

# **CEDRE**

Cycle des Évaluations Disciplinaires Réalisées sur Échantillons

## **Rapport technique**

Mathématiques 2014

École

Auteurs :

Etienne DALIBARD

Saskia KESKPAIK

Marion LE CAM

Jean-Marc PASTOR

Thierry ROCHER

Bureau de l'évaluation des élèves

DEPP - Direction de l'évaluation, de la prospective et de la performance

Ministère de l'éducation nationale, de l'enseignement supérieur et de la recherche

Décembre 2015

## Table des matières

<b>Introduction</b>	<b>3</b>
<b>1 Cadre d'évaluation</b>	<b>4</b>
1.1 Objectifs . . . . .	4
1.2 Les compétences et connaissances visées . . . . .	5
1.3 Construction du test . . . . .	11
1.4 Passation des évaluations . . . . .	14
<b>2 Sondage</b>	<b>15</b>
2.1 Méthodes . . . . .	15
2.2 Echantillonnage . . . . .	18
2.3 Etat des lieux de la non-réponse . . . . .	21
2.4 Redressement . . . . .	24
2.5 Précision . . . . .	24
<b>3 Analyse des items</b>	<b>26</b>
3.1 Méthodologie . . . . .	26
3.2 Codage des réponses aux items . . . . .	29
3.3 Résultats . . . . .	32
<b>4 Modélisation</b>	<b>34</b>
4.1 Méthodologie . . . . .	34
4.2 Résultats . . . . .	40
4.3 Calcul des scores . . . . .	41
4.4 Courbes d'information . . . . .	42
<b>5 Construction de l'échelle</b>	<b>43</b>
5.1 Méthode . . . . .	43
5.2 Caractérisation des groupes de niveaux . . . . .	43
5.3 Exemples d'items . . . . .	46
<b>6 Variables contextuelles et non cognitives</b>	<b>52</b>
6.1 Variables sociodémographiques et indice de position sociale . . . . .	52
6.2 Élaboration des questionnaires de contexte . . . . .	53
6.3 Construction des scores factoriels et des indicateurs . . . . .	53
6.4 Motivation des élèves face à la situation d'évaluation . . . . .	54
<b>7 Annexe</b>	<b>56</b>
<b>Références</b>	<b>59</b>



## 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 : 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 Les compétences et connaissances visées

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

L'évaluation CEDRE en fin d'école en mathématiques vise à faire le point sur les acquis des élèves et à mesurer l'évolution de ces connaissances et compétences entre deux prises d'information (2008 et 2014).

Les documents de référence pour la construction des items sont le programme officiel en vigueur à partir de la rentrée scolaire 2012-2013 parus au BO spécial n°6 du 28 août 2008 ainsi que les grilles de référence du socle commun. A noter que la première prise d'informations en 2008 se basait sur le programme en cours à l'époque, c'est-à-dire celui de 2002. Si ceci n'a pas entraîné de changement pour l'approche par compétences, il n'en a pas été de même pour les connaissances. En effet, le champ exploitation de données numérique a disparu en tant que tel dans le programme de 2008 ; les éléments constitutifs de ce champ sont ventilés dans les champs organisation et gestion de données (O.G.D.) qui intègre la proportionnalité et dans les autres champs mathématiques.

### 1.2.1 Les compétences évaluées

En 2008 comme en 2014, la même grille de compétences est utilisée quel que soit le champ mathématique concerné (cf. tableau 2). Elle permet d'assurer un point de comparaison entre les deux prises d'informations.

Tableau 2 – Définition des compétences évaluées

Compétence évaluée	Définition
Identifier	Reconnaître la dimension mathématique d'un énoncé. L'élève choisit un résultat parmi les propositions d'un questionnaire à choix multiple.
Exécuter	Répondre immédiatement à un stimulus direct. L'élève écrit sa réponse dans un champ-libre.
Traiter	Analyser et comprendre des données ; traiter ces données (le brouillon était autorisé). L'élève choisit sa réponse parmi les propositions d'un questionnaire à choix multiple.
Produire	Analyser et comprendre des données ; traiter ces données. L'élève construit sa réponse dans un cadre de recherche.
Contrôler/Valider	Analyser des démarches d'élèves proposées et vérifier leur véracité. L'élève choisit sa réponse parmi les propositions d'un questionnaire à choix multiple.

### 1.2.2 Les connaissances évaluées

Les connaissances sont réparties dans les différents champs mathématiques.

En 2008, six champs mathématiques (référence au programme de 2002) étaient disponibles :

- Connaissance des nombres entiers naturels
- Fractions et nombres décimaux
- Calcul
- Espace et géométrie
- Grandeurs et mesures
- Exploitation de données numériques

En 2014, six champs mathématiques (référence au programme de 2008) étaient disponibles :

- Nombres entiers naturels
- Nombres décimaux
- Calcul
- Géométrie
- Grandeurs et mesures
- Organisation Gestion de données



### 1.2.3 Particularité de l'évaluation 2014

L'évaluation CEDRE en fin d'école comporte deux volets : le premier concerne des items présentés sur un cahier d'évaluation en version « papier-crayon » ; le second correspond à celui sur ordinateur. Notons que les analyses présentées dans ce documents reposent sur les épreuves « papier-crayon ». Nous donnons néanmoins quelques éléments sur les épreuves numériques qui font l'objet d'études séparées.

#### 1er volet : évaluation « papier-crayon »

Il s'agit pour l'élève de répondre à des unités présentées sur un support « papier-crayon ». Une unité se compose d'un ou de plusieurs documents que l'élève devra utiliser pour répondre aux questions. L'unité 6 (exemple 1 dans l'encadré « exemples d'items ») se compose de quatre questions notées respectivement situation 1, 2, 3 et 4. L'unité 13 (exemple 2 dans l'encadré « exemples d'items ») se compose d'un document et d'une question. Les unités sont regroupées dans des blocs ; les blocs dans des cahiers (tableau 3)

#### 2nd volet : évaluation « numérique »

Les technologies de l'information et de la communication numériques ont apporté une nouvelle dimension à l'acte de lire, de comprendre et de réaliser des exercices sur un écran numérique. Dès lors, l'élève doit développer de nouvelles pratiques lui permettant d'acquérir des habiletés spécifiques qu'un des volets de cette étude a voulu approcher.

Dans les épreuves numériques, il s'agit pour l'élève de répondre à des unités présentées sur un support électronique. Plusieurs cas se présentent :

- L'unité propose directement une question à l'élève (exemple 3). Il doit répondre dans le champ de réponse prévu à cet effet. Ce type d'item est particulièrement adapté au calcul mental et plus largement aux stimuli qui entraîne une réponse de l'ordre de l'automatisme.
- L'unité propose un document à l'élève et que question afférente à ce dernier (exemple 4). Ce type d'item est adapté pour la dématérialisation, c'est-à-dire le passage d'une unité « papier » vers une unité numérique. Il est alors possible de mesurer l'écart de performance entre les deux supports.
- L'unité propose un document multimédia à l'élève et une question qui s'y rapporte (exemple 5). Ce type d'item permet de tester les habiletés nouvelles induites par le support numérique.
- Les unités sont ventilées dans des situations ; les situations dans des modules.

## Exemples d'items

### Exemple 1 : grandeurs et mesures

#### Unité 6

Pour chaque situation, indique la réponse qui te semble correcte.

##### Situation 1

Longueur d'une piscine :

- 1  25 cm
- 2  25 m
- 3  25 km

ESMIM870101  
23

##### Situation 2

La longueur d'un stylo :

- 1  13 mm
- 2  13 cm
- 3  13 m

ESMIM870201  
24

##### Situation 3

Le poids d'un bébé à la naissance :

- 1  3 g
- 2  3 kg
- 3  30 kg

ESMIM870301  
25

##### Situation 4

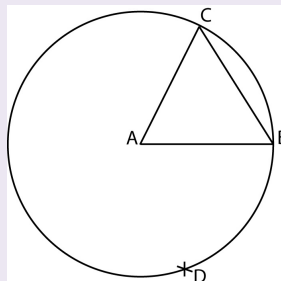
La capacité d'une grande bouteille d'eau :

- 1  1,5 mL
- 2  1,5 cL
- 3  1,5 L

ESMIM870401  
26

**Exemple 2 : géométrie****Unité 13**

Observe la figure ci-dessous.



*ABC est un triangle équilatéral - D est un point du cercle - A est le centre du cercle*

**Observation 1**

Sans utiliser d'instrument de mesure, réponds par "Vrai" ou "Faux".

	Vrai	Faux
AC = BC	<input type="checkbox"/> 1	<input type="checkbox"/> 2
AB = AD	<input type="checkbox"/> 1	<input type="checkbox"/> 2
AD = CB	<input type="checkbox"/> 1	<input type="checkbox"/> 2
AC = DB	<input type="checkbox"/> 1	<input type="checkbox"/> 2

E5MTG960101  
48

E5MTG960102  
49

E5MTG960103  
50

E5MTG960104  
51

## Exemples d'items numériques

### Exemple 3

Bloc N1-01

- [Accueil](#)
- [Question 1](#)
- [Question 2](#)
- [Question 3](#)
- [Question 4](#)

**Question 1**

11 x 3

### Exemple 4

TEDN10-64

- [Site](#)
- [Accueil](#)
- [Question 1](#)
- [Question 2](#)
- [Question 3](#)

**Question 1**

Quel jour de la semaine fait-il le plus froid à 8h ?

**Relevé de températures**

Jour de la semaine	Température à 8h	Température à 14h
Lundi	1	8
Mardi	3	6
Mercredi	5	9
Judi	6	11
Vendredi	7	10
Samedi	8	9
Dimanche	2	7

- lundi
- mardi
- mercredi
- jeudi
- vendredi
- samedi
- dimanche

### Exemple 5

Angle 1

- [Site](#)
- [Accueil](#)
- [Question 1](#)

**Question 1**

A chacune des étapes, observe l'angle en A.

Indique si l'angle est aigu, droit ou obtus.

Pour voir la consigne. [Cliquer ici](#)

	un angle aigu	un angle droit	un angle obtus
A l'étape 1 l'angle A est ...	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
A l'étape 2 l'angle A est ...	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
A l'étape 3 l'angle A est ...	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

## 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 chef de projet, l'inspection et l'inspection générale. Un item proposé par un concepteur, pédagogue de terrain ayant une bonne connaissance des pratiques de classe, fait l'objet d'une discussion contradictoire jusqu'à aboutir à un consensus. L'item est alors soumis à un « cobayage », c'est-à-dire une passation auprès d'une ou plusieurs classes pour estimer sa difficulté et recueillir les réactions des élèves.

Un équilibre de proportion entre les items considérés comme étant de difficulté « facile », « moyenne » ou « difficile » est recherché. Les items des six domaines sont pour certains identiques à ceux proposés en 2008 afin d'assurer une comparabilité de qualité.

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 donnent lieu à un regroupement ultérieur de leurs propositions. Dans le cas de ces séries, des seuils statistiques ont été établis pour valider les réponses des élèves. Les réponses des formats « réponse libre de l'élève » sont corrigées par des experts via une interface Internet. 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**

L'évaluation CEDRE 2014 est constituée de 13 cahiers tournants intégrant un ensemble de 13 blocs d'évaluation (cf. tableau 3) contenant des items de 2008 repris à l'identique et de nouveaux items qui ont fait l'objet d'une expérimentation en 2013. Au total l'évaluation de 2014 est constituée de 209 items, dont 119 d'ancrage (identiques à 2008) soit 57 %, avant regroupements des items de type

« vrai/faux » et analyses psychométriques (cf. parties 3 et 4 du rapport).

Tableau 3 – Répartition des blocs dans les cahiers pour l'évaluation CEDRE mathématiques école 2014

Cahier	Séquence 1		Séquence 2	
	Bloc 1	Bloc 2	Bloc 3	Bloc 4
C1	B5	B6	B12	B7
C2	B4	B12	B3	B8
C3	B6	B3	B2	B9
C4	B12	B2	B1	B13
C5	B3	B1	B7	B11
C6	B2	B7	B8	B10
C7	B1	B8	B9	B5
C8	B7	B9	B13	B4
C9	B8	B13	B11	B6
C10	B9	B11	B10	B12
C11	B13	B10	B5	B3
C12	B11	B5	B4	B2
C13	B10	B4	B6	B1

La méthodologie des cahiers tournants permet d'évaluer un nombre important d'items sans allonger le temps de passation. Les items sont ainsi repartis dans des blocs d'une durée de 20 minutes et les blocs sont ensuite distribués dans les cahiers tout en respectant certaines contraintes telles que chaque bloc devant se retrouver un même nombre de fois au total et chaque association de blocs doit figurer au moins une fois dans un cahier. Ce dispositif, couramment utilisé dans les évaluations bilans, notamment les évaluations internationales, permet d'estimer la probabilité de réussite de chaque élève à chaque item sans que chaque élève ait passé l'ensemble des items.

Au final, pour l'évaluation CEDRE 2014, chaque cahier comprend deux séquences cognitives de 45 minutes chacune. Elles sont complétées par une troisième séquence de 30 minutes (questionnaire de contexte), identique dans tous les cahiers (cf. partie 6 du rapport).

### 1.3.3 Constitution des blocs numériques

L'évaluation numérique est également constituée de 13 modules reprenant la méthodologie des cahiers tournants intégrant un ensemble de 4 unités d'évaluation contenant trois types d'items : item impliquant une réponse directe à un stimulus (calcul mental) ; item reproduisant à l'identique un item existant au

format « papier » (dématérialisation) et item utilisant les ressources « multimédia » (animation).

## 1.4 Passation des évaluations

La passation de l'évaluation finale a eu lieu en mai 2014. Comme en 2008, cette évaluation a été précédée d'une expérimentation l'année  $n - 1$  de façon à tester un grand nombre d'items auprès d'un échantillon réduit d'établissements.

Dans chaque école, le directeur ou la directrice a été désigné 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 mêmes conditions quel que soit l'établissement.

Chaque séquence était passée dans une demi-journée. Les deux premières séquences interrogeaient les élèves sur leurs connaissances et compétences en mathématiques alors que la troisième séquence correspondait à la réponse à un questionnaire de contexte permettant d'éclairer les réponses des élèves et de nuancer certaines différences de niveaux qui peuvent apparaître (notamment entre types d'écoles fréquentées).

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

Une fois l'évaluation terminée, les cahiers et questionnaires étaient renvoyés dans des conditionnements prévus à cet effet, pré affranchis et pré étiquetés. Aucun travail de correction n'a été demandé aux écoles.



## 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 connaît 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 :

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

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  $\hat{Y}$  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  $V(\hat{Y})$ . En absence de formule théorique pour calculer  $V(\hat{Y})$ , il existe plusieurs procédures permettant de l'estimer, c'est-à-dire de calculer  $\hat{V}(\hat{Y})$ , 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 réplification, 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  $\hat{Y}$  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  $\hat{Y}$ , 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  $\hat{Y}$  est alors obtenue en calculant la variance d'échantillonnage de  $g_i e_i$ .

## 2.2 Echantillonnage

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

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

### 2.2.1 Echantillon 2008

#### Modalités de sélection

Le tirage des écoles est stratifié selon six strates. Ensuite, tous les élèves de CM2 des écoles sélectionnées sont interrogés.

#### Stratification

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

1. Écoles publiques hors ZEP (14 élèves ou plus)
2. Écoles publiques en ZEP (14 élèves ou plus)
3. Écoles privées (14 élèves ou plus)

4. Écoles publiques hors ZEP (entre 6 et 13 élèves)
5. Écoles publiques en ZEP (entre 6 et 13 élèves)
6. Écoles privées (entre 6 et 13 élèves)

L'échantillon de 2008 visait environ 4 000 élèves.

### Base de sondage

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

Tableau 4 – Répartition base de sondage - CEDRE mathématiques CM2 2008

	Nb écoles	Nb élèves
Écoles publiques hors ZEP (14 élèves ou plus)	13 501	434 132
Écoles publiques en ZEP (14 élèves ou plus)	2 613	92 593
Écoles privées (14 élèves ou plus)	2 919	99 492
Écoles publiques hors ZEP (entre 6 et 13 élèves)	7 012	64 676
Écoles publiques en ZEP (entre 6 et 13 élèves)	357	3 381
Écoles privées (entre 6 et 13 élèves)	1 208	11 333
<b>Total</b>	<b>27 610</b>	<b>705 607</b>

### Échantillon

Le tableau 5 présente la répartition de l'échantillon dans les différentes strates. Au total, 155 écoles ont été sélectionnées. Les élèves de ZEP et du privé ont été intentionnellement sur-représentés afin de calculer des indicateurs plus robustes selon le secteur de scolarisation.

Tableau 5 – Répartition dans l'échantillon - CEDRE mathématiques CM2 2008

	Nb écoles	Nb élèves
Écoles publiques hors ZEP (14 élèves ou plus)	38	1 213
Écoles publiques en ZEP (14 élèves ou plus)	39	1 291
Écoles privées (14 élèves ou plus)	37	1 330
Écoles publiques hors ZEP (entre 6 et 13 élèves)	20	217
Écoles publiques en ZEP (entre 6 et 13 élèves)	5	47
Écoles privées (entre 6 et 13 élèves)	16	154
<b>Total</b>	<b>155</b>	<b>4 252</b>

### 2.2.2 Echantillon 2014

#### Champ

Le champ est celui des élèves scolarisés en classe de CM2 des écoles publiques et privées sous contrat en France métropolitaine.

Sont donc exclus du champ :

- Les TOM.
- Les écoles hors contrat.
- Les écoles à l'étranger.
- Les écoles spécialisées.
- Pour des raisons de coût, les écoles de moins de 6 élèves de CM2 sont exclues du champ.
- Les DOM.

Le tableau 6 récapitule ces exclusions :

Tableau 6 – Exclusions - CEDRE mathématiques CM2 2014

	Ecoles	Elèves
Ecoles accueillant des CM2 hors TOM	32 747	803 374
On retire les écoles hors contrat	32 467	800 822
On retire les écoles spécialisées	32 449	800 308
On retire les petites écoles (<6 CM2)	29 907	791 185
Exclusion des DOM	28 999	752 615
<b>Base CM2 CEDRE</b>	<b>28 999</b>	<b>752 615</b>

#### Base de sondage

Trois strates sont considérées (tableau 7).

Tableau 7 – Répartition base de sondage - CEDRE mathématiques CM2 2014

	Ecoles	Elèves	Nb moyen de CM2 par école
1. Public hors EP	21 736	541 675	24.9
2. EP	3 086	97 452	31.6
3. Privé	4 177	113 488	27.2
<b>Total</b>	<b>28 999</b>	<b>752 615</b>	

### Modalités de sélection

L'échantillon est stratifié selon les trois strates. L'éducation prioritaire et le privé ont été surreprésentés pour avoir des statistiques plus robustes par strates. Une fois les écoles tirées au sort, tous les élèves de CM2 des écoles sélectionnées sont interrogés : c'est un sondage par grappe. L'échantillon a été tiré à l'aide de la macro CUBE.

En outre, les écoles des échantillons « Expérimentation lecture », TIMSS CM1 et Socle C1 C3 CE1 ont été retirées de la base de tirage pour qu'il n'y ait pas de recouvrement.

### Échantillon

En 2014, l'échantillon visait environ 8 000 élèves ce qui correspond à 121 écoles dans la strate 1, 95 écoles pour la strate 2 et 74 écoles pour la strate 3 (tableau 8).

Tableau 8 – Répartition dans l'échantillon - CEDRE mathématiques CM2 2014

	Nb écoles	Nb élèves attendus
1. Public hors EP	121	2 986
2. EP	95	2 968
3. Privé	74	1 998
<b>Total</b>	<b>290</b>	<b>7 952</b>

## 2.3 Etat des lieux de la non-réponse

### 2.3.1 Non-réponse totale

Parmi la non-réponse totale, nous distinguons selon la non-réponse d'écoles entières ou la non-réponse d'élèves dans les écoles participantes. Les chiffres suivants ont été observés pour 2014. Tout d'abord, 92,4 % des écoles de l'échantillon ont répondu à l'évaluation (tableau 9).

Au final, 91 % des effectifs attendus ont participé (tableau 10).

Tableau 9 – Non-réponse des écoles - CEDRE mathématiques CM2 2014

strate	Nb écoles attendues	Nb écoles répondantes	% de écoles répondantes
1- public hors EP	121	113	93,4%
2- EP	95	91	95,8%
3- privé	74	64	86,5%
<b>Total</b>	<b>290</b>	<b>268</b>	<b>92,4%</b>

Tableau 10 – Non-réponse globale - CEDRE mathématiques CM2 2014

strate	Nb élèves attendus	Nb élèves répondants	% élèves répondants
1- public hors EP	2 986	2 745	91,9%
2- EP	2 968	2 711	91,3%
3- privé	1 998	1 778	89,0%
<b>Total</b>	<b>7 952</b>	<b>7 234</b>	<b>91,0%</b>

### 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").

### CEDRE 2008



Les cahiers élèves sont composés de trois séquences qui devaient être passées trois jours différents. La non réponse terminale a été étudiée par séquence et par cahier. Parmi les élèves ayant de la non réponse terminale, il y en a en moyenne 3,9 pour la 1<sup>re</sup> séquence, 7,9 pour la 2<sup>e</sup> séquence et 3,8 pour la 3<sup>ème</sup> séquence.

Si un élève a passé moins de 50 % d'une séquence, on considère qu'il n'a pas vu la séquence (code « s »).

Au final, on considère que :

- 57 élèves n'ont pas vu la séquence 1 dont :
  - 51 n'ont répondu à aucun item de la séquence
  - 6 ont répondu à moins de 50 % de la séquence
- 56 élèves n'ont pas vu la séquence 2 dont :
  - 46 n'ont répondu à aucun item de la séquence
  - 10 ont répondu à moins de 50 % de la séquence
- 88 élèves n'ont pas vu la séquence 3 dont :
  - 83 n'ont répondu à aucun item de la séquence
  - 5 ont répondu à moins de 50 % de la séquence

Les élèves dont les trois séquences sont codées en « s » sont considérés comme de la non réponse totale. C'est le cas pour 1 élève.

#### **CEDRE 2014**

Les cahiers élèves sont composés de deux séquences. La non réponse terminale a été étudiée par séquence et par cahier. Parmi les élèves ayant de la non réponse terminale, il y en a en moyenne 6,2 pour la 1<sup>ère</sup> séquence et 8,8 pour la 2<sup>ème</sup> séquence.

Si un élève a passé moins de 50 % d'une séquence, on considère qu'il n'a pas vu la séquence (code « s »).

Au final, on considère que :

- 60 élèves n'ont pas vu la séquence 1 dont :
  - 22 n'ont répondu à aucun item de la séquence
  - 38 ont répondu à moins de 50 % de la séquence
- 122 élèves n'ont pas vu la séquence 2 dont :
  - 57 n'ont répondu à aucun item de la séquence
  - 65 ont répondu à moins de 50 % de la séquence

Les élèves dont les quatre séquences sont codées en « s » sont considérés comme de la non réponse totale. C'est le cas pour 25 élèves.

## 2.4 Redressement

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

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

Le tableau 11 montre que l'ampleur du calage est très réduit.

Tableau 11 – Comparaison entre les marges de l'échantillon et les marges dans la population

	Modalité ou variable	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population
Retard	1	88 888.51	85 578	11.81	11.37
	2	663 726.43	667 037	88.19	88.63
Sexe	1	391 603.43	384 042	52.03	51.03
	2	361 011.51	368 573	47.97	48.97
Strate	1	541 674.95	541 675	71.97	71.97
	2	97 452.00	97 452	12.95	12.95
	3	113 488.00	113 488	15.08	15.08

## 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 mathématiques CM2

Année	Score moyen	Erreur standard
2008	250	2.1
2014	248,6	1.6

Pour savoir si l'évolution entre 2008 et 2014 est significative, il faut donc calculer la valeur suivante :

$$\frac{|\hat{Y}_{2014} - \hat{Y}_{2008}|}{\sqrt{se_{\hat{Y}_{2014}}^2 + se_{\hat{Y}_{2008}}^2}} \quad (2)$$

Avec une valeur de 0,98 (inférieure à 1,96), cela signifie que la baisse du score moyen observée entre 2008 et 2014 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 – Répartition en % dans les groupes de niveaux - CEDRE mathématiques CM2

Année	Groupe < 1	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5
2008	2.6	12.4	25.5	31.3	18.2	10.0
2014	3.7	12.6	26.1	28.6	18.8	10.2

Tableau 14 – Erreurs standards des répartitions en % dans les groupes de niveaux - CEDRE mathématiques CM2

Année	Groupe < 1	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5
2008	0.4	1.1	1.0	1.0	1.0	1.1
2014	0.4	0.6	0.8	0.6	0.7	0.8

### *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 :

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

Dans le cas d'un sondage en grappes, la précision est dégradée en comparaison d'un sondage aléatoire simple (tableau 15). Cela signifie qu'en 2014, un sondage aléatoire simple avec un effectif 3,4 fois moins important aurait conduit au même niveau de précision.

Tableau 15 – Effet du plan de sondage - CEDRE mathématiques CM2 2014

Année	Erreur Standard	Erreur SAS	<i>Design Effect</i>
2008	2.1	0.77	7.44
2014	1.6	0.56	8.16

## 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  $i$  ( $i = 1, \dots, n$ ) à l'item  $j$  ( $j = 1, \dots, J$ ). Dans notre cas, les items sont dichotomiques, c'est-à-dire qu'ils ne prennent que deux modalités (la réussite ou l'échec) :

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

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

$$p_j = \frac{1}{n} \sum_{i=1}^n Y_i^j \quad (5)$$

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  $i$  :

$$S_i = \sum_{j=1}^J Y_i^j \quad (6)$$

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<sup>1</sup>.

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 :

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

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<sup>2</sup>.

### 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

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

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)<sup>3</sup>. 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 :

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

La réponse  $y_{ij}$  de l'élève  $i$  à l'item  $j$  est incorrecte tant que la variable latente  $Z_j$  reste en deçà 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 (cf. l'exemple 2 donné au paragraphe 1.2.3). 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.

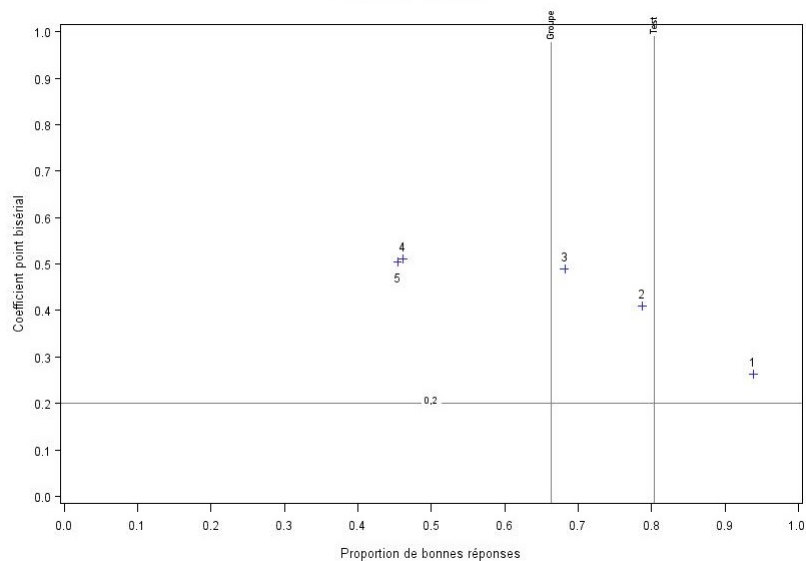
Le regroupement de ces items consiste à faire la somme des indicatrices de réussite et à déterminer un seuil de maîtrise. Une visualisation graphique est utilisée pour fixer les scores « seuils » (cf. figure 1). Ce graphique représente le taux de réussite pour chaque seuil possible en fonction de la 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.

### 3.2.3 Traitement des données et correction des questions ouvertes

Tous les cahiers recueillis dans le cadre de cette opération ont été scannés par une société extérieure. Les réponses aux questions à choix multiples ainsi que les grilles d'évaluation remplies par les professeurs lors des séquences de travaux pratiques ont été numérisées et les codes de réponses stockés dans un fichier. En ce qui concerne les questions ouvertes, demandant une rédaction plus ou moins longue de la part des élèves (explication, schématisation...), elles ont été découpées en « imagettes » puis transmises au ministère afin d'être intégrées dans un logiciel de correction à distance (cf. encadré « AGATE »). Celui-ci nécessite la formation technique des correcteurs et l'élaboration d'un cahier des charges strict de corrections pour limiter la subjectivité des corrections. Une fois la correction terminée, les codes saisis par les correcteurs ont été stockés dans un fichier puis associés à ceux issus des réponses aux QCM.



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 biserial 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).

### 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'images (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 image, il arrive parfois qu'il y ait une différence de codage. Cette image 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'images 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

Lorsque l'on calcule les indices de discrimination sur l'ensemble des items, aucun item n'est apparu faiblement discriminant (i.e. tous les *r-bis point* sont supérieurs à 0,2).

### 3.3.2 Dimensionnalité

Le tableau 16 présente les résultats de l'analyse factorielle des items effectuée sur l'année 2014.

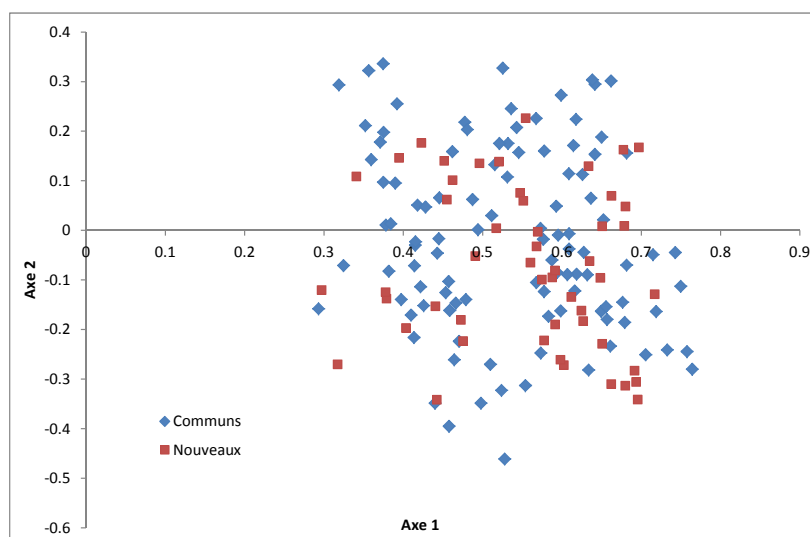
La structure des items est fortement unidimensionnelle : le « poids » de la première dimension est très important (valeur propre de 65,2 contre 6,5 pour

Tableau 16 – ACP (CEDRE mathématiques CM2 2014)

	Valeur Propre	Difference	Proportion	Proportion cumulee
1	65.2	58.7	0.31	0.31
2	6.5	1.0	0.03	0.34
3	5.5	0.6	0.03	0.37

la deuxième dimension). En outre, il n'apparaît pas de différenciation sur la deuxième dimension, entre les items nouveaux de 2014 et les items repris de 2008, comme le montre la figure 2

Figure 2 – Premier plan factoriel des items de 2014 (CEDRE mathématiques CM2)



Note de lecture : Le graphique représente le premier plan factoriel de l'ACP réalisée à partir des coefficients de corrélations tétrachoriques. L'axe des ordonnées représente la première dimension et l'axe des abscisses la deuxième dimension. Les losanges bleus représentent les items communs entre 2008 et 2014. Les carrés rouges représentent les items nouveaux de 2014.

## 4 Modélisation

### 4.1 Méthodologie

#### 4.1.1 Modèle de réponse à l'item

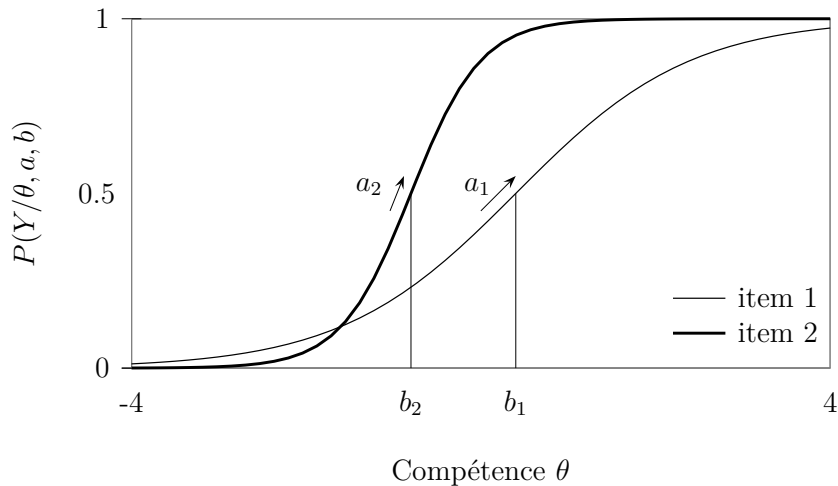
Le modèle de mesure utilisé est un modèle de réponse à l'item à deux paramètres avec une fonction de lien logistique (MRI 2PL) :

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

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

Figure 3 – Modèle de réponse à l'item - 2 paramètres



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

L'avantage de ce type de modélisation, c'est de séparer deux concepts-clé, à savoir la difficulté de l'item et le niveau de compétence de l'élève. Les MRI ont un intérêt pratique pour la construction de tests et la comparaison entre différents groupes d'élèves : si le modèle est bien spécifié sur un échantillon donné, les paramètres des items – en particulier leurs difficultés – peuvent être considérés comme fixes et applicables à d'autres échantillons dont il sera alors possible de déduire les paramètres relatifs aux élèves – en particulier, leur niveau de compétence. 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 3).

#### 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 2008).

Sous l'hypothèse d'indépendance locale des items<sup>4</sup>, la fonction de vraisemblance

---

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).

s'écrit :

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

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 priori* (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 :

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

avec  $g$  la densité de  $\theta$ .

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

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

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 :

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

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 :

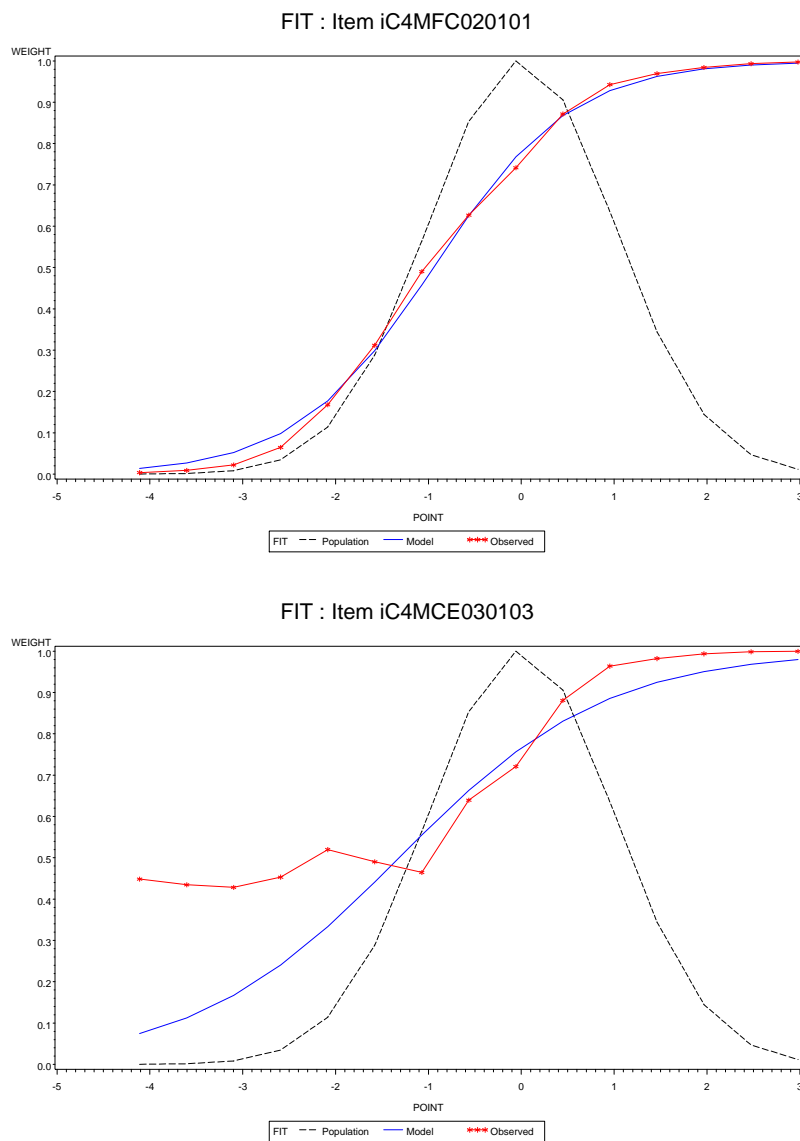
$$\frac{\partial \ln L}{\partial \theta_i} + \frac{J}{2I} = 0 \quad (15)$$

#### 4.1.3 Indice d'ajustement (FIT)

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

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

Figure 4 – 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 :

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

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 :

$$P(Y | Z, G) \neq P(Y | Z) \quad (18)$$

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 :

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

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  $\theta$  (cf. l'illustration plus loin dans la section 4.4). 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

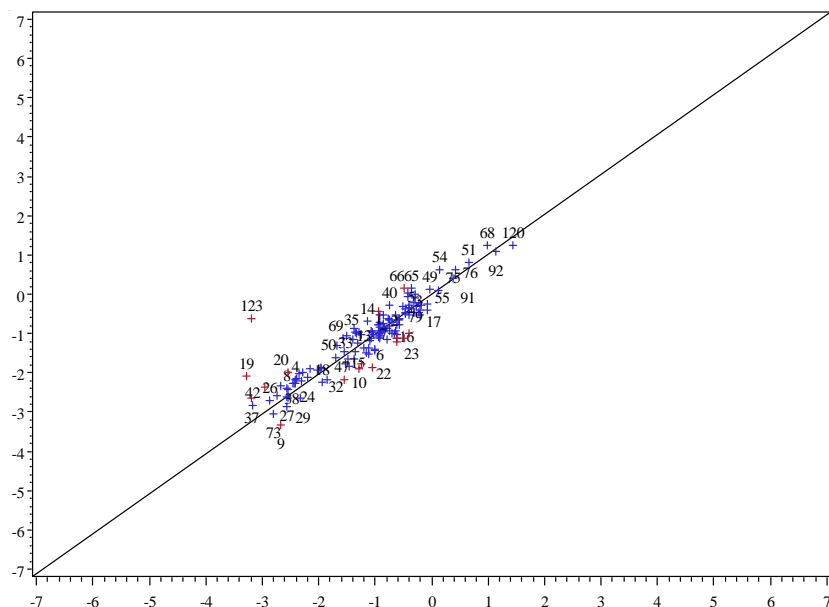
### 4.2.1 Identification des fonctionnements différentiels d'items (FDI)

L'analyse des FDI a permis de détecter 17 items : 10 items en faveur de 2014, 7 items en faveur de 2008 (figure 5). Ils ont été éliminés des calculs.

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

L'analyse des ajustements (FIT) a conduit à supprimer 14 items, dont 4 sont des items communs aux deux années, 9 sont des items de 2008 et 1 item de 2014.

Figure 5 – Comparaison des paramètres de difficulté 2008-2014 (CEDRE mathématiques CM2)



Note de lecture : Les points sont les items. En abscisse figure la valeur des paramètres de difficulté estimés en 2008, et en ordonnée la la valeur des paramètres de difficulté estimés et ajustés pour l'année 2014. Les items présentant un FDI apparaissent en rouge.

### 4.2.3 Bilan de l'analyse des items

Au départ, il y avait :

- 140 items communs
- 218 items de 2008
- 130 items de 2014

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

- 119 items communs
- 209 items de 2008
- 129 items de 2014

## 4.3 Calcul des scores

Comme indiqué précédemment, une analyse conjointe des données (2008 et 2014) 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 2008. Le tableau 17 présente les résultats obtenus.

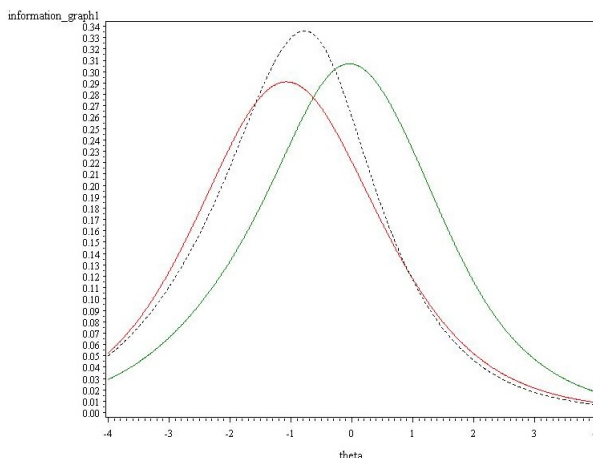
Tableau 17 – Niveaux de compétences CEDRE mathématiques école (moyennes et écarts-type)

annee	Nb élèves	Moyenne	Ecart-Type
2008	3 808	250.0	50.0
2014	7 234	248.6	52.3

#### 4.4 Courbes d'information

La figure 6 représente les courbes d'information pour les items des évaluations 2008 et 2014. La courbe rouge représente les items 2008 non repris en 2014, la courbe verte les items nouveaux de 2014 et la courbe en pointillé les items communs aux deux évaluations. Il ressort que les items nouveaux de 2014 sont globalement plus difficiles que les items de 2008, mais l'ancrage a été réalisé sur des items communs présentant des difficultés équilibrées par rapport aux deux années.

Figure 6 – Courbe d'information du test (CEDRE mathématiques école)



Note de lecture : la courbe rouge représente la courbe d'information des items de 2008 non repris en 2014, la courbe verte celle des items nouveaux de 2014 et la courbe en pointillé celle des items communs aux deux évaluations.

## 5 Construction de l'échelle

### 5.1 Méthode

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

Les scores en mathématiques estimés selon le modèle de réponse à l'item présenté dans la partie précédente ont été standardisés de manière à obtenir une moyenne de 250 et un écart-type de 50 pour l'année 2008. Puis, comme le montre la figure 7, 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.

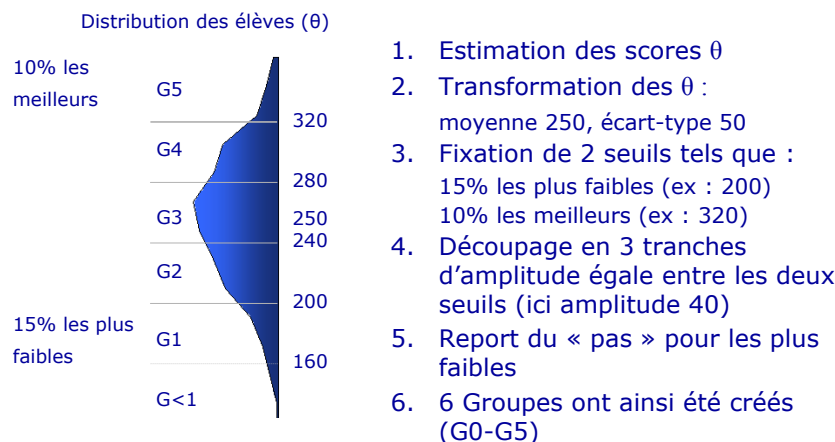
### 5.2 Caractérisation des groupes de niveaux

A partir de cette correspondance entre les items et les groupes, une description qualitative et synthétique des compétences maîtrisées par les élèves des différents groupes est proposée. Ces principaux résultats sont présentés dans une Note d'information (Dalibard & Pastor, 2015).

#### **Groupe $< 1$ (3,7 % des élèves)**

Ces élèves peuvent répondre ponctuellement à quelques items simples. Les réussites observées se fondent essentiellement sur des situations ayant trait à la vie courante - « estimer la taille d'objets usuels » -, à des pratiques scolaires ancrées - « repérer si une figure est symétrique par rapport à un axe vertical » -, donner une réponse par lecture directe - « lecture d'un nombre sur une règle graduée »

Figure 7 – Principes de construction de l'échelle



–. Ils maîtrisent très peu de compétences ou de connaissances exigibles en fin d'école primaire.

#### Groupe 1 (12,6 % des élèves)

Ces élèves ont des connaissances des nombres qui leur permettent la mise en œuvre d'opérations (additions et soustractions), néanmoins l'utilisation des retenues dans la soustraction n'est pas acquise. La construction du nombre en classes n'est pas solide, ils maîtrisent la « comptine » des nombres mais ils ont des difficultés en dehors de l'ordre croissant. Les réussites observées s'appuient essentiellement sur des automatismes scolaires. Certains de ces mécanismes leur permettent de réussir des problèmes additifs directs qui ne nécessitent qu'une seule étape pour leur résolution. Ils sont capables de mettre en œuvre des instruments de mesure pour comparer des segments. Ils maîtrisent la lecture de l'heure.

#### Groupe 2 (26,1 % des élèves)

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

**Groupe 3 (28,6 % des élèves)**

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

**Groupe 4 (18,8 % des élèves)**

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

**Groupe 5 (10,2 % des élèves)**

Ces élèves manient habilement les concepts mathématiques de fin d'école primaire. Cela leur permet de prendre du recul dans les situations nouvelles proposées, de gérer une masse d'information plus grande, de sélectionner les éléments utiles de ceux accessoires, d'imaginer des solutions et de produire un travail en autonomie. Quelques items leur résistent : il s'agit d'items dont les notions seront revues ultérieurement au collège (formules de solides ou calcul de vitesse moyenne). Ces élèves font preuve d'expertise dans les compétences et connaissances de fin d'école primaire, ils maîtrisent tous les champs du programme et font preuve de capacité d'abstraction, de rigueur et de précision. Ces élèves ont acquis l'ensemble des connaissances et des compétences exigibles en fin d'école primaire.

### 5.3 Exemples d'items

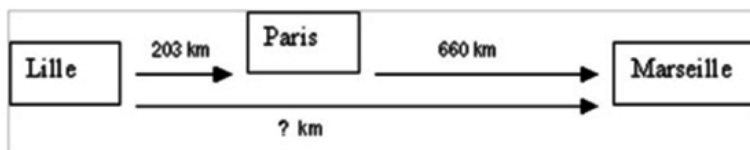
#### 5.3.1 Item caractéristique du groupe < 1

Les élèves de ce groupe résolvent ce problème de type additif (figure 8). Il est nécessaire de prendre de l'information dans le schéma et de trouver la situation finale :

- Le contexte est familier ; à partir de deux distances il est demandé d'en déduire la distance finale. La présentation du problème privilégie la schématisation et limite le recours au langage écrit.
- La connaissance mathématique mise en jeu correspond à l'utilisation d'une addition sans retenue. Nous sommes dans le cas générique de  $A+B=C$  avec A et B sont connus et C est l'inconnue.
- La tâche de l'élève consiste à traiter le problème puis à choisir parmi les propositions (QCM), la bonne réponse.

Figure 8 – Exemple groupe < à 1

Quelle est la distance entre Lille et Marseille ?



1	<input type="checkbox"/>	457 km	
2	<input type="checkbox"/>	473 km	
3	<input type="checkbox"/>	663 km	
4	<input type="checkbox"/>	863 km	ESMTC10101



### 5.3.2 Item caractéristiques du groupe 1

Les élèves de ce groupe résolvent ce problème de type additif (figure 9). Il s'agit de retrouver la situation initiale :

- Le contexte est familier ; Il correspond à des situations d'échange souvent présentes dans les pratiques scolaires et positionne les éléments du problème dans la vie courante. L'énoncé est court et propose un niveau de lecture très simple.
- La connaissance mathématique mise en jeu correspond à l'utilisation d'une addition avec retenue. Nous sommes dans le cas  $A+B=C$  avec B et C connus et A inconnue.
- La tâche de l'élève consiste à traiter le problème puis à choisir parmi les propositions (QCM), la bonne réponse.

Figure 9 – Exemple groupe 1

**Madame Durand va chez le garagiste pour payer sa facture.**

**La facture est de 236€.**

**Le garagiste lui rend 14€.**

**Combien avait-elle donnée ?**

200 euros

250 euros

300 euros

350 euros

### 5.3.3 Item caractéristique du groupe 2

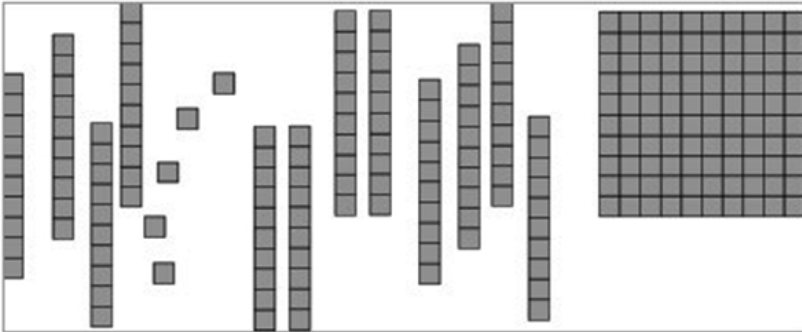
Cet exemple (figure 10) atteste d'une première connaissance stable du système de numération décimale. En effet, il nécessite la compréhension du groupement par dix et de la connaissance des classes de puissance de dix.

- Le contexte est familier ; Il correspond à des situations de « groupement/codage ».
- Les élèves utilisent régulièrement du matériel pédagogique leur permettant de mettre en œuvre des groupements par dix ou par cent.
- La connaissance mathématique mise en jeu correspond aux entiers naturels inférieurs à mille.
- La tâche de l'élève consiste à choisir parmi traiter les informations spatiales et à écrire sa réponse dans un peigne de codage.

Figure 10 – Exemple groupe 2

Une barre contient dix petits carrés.  
 Une plaque contient cent petits carrés.

**Question 3- Traiter - Nombres entiers naturels 99**



Sur le dessin ci-dessus, il y a  petits carrés.

0291124200001

### 5.3.4 Item caractéristique du groupe 3

Les élèves de ce groupe réussissent cet item (figure 11) ayant trait à la proportionnalité avec deux comme facteur de proportionnalité.

- Le contexte semble familier ; Il correspond à une recette de cuisine. La recherche correspond au doublement des proportions de base.
- La connaissance mathématique mise en jeu correspond dans l'ensemble des entiers naturels à comprendre la proportion qui s'établit entre double de fruit qui implique double de sucre. Il faut qu'ils aient repéré qu'entre 800 gr et 1 600 gr existe un facteur deux.
- La tâche de l'élève consiste à traiter le problème puis à choisir parmi les propositions (QCM), la bonne réponse.

Figure 11 – Exemple groupe 3

**Pour faire une salade de fruits on a utilisé la recette suivante :**

**800 gr de fruits pour 160 gr de sucre.**

**Avec la même recette, combien faut-il de sucre pour**

**1 600gr de fruits ?**

**80 gr**

**160 gr**

**320 gr**

**400 gr**

### 5.3.5 Item caractéristique du groupe 4

Les élèves de ce groupe réussissent cet item (figure 12) ayant trait à la proportionnalité avec un quart comme facteur de proportionnalité.

- Le contexte semble familier ; il correspond à une recette de cuisine. La recherche correspond au quart des proportions de base.
- La connaissance mathématique mise en jeu correspond dans l'ensemble des entiers naturels à comprendre la proportion qui s'établit entre quatre fois moins de fruit et l'ensemble. Il faut que les élèves aient repéré qu'entre 40 gr et 160 gr existe un facteur un quart.
- La tâche de l'élève consiste à traiter le problème puis à choisir parmi les propositions (QCM), la bonne réponse.

Figure 12 – Exemple groupe 4

**Pour faire une salade de fruits on a utilisé la recette suivante :**

**800 gr de fruits pour 160 gr de sucre.**

**Avec la même recette, combien faut-il de fruits pour  
40 gr de sucre ?**

100 gr

160 gr

200 gr

600 gr

### 5.3.6 Items caractéristiques du groupe 5

Les élèves de ce groupe réussissent cet item (figure 13) ayant trait à la proportionnalité avec trois quarts comme facteur de proportionnalité.

- Le contexte semble familier ; Il correspond à une recette de cuisine. La recherche correspond à un quart de moins des proportions de base.
- La connaissance mathématique mise en jeu correspond dans l'ensemble des entiers naturels à comprendre la proportion qui s'établit entre un quart de moins de sucre et l'ensemble. Il faut que les élèves aient repéré qu'entre 120 gr et 1 60 gr existe un facteur un trois quarts.
- La tâche de l'élève consiste à traiter le problème puis à choisir parmi les propositions (QCM), la bonne réponse.

Figure 13 – Exemple groupe 5

**Pour faire une salade de fruits on a utilisé la recette suivante :**

**800 gr de fruits pour 160 gr de sucre.**

**Avec la même recette, combien faut-il de fruits pour**

**120 gr de sucre ?**

**400 gr**

**500 gr**

**600 gr**

**700 gr**

## 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 (Dalibard & Pastor, 2015).

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, à paraître).

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 2014, la PCS des parents est disponible pour chaque élève, mais elle ne l'était pas en 2008.

Pour chaque établissement des échantillons de 2008 et 2014, 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 18).

Tableau 18 – Score moyen selon l'indice de position sociale moyen de l'établissement (CEDRE mathématiques école 2008-2014)

Indice moyen de l'etab.	Année	Score moyen	Ecart type
<b>Groupe 1 (25 % les plus défavorisés)</b>	2008	240	50
	2014	<b>229</b>	48
<b>Groupe 2</b>	2008	249	53
	2014	245	51
<b>Groupe 3</b>	2008	249	49
	2014	254	51
<b>Groupe 4 (25 % les plus favorisés)</b>	2008	261	46
	2014	266	<b>52</b>

## 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 en ligne était adressé aux enseignants des classes participantes à l'évaluation.

Pour pouvoir observer des évolutions dans le temps ainsi que de comparer les résultats à d'autres études, certaines questions ont été reprises des cycles précédents du CEDRE et d'autres évaluations nationales et/ou internationales, telles que PISA (Programme international pour le suivi des acquis des élèves).

Le questionnaire enseignant interroge les enseignants sur leur formation initiale et continue, leur ancienneté et parcours. Ce questionnaire inclut aussi des questions sur l'enseignement des mathématiques, sur les pratiques pédagogiques et sur les modalités d'évaluation utilisées par les enseignants.

Le questionnaire élève interroge des dimensions dites conatives intéressantes à mettre en lien avec le score obtenu à l'épreuve - l'intérêt pour les mathématiques, la perception de soi et l'anxiété en mathématiques. Ce questionnaire inclut également des questions concernant les attitudes de l'entourage vis-à-vis des mathématiques. Enfin, 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 socioprofessionnelle des parents mais aussi le parcours de l'élève (raccourcissement de cycle ou maintien dans un cycle, orientation retenue, etc.).

## 6.3 **Construction des scores factoriels et des indicateurs**

Les items correspondants à des dimensions conatives font d'abord l'objet d'une analyse factorielle exploratoire en facteurs corrélés permettant d'explorer la structure des items (Keskpaik, 2011). Les différentes dimensions sont validées puis un indice est calculé pour chacune d'entre elle, en considérant le premier axe d'une Analyse en Composantes Principales (ACP).

Le tableau 19 présente en guise d'illustration les items d'une de ces dimensions, en l'occurrence l'anxiété vis-à-vis des mathématiques.

Ces scores factoriels peuvent ensuite être utilisés dans des analyses secondaires. Notamment, dans des modèles de régression linéaire et de multiniveau.

Tableau 19 – Exemple de variable conative - l'anxiété vis-à-vis des mathématiques (CEDRE mathématiques école 2014)

Question	1er Axe ACP
Je me sens perdu(e) quand j'essaie de résoudre un problème en mathématiques	0,79
Je deviens très nerveux(se) quand je travaille à des problèmes de mathématiques	0,76
Je m'inquiète souvent en pensant que j'aurai des difficultés en mathématiques	0,74
Je suis très inquiet(e) quand j'ai un travail en mathématiques à faire	0,72
Je m'inquiète à l'idée d'avoir de mauvaises notes en mathématiques	0,66
J'aime bien résoudre des problèmes	0,52

Note de lecture : Les élèves devaient répondre à ces questions sur échelle dite de Lickert, de Tout à fait d'accord à Pas du tout d'accord. Pour faciliter l'interprétation des indicateurs, les échelles de certains items ont été inversés. Ainsi, plus la valeur de l'indicateur est élevé, et plus grande est « l'adhésion » de l'élève à la dimension correspondante.

#### 6.4 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 14) a été introduit dans plusieurs évaluations conduites au niveau national par la DEPP, y compris dans CEDRE mathématiques. 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 20 présente les grands résultats de cet instrument.



Tableau 20 – Résultats de l'instrument de mesure de la motivation au test (CEDRE mathématiques 2014)

		%
Comment as-tu trouvé les exercices de cette évaluation ?	Très faciles	17,2
	Faciles	64,0
	Difficiles	17,6
	Très difficiles	1,3
Je me suis bien appliqué(e) pour faire cette évaluation	Pas du tout d'accord	4,3
	Pas d'accord	6,6
	D'accord	54,5
Je me suis autant appliqué(e) pour faire cette évaluation que le travail quotidien de classe	Tout à fait d'accord	34,5
	Pas du tout d'accord	7,8
	Pas d'accord	17,5
	D'accord	39,7
	Tout à fait d'accord	35,0

Figure 14 – 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	<input type="checkbox"/> 1	<input type="checkbox"/> 2	<input type="checkbox"/> 3	<input type="checkbox"/> 4
Je me suis autant appliqué(e) à faire cette évaluation que le travail quotidien de classe	<input type="checkbox"/> 1	<input type="checkbox"/> 2	<input type="checkbox"/> 3	<input type="checkbox"/> 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.
- Dalibard, E., & Pastor, J. (2015). CEDRE 2014 - mathématiques en fin d'école primaire : les élèves qui arrivent au collège ont des niveaux très hétérogènes. *Note d'information*, 18.
- Keskpaik, S. (2011). L'analyse factorielle exploratoire. *Document de travail - série Méthodes*, M03.
- Keskpaik, S., & Rocher, T. (2015). La motivation des élèves français face à des évaluations à faibles enjeux. comment la mesurer ? son impact sur les réponses. *Education et formations*, 85-86, 119-139.
- Rocher, T. (1999). *Psychométrie et théorie des sondages* (Mémoire de Master non publié). Université Paris VI.
- Rocher, T. (2013). *Mesure des compétences : les méthodes se valent-elles ? questions de psychométrie dans le cadre de l'évaluation de la compréhension de l'écrit* (Thèse de doctorat non publiée). Université Paris-Ouest.
- Rocher, T. (2015). Mesure des compétences : méthodes psychométriques utilisées dans le cadre des évaluations des élèves. *Éducation et Formations*, 86-87, 37-60.
- Rocher, T. (à paraître). Construction d'un indice de position sociale des élèves. *Éducation et Formations*, 90.
- 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. échantillonnage 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

1	Cycle des évaluations disciplinaires réalisées sur échantillons (CEDRE) depuis 2003 . . . . .	5
2	Définition des compétences évaluées . . . . .	6
3	Répartition des blocs dans les cahiers pour l'évaluation CEDRE mathématiques école 2014 . . . . .	13
4	Répartition base de sondage - CEDRE mathématiques CM2 2008	19
5	Répartition dans l'échantillon - CEDRE mathématiques CM2 2008	19
6	Exclusions - CEDRE mathématiques CM2 2014 . . . . .	20
7	Répartition base de sondage - CEDRE mathématiques CM2 2014	20
8	Répartition dans l'échantillon - CEDRE mathématiques CM2 2014	21
9	Non-réponse des écoles - CEDRE mathématiques CM2 2014 . . .	22
10	Non-réponse globale - CEDRE mathématiques CM2 2014 . . . .	22
11	Comparaison entre les marges de l'échantillon et les marges dans la population . . . . .	24
12	Scores moyens et erreurs standard associées - CEDRE mathématiques CM2 . . . . .	24
13	Répartition en % dans les groupes de niveaux - CEDRE mathématiques CM2 . . . . .	25
14	Erreurs standards des répartitions en % dans les groupes de niveaux - CEDRE mathématiques CM2 . . . . .	25
15	Effet du plan de sondage - CEDRE mathématiques CM2 2014 . .	25
16	ACP (CEDRE mathématiques CM2 2014) . . . . .	33
17	Niveaux de compétences CEDRE mathématiques école (moyennes et écarts-type) . . . . .	42
18	Score moyen selon l'indice de position sociale moyen de l'établissement (CEDRE mathématiques école 2008-2014) . . . . .	52
19	Exemple de variable conative - l'anxiété vis-à-vis des mathématiques (CEDRE mathématiques école 2014) . . . . .	54
20	Résultats de l'instrument de mesure de la motivation au test (CEDRE mathématiques 2014) . . . . .	55

## Table des figures

1	Représentation graphique utilisée pour le regroupement d'items .	31
2	Premier plan factoriel des items de 2014 (CEDRE mathématiques CM2) . . . . .	33
3	Modèle de réponse à l'item - 2 paramètres . . . . .	34
4	Exemples d'ajustements (FIT) . . . . .	38

5	Comparaison des paramètres de difficulté 2008-2014 (CEDRE mathématiques CM2) . . . . .	41
6	Courbe d'information du test (CEDRE mathématiques école) . .	42
7	Principes de construction de l'échelle . . . . .	44
8	Exemple groupe < à 1 . . . . .	46
9	Exemple groupe 1 . . . . .	47
10	Exemple groupe 2 . . . . .	48
11	Exemple groupe 3 . . . . .	49
12	Exemple groupe 4 . . . . .	50
13	Exemple groupe 5 . . . . .	51
14	Instrument de mesure de la motivation au test . . . . .	55