items de 2006

— 43 items de 2012
— 73 items de 2017
— 10 items d'ancrage 2006-2012
— 43 items d'ancrage 2012-2017
— 89 items d'ancrage 2006-2012-2017

411 items sont donc conservés dans l'analyse, dont 205 utilisés dans l'évaluation 2017.

## 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 $\theta$ des élèves. Afin de lever l'indétermination du modèle, la moyenne des $\theta$ a été fixée à 250 et leur écart-type à 50, pour l'échantillon de 2006. Le tableau 16 présente les résultats obtenus.

Tableau 16 – Niveaux de compétences (moyennes des scores et écarts-types) - CEDRE 2017 Histoire-géographie et enseignement moral et civique École

| Année | Score moyen | Écart-type |
|---|---|---|
| 2006 | 250 | 50 |
| 2012 | 250.7 | 51.9 |
| 2017 | 252.2 | 52.5 |

# 5 Construction de l'échelle

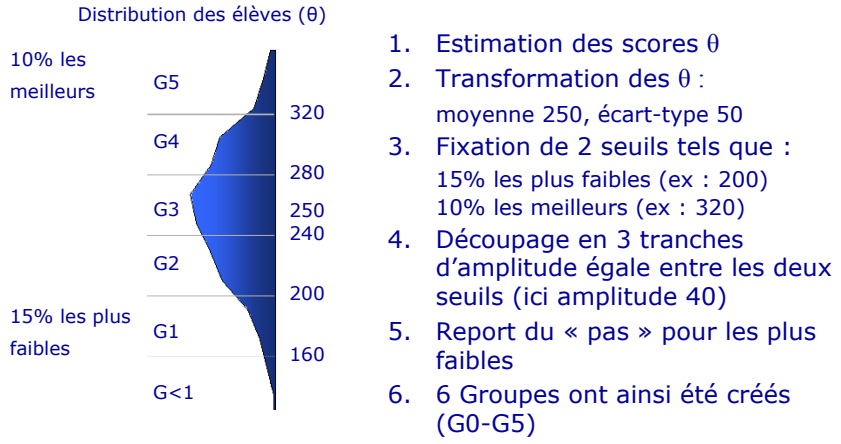
## 5.1 Méthode

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 Histoire-géographie et enseignement moral et civique 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 2006. 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

A 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 $< 1$ (2,6 % des élèves)

Bien que capables de répondre très ponctuellement à quelques questions, ces élèves ne maîtrisent aucune des connaissances et des compétences attendues en fin d'école primaire. Des réussites sont constatées sur des sujets très souvent abordés à l'école, comme la préhistoire, et sur la capacité à repérer des caractéristiques explicites dans les documents proposés. On observe très peu de réussites dans le domaine du traitement de l'information, exceptées quelques mises en relation entre une image et un texte lorsque celui-ci est simple. Ces élèves utilisent leurs compétences pour " mémoriser " (utilisation ponctuelle de la mémoire de rappel) et " analyser des images " (identification d'une image connue).

### Groupe 1 (12,8 % des élèves)

Ces élèves obtiennent des réussites pour des connaissances mettant en jeu des documents iconographiques largement exploités en classe. Ils utilisent leurs compétences pour " mémoriser " (utilisation de la mémoire de rappel) et " analyser des images " (identification d'images connues, utilisation du descriptif de l'image pour répondre à des questions demandant des caractéristiques d'un personnage, d'un lieu ou d'un événement). Ils font preuve de l'acquisition de quelques connaissances ponctuelles établies essentiellement en histoire.

**Groupe 2 (26,4 % des élèves)**

Ces élèves sont capables d'utiliser les légendes accompagnant les documents proposés. Ils répondent aux questions mettant en jeu des consignes simples et qui ont trait à des documents courts facilement identifiables et dont le prélèvement d'informations est simple à réaliser. Ils butent sur les supports exigeant une lecture prolongée. Ils utilisent leur compétence pour " mémoriser " (utilisation plus efficiente de la mémoire de rappel), "analyser des textes " (utilisation du paratexte des documents), " analyser des images " (description de l'image permettant de dégager des éléments) et " établir des liens " (mise en relation de deux documents par la liaison terme à terme). En histoire, ils reconnaissent des personnages, des événements, des lieux sans pour autant les associer aux bonnes périodes ou sans pouvoir les relier à la notion historique. En géographie, ils savent décrire un paysage mais ne sont pas capable de l'interpréter.

**Groupe 3 (27,3 % des élèves)**

Ces élèves utilisent leur capacité de mémorisation pour réactiver des notions vues antérieurement, notamment avec les documents patrimoniaux. Ils montrent des habiletés à produire des inférences simples sur les informations explicites, aisément repérables, contenues dans deux documents différents en prenant appui sur un vocabulaire courant. Ils peinent cependant à effectuer un traitement pertinent sur ces informations pour les interpréter et accéder à leur pleine compréhension. Ils utilisent leurs compétences pour " mémoriser " (mémoire de rappel accrue), "analyser des textes " (avec des textes courts et utilisant un vocabulaire simple), " analyser des images " (description et quelques inférences), " établir des liens " (mise en relation de deux documents) et " conceptualiser " (notions existantes mais inégalement construites). En histoire, ces élèves ont des connaissances au sujet des documents iconographiques "patrimoniaux". En géographie, ils répondent correctement aux items mettant en oeuvre une lecture de carte. Ils se sont appropriés les outils de base tels les légendes et les tableaux de données à entrées multiples.

**Groupe 4 (18,8 % des élèves)**

Ces élèves manifestent des compétences robustes en histoire et en géographie. Ils sont capables de mises en relation complexes entre deux documents pour accéder à une synthèse leur permettant de répondre à des questions exigeant un très bon niveau de maîtrise de la langue. Ces élèves utilisent leurs compétences pour " mémoriser " (mémoire de rappel efficiente), "analyser des textes " (inférences systématiques), " analyser des images " (les interpréter), " établir des liens " (mise en relation de deux documents ou plus), " conceptualiser " (notions connues dans l'ensemble des domaines). En histoire, ils identifient les personnages majeurs et leurs rôles, les lieux des faits historiques et ils peuvent les situer dans leur

période. En géographie, ils repèrent l'évolution des paysages. Ils sont capables de localiser.

**Groupe 5 (12,1 % des élèves)**

Ces élèves affichent des réussites systématiques aux questions demandant une réponse construite . Ils savent utiliser les connaissances historiques qu'ils ont assimilées pour catégoriser les documents proposés notamment dans le domaine de l'histoire des arts. Ils savent interpréter et intégrer des informations contenues dans des documents hétérogènes. Ils utilisent leurs compétences pour " mémoriser " (mémoire de rappel efficiente), "analyser des textes " (analyse interprétative), " analyser des images " (interprétation), " établir des liens " (mise en relation de plusieurs documents), " conceptualiser " (solides notions dans tous les domaines). En histoire, ces élèves situent correctement les faits historiques sur la frise chronologique repère et ils les associent aux événements significatifs de la période concernée. En géographie, ils savent, à partir d'éléments, catégoriser une notion et la replacer dans son contexte.

## 5.3 Exemples d'items

### 5.3.0.a Item caractéristique des groupes < 1 et 1

Figure 7 – Exemple groupe < à 1

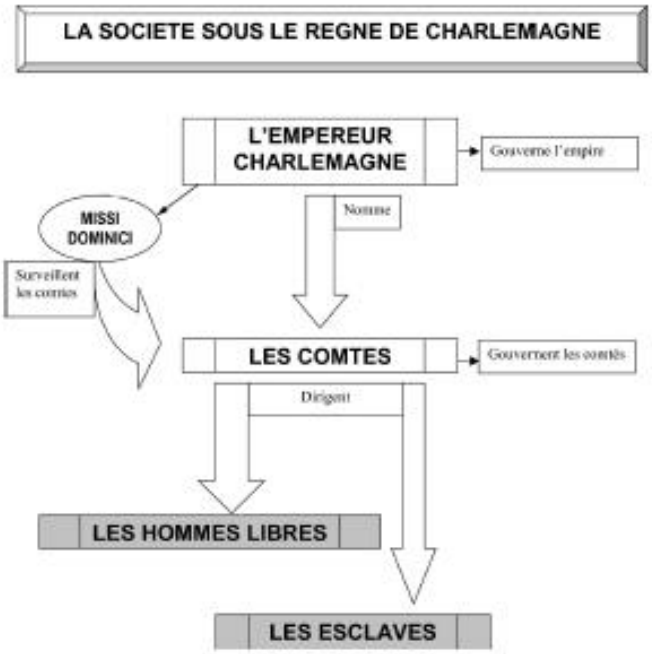

| | Vrai | Faux |
|---|---|---|
| En 1910, il y a des voitures et des immeubles. | ☐1 | ☐2 |
| Depuis 1910, on a construit de grands immeubles. | ☐1 | ☐2 |
| En 1992, on reconnaît la rue qui existait en 1910. | ☐1 | ☐2 |
| En 1971, la circulation est plus importante. | ☐1 | ☐2 |

Les élèves des groupes <1 et 1 répondent ponctuellement à quelques questions. Ils utilisent leurs compétences pour " mémoriser " (utilisation ponctuelle de la mémoire de rappel) et " analyser des images " (identification d'une image connue). Ils réussissent principalement les items relatifs à des périodes maintes fois vues à l'école. Ils prélèvent des informations immédiatement accessibles dans une image mais restent le plus souvent uniquement dans l'aspect descriptif. Le document proposé à l'observation des élèves est constitué de 3 images de la ville de Bobigny à différentes périodes de son histoire. En dessous, sont notés les dates des prises de vue. Dans cet item, les élèves doivent effectuer quatre allers-retours entre le document et la série de propositions. Nous rappelons que pour obtenir le crédit (la bonne réponse), il faut que les élèves répondent correctement à toutes les propositions. Ces élèves font preuve d'une observation correcte des images les conduisant à répondre correctement aux propositions.

### 5.3.0.b Item caractéristique du groupe 2

Le document proposé est un schéma de l'organisation de la société sous le règne de Charlemagne. Les élèves doivent lire celui-ci en comprenant le sens des flèches. Comme pour l'exemple précédent, le format de l'item est caractérisé par une

Figure 8 – Exemple groupe 2

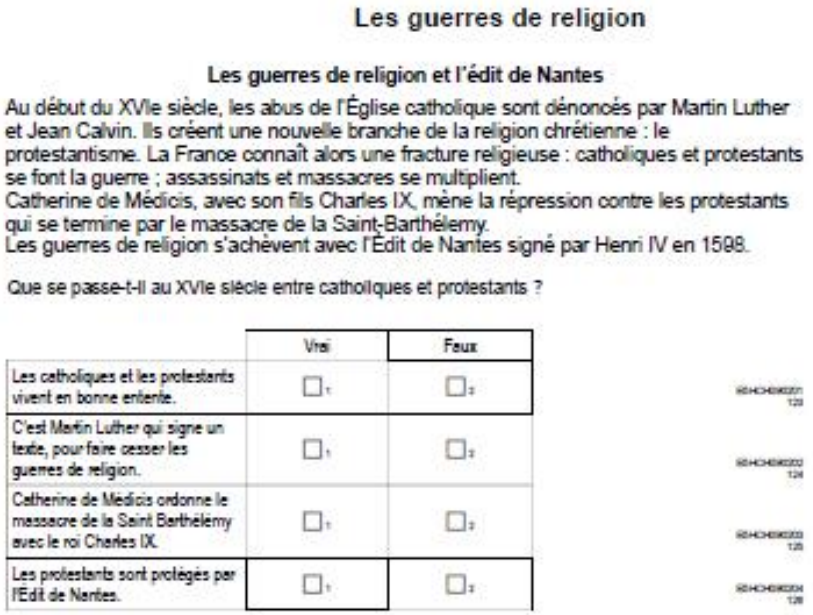



| | Vrai | Faux |
|---|---|---|
| Les comtes surveillent l'Empire. | ☐1 | ☐2 |
| Les comtes gouvernent les comtés. | ☐1 | ☐2 |
| Charlemagne gouverne l'Empire. | ☐1 | ☐2 |
| Les Missi Dominici gouvernent les comtés. | ☐1 | ☐2 |

suite de quatre propositions. Les élèves doivent répondre correctement à toutes les propositions pour obtenir le crédit (bonne réponse). Nous rappelons que ces élèves sont capables d'utiliser les légendes accompagnant les documents proposés. Ils répondent aux questions mettant en jeu des consignes simples et qui ont trait à des documents courts facilement identifiables et dont le prélèvement d'informations est simple à réaliser. Dans le cas présent, ils montrent qu'ils sont capables de comprendre le fonctionnement de ce document et d'effectuer des prélèvements les conduisant à une réponse exacte.

#### 5.3.0.c Item caractéristique du groupe 3

Figure 9 – Exemple groupe 3



**Les guerres de religion**

**Les guerres de religion et l'édit de Nantes**

Au début du XVIe siècle, les abus de l'Église catholique sont dénoncés par Martin Luther et Jean Calvin. Ils créent une nouvelle branche de la religion chrétienne : le protestantisme. La France connaît alors une fracture religieuse : catholiques et protestants se font la guerre ; assassinats et massacres se multiplient.
Catherine de Médicis, avec son fils Charles IX, mène la répression contre les protestants qui se termine par le massacre de la Saint-Barthélemy.
Les guerres de religion s'achèvent avec l'Édit de Nantes signé par Henri IV en 1598.

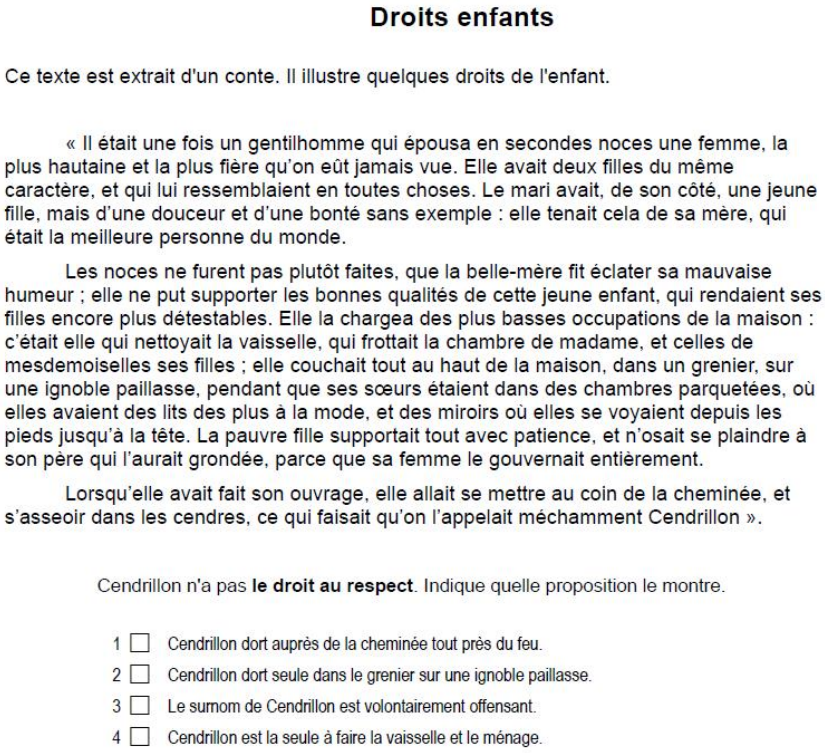
Que se passe-t-il au XVIe siècle entre catholiques et protestants ?

| | Vrai | Faux |
|---|---|---|
| Les catholiques et les protestants vivent en bonne entente. | ☐1 | ☐2 |
| C'est Martin Luther qui signe un texte, pour faire cesser les guerres de religion. | ☐1 | ☐2 |
| Catherine de Médicis ordonne le massacre de la Saint Barthélemy avec le roi Charles IX. | ☐1 | ☐2 |
| Les protestants sont protégés par l'Édit de Nantes. | ☐1 | ☐2 |

Le document proposé est un texte décrivant la situation de la France au XVIe siècle. Nous rappelons que les élèves du groupe 3 montrent des habiletés à produire des inférences simples sur les informations explicites, aisément repérables, contenues dans deux documents différents en prenant appui sur un vocabulaire courant. Dans ce cas, il faut que les élèves aient lu attentivement le texte. Comme dans les exemples précédents, le format de l'item est caractérisé par une suite de quatre propositions. Les élèves doivent répondre correctement à toutes les propositions pour obtenir le crédit (bonne réponse). En observant les propositions, nous remarquons qu'elles nécessitent de la part des élèves des inférences.

#### 5.3.0.d Item caractéristique du groupe 4

Figure 10 – Exemple groupe 4



**Droits enfants**

Ce texte est extrait d'un conte. Il illustre quelques droits de l'enfant.

« Il était une fois un gentilhomme qui épousa en secondes noces une femme, la plus hautaine et la plus fière qu'on eût jamais vue. Elle avait deux filles du même caractère, et qui lui ressemblaient en toutes choses. Le mari avait, de son côté, une jeune fille, mais d'une douceur et d'une bonté sans exemple : elle tenait cela de sa mère, qui était la meilleure personne du monde.

Les noces ne furent pas plutôt faites, que la belle-mère fit éclater sa mauvaise humeur ; elle ne put supporter les bonnes qualités de cette jeune enfant, qui rendaient ses filles encore plus détestables. Elle la chargea des plus basses occupations de la maison : c'était elle qui nettoyait la vaisselle, qui frottait la chambre de madame, et celles de mesdemoiselles ses filles ; elle couchait tout au haut de la maison, dans un grenier, sur une ignoble paillasse, pendant que ses sœurs étaient dans des chambres parquetées, où elles avaient des lits des plus à la mode, et des miroirs où elles se voyaient depuis les pieds jusqu'à la tête. La pauvre fille supportait tout avec patience, et n'osait se plaindre à son père qui l'aurait grondée, parce que sa femme le gouvernait entièrement.

Lorsqu'elle avait fait son ouvrage, elle allait se mettre au coin de la cheminée, et s'asseoir dans les cendres, ce qui faisait qu'on l'appelait méchamment Cendrillon ».

Cendrillon n'a pas **le droit au respect**. Indique quelle proposition le montre.

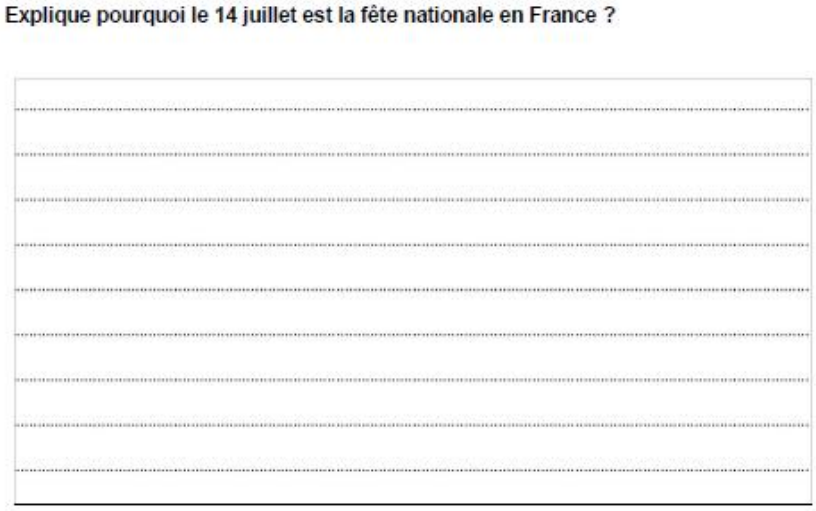
1 ☐ Cendrillon dort auprès de la cheminée tout près du feu.
2 ☐ Cendrillon dort seule dans le grenier sur une ignoble paillasse.
3 ☐ Le surnom de Cendrillon est volontairement offensant.
4 ☐ Cendrillon est la seule à faire la vaisselle et le ménage.

Le document proposé aux élèves est un extrait de comte. Il correspond à un texte relativement long et de lecture difficile. Même si la plupart des élèves connaissent ce comte, les questions posées nécessitent une lecture fine de ce dernier et du recul. Par ailleurs la question posée faisant référence aux droits de l'enfant, les élèves doivent activer les connaissances qu'ils ont dans ce domaine de l'éducation civique. Toutefois, cette notion du " droit au respect " est quotidienne à l'école et sans doute connue " en actes ". Le format de question est une QCM. Les élèves ont le choix entre quatre propositions dont une seule est la bonne. C'est au niveau quatre de l'échelle que cette question est réussie ; nous rappelons que ces élèves manifestent des compétences robustes en histoire et en géographie. Ils sont capables de mises en relation complexes entre deux documents pour accéder à une synthèse leur permettant de répondre à des questions exigeant un très bon niveau de maîtrise de la langue. Ils sont capables de prendre du recul par rapport aux informations traitées et de faire le lien avec leurs connaissances

et la notion abordée.

#### 5.3.0.e Item caractéristique du groupe 5

Figure 11 – Exemple groupe 5




Trois questions portaient sur les commémorations du 14 juillet, du 8 mai et du 11 novembre. A chaque fois nous demandions aux élèves d'expliquer pourquoi ces dates étaient commémorées. Le format de l'item était un champ libre et nous attendions une réponse rédigée par les élèves. Ces trois items sont représentatifs des élèves du groupe 5. Nous rappelons que ces élèves affichent des réussites systématiques aux questions demandant une réponse construite. Ils savent utiliser les connaissances historiques qu'ils ont assimilées pour catégoriser les documents proposés notamment dans le domaine de l'histoire des arts. Ces élèves font preuve de connaissances et de capacité de restitution écrite de celles-ci afin d'obtenir une réponse correcte.

# 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 (Ninnin & Pastor, 2018).

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 2006, 2012 et 2017, 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 17).

Tableau 17 – Score moyen selon l'indice de position sociale moyen de l'établissement (CEDRE 2006-2012-2017)

| Indice moyen école | Année | Répartition (%) | Score moyen | Écart type |
|---|---|---|---|---|
| 1er quart | 2006 | 25.0 | 230 | 45 |
| 1er quart | 2012 | 24.6 | 229 | 49 |
| 1er quart | 2017 | 25.0 | **237** | 47 |
| 2e quart | 2006 | 24.8 | 245 | 48 |
| 2e quart | 2012 | 24.8 | 250 | 49 |
| 2e quart | 2017 | 24.6 | **243** | 51 |
| 3e quart | 2006 | 24.8 | 257 | 48 |
| 3e quart | 2012 | 25.3 | 258 | 51 |
| 3e quart | 2017 | 25.0 | 254 | 52 |
| 4e quart | 2006 | 25.6 | 268 | 51 |
| 4e quart | 2012 | 25.3 | 266 | 51 |
| 4e quart | 2017 | 25.3 | **275** | 52 |

Note de lecture : en 2017, le score moyen des élèves appartenant au quart des écoles les plus défavorisées (1er quart) augmente de 8 points par rapport à 2012. 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 interessantes 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, les élèves sont demandés 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 18 présente les grands résultats de cet instrument.

Tableau 18 – Résultats de l'instrument de mesure de la motivation au test (CEDRE 2017)

| | | % |
|---|---|---|
| Comment as-tu trouvé les exercices de cette évaluation ? | Très faciles | 9.7 |
| | Faciles | 48.6 |
| | Difficiles | 36.7 |
| | Très difficiles | 5.0 |
| Je me suis bien appliqué(e) pour faire cette évaluation | Pas du tout d'accord | 2.5 |
| | Pas d'accord | 9.3 |
| | D'accord | 55.5 |
| | Tout à fait d'accord | 32.7 |
| Je me suis autant appliqué(e) pour faire cette évaluation que le travail quotidien de classe | Pas du tout d'accord | 7.2 |
| | Pas d'accord | 19.0 |
| | D'accord | 40.2 |
| | Tout à fait d'accord | 33.6 |

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.

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.

Ninnin, L.-M., & Pastor, J.-M. (2018). CEDRE 2006-2012-2017, histoire, géographie, enseignement moral et civique en fin d'école primaire : stabilité des résultats depuis onze ans. *Note d'information*, *16*.

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 n° 3*, 427-450.

# Liste des tableaux

# Table des figures