# CEDRE

Cycle des Évaluations Disciplinaires Réalisées sur Échantillons

# Rapport technique

Sciences expérimentales 2013

Collège

Auteurs :
Anaïs BRET
Emilie GARCIA
Thierry ROCHER
Léa ROUSSEL
Ronan VOURC'H

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

Février 2015

# Table des matières

# Introduction

La DEPP met en place des dispositifs d'évaluation des acquis des élèves reposant sur des épreuves standardisées. Elle est également maître d'œuvre pour la France de diverses évaluations internationales. Ces programmes d'évaluations dites « bilans » sont des outils pour le pilotage d'ensemble du système éducatif. Ainsi, les évaluations du CEDRE révèlent, en référence aux programmes scolaires, les objectifs atteints et ceux qui ne le sont pas. Ces évaluations doivent permettre d'agir au niveau national sur les programmes des disciplines, sur les organisations des enseignements, sur les contextes de l'enseignement, sur des populations caractérisées.

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

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

# 1 Cadre d'évaluation

## 1.1 Objectifs

Le cycle des évaluations disciplinaires réalisées sur échantillons (CEDRE) établit des bilans nationaux des acquis des élèves en fin d'école et en fin de collège. Il couvre les compétences des élèves dans la plupart des domaines disciplinaires en référence aux programmes scolaires. La présentation des résultats permet de situer les performances des élèves sur des échelles de niveau allant de la maîtrise pratiquement complète de ces compétences à une maîtrise bien moins assurée, voire très faible, de celles-ci. Renouvelées tous les six ans (tous les cinq ans à partir de 2012), ces évaluations permettent de répondre à la question de l'évolution du « niveau des élèves » au fil du temps.

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

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

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

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

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

## 1.2 Compétences visées

Les compétences permettant de cerner les acquis des élèves ont été retenues selon les finalités assignées à l'enseignement de la physique-chimie et des sciences de la vie et de la Terre. Une évaluation en sciences a pour objet de confronter les résultats du fonctionnement pédagogique du système éducatif aux objectifs qui lui sont assignés. L'évaluation doit envisager toutes les compétences inscrites au programme et l'ensemble des niveaux du collège.

A partir de ces finalités, trois types de compétences ont été retenus avec leurs composantes internes : les connaissances, les capacités et les attitudes [1].

### 1.2.1 Connaissances

- Montrer des connaissances : reconnaître une définition, une formule, une loi, une valeur
- Mobiliser ses connaissances en situation

### 1.2.2 Capacités

#### Mener une démarche

- Formuler un problème scientifique, à partir d'une situation donnée
- Identifier le caractère scientifique d'un problème
- Formuler une hypothèse
- Faire la différence entre une simulation et la réalité
- Proposer une expérience, son principe ou un modèle permettant de : formuler une conséquence vérifiable d'une hypothèse ; valider ou infirmer une hypothèse ; élaborer un protocole
- Établir une relation de cause à effet
- Comprendre qu'un effet peut avoir plusieurs causes agissant simultanément, percevoir qu'il peut exister des causes non apparentes ou inconnues
- Conclure sur la validité d'une hypothèse, conclure à partir de résultats

#### Manipuler et expérimenter

- Adapter un protocole
- Proposer la modification d'une expérience ou de paramètres pour atteindre un objectif

1. La terminologie utilisée fait donc référence aux définitions en vigueur avant le décret du 11 juillet 2006 sur le socle commun de connaissances et de compétences.

- Mettre en oeuvre un protocole : réaliser un réglage ; adapter la valeur d'une grandeur s'organiser dans le temps et dans l'espace pour mener à bien des expériences ; maîtriser des gestes manipulatoires ; Respecter les consignes de sécurité et d'organisation
- Établir un résultat expérimental
- Écarter une valeur erronée
- Exprimer un résultat avec une unité adaptée
- Associer une incertitude à une mesure
- Apprécier la nature et la validité d'un résultat statistique

**Exprimer et exploiter des données**

- Analyser des résultats expérimentaux, les confronter à des résultats théoriques attendus
- Utiliser correctement des connecteurs logiques
- Faire l'analyse critique d'un montage, d'un ordre de grandeur obtenu, d'un résultat
- Exprimer et (ou) exploiter les résultats sous la forme d'un schéma : lire un schéma ; traduire des informations par un schéma ; choisir un schéma parmi plusieurs propositions
- Exprimer et (ou) exploiter les résultats sous la forme d'un dessin scientifique
- Exprimer et (ou) exploiter les résultats sous la forme d'un modèle mathématique
- Exprimer et (ou) exploiter les résultats sous la forme d'un tableau : compléter, réaliser un tableau de mesures ; extraire des informations d'un tableau
- Exprimer et (ou) exploiter les résultats sous la forme d'un graphique : repérer des valeurs sur un graphique ; reconnaître sur un graphique le sens de variation d'une grandeur ; extraire des informations d'un graphique ; construire un graphique
- Exploiter un texte : extraire des informations pertinentes d'un texte ; identifier une question scientifique qui se pose dans un texte ; trier des informations d'un texte ; mettre en relation les informations extraites d'un texte pour répondre à une question ; utiliser un vocabulaire scientifique correct et adapté
- Exploiter une photo, un film, un croquis

### 1.2.3 Attitudes

- Observer des règles élémentaires de sécurité dans les domaines de la biologie, de la chimie et dans l'usage de l'électricité
- Faire preuve de responsabilité face à l'environnement, au monde vivant, à la santé

– Faire preuve d'esprit critique : distinguer le prouvé, le probable ou l'incertain ; distinguer prédiction et prévision ; situer un résultat ou une information dans son contexte
– Manifester de l'intérêt pour les progrès scientifiques et techniques

La déclinaison de ces domaines de compétences en composantes a permis de construire l'évaluation en lien avec les programmes en prenant en compte ce que chaque composante apporte à la maîtrise de la compétence. Un item met forcément en oeuvre différentes compétences que l'élève doit utiliser afin de répondre à la question. Il a donc été important de débattre entre concepteurs sur la compétence principale visée par l'item. Dans le cadre d'une évaluation bilan, ce découpage permet de positionner l'élève selon différents niveaux d'acquisition dans les différentes compétences et ainsi construire une échelle de performance pour les différents groupes de niveau.

## 1.3 Construction du test

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

### 1.3.1 Elaboration des questionnaires

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

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

### Exemples d'items

### Exemple 1 : série de vrai-faux

**Courir un 100 mètres**

Le 5 juillet 2005, Ronald Pognon remporte une course au meeting de Lausanne en 9 s 99. Il devient alors le premier Français à passer sous la barre des dix secondes.

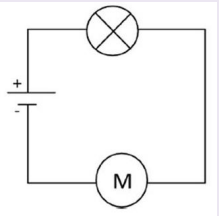
**Question 1**

Pour produire de l'énergie lors du départ de la course, les muscles des jambes du coureur ont besoin d'un apport important de :

Cocher vrai ou faux pour chaque proposition.

| | Vrai | Faux | |
|---|---|---|---|
| dioxygène. | ☐ 1 | ☐ 2 | C3XCV030101 |
| dioxyde de carbone. | ☐ 1 | ☐ 2 | C3XCV030102 |
| nutriments. | ☐ 1 | ☐ 2 | C3XCV030103 |

### Exemple 2 : question ouverte




**Question 4**

Dans le cadre ci-dessous, reproduire le schéma en y ajoutant l'appareil qui permet de mesurer l'intensité du courant qui circule dans le circuit. Préciser les bornes de cet appareil de mesure.

C3XEP190401

L'évaluation CEDRE sciences 2013 se présente sous la forme de QCM et de questions ouvertes (voir l'encadré « Exemples d'items »). Elle est composée d'items repris à l'identique par rapport à 2007 (103 items soit environ 60 % du test) et d'items nouveaux (72 items).

La spécificité de l'évaluation des sciences dans CEDRE, par rapport aux évaluations internationales et aux évaluations nationales d'autres pays, est la prise en compte des capacités expérimentales des élèves. En effet, l'évaluation comporte deux types d'épreuves : une épreuve « papier-crayon » et des travaux pratiques.

### 1.3.2 Constitution des cahiers

L'évaluation de 2013 est composée d'items de 2007, repris à l'identique permettant une comparaison diachronique, et d'items nouveaux expérimentés en 2012. L'évaluation de 2013 est composée de 13 blocs et 4 travaux pratiques répartis dans 17 cahiers selon la méthodologie des cahiers tournants. Chaque cahier est constitué de 4 blocs. 13 cahiers contiennent uniquement des épreuves de type « papier crayon », ces cahiers sont numérotés de 1 à 13 (tableau 2).

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

| *Cahier* | *Séquence 1* | | *Séquence 2* | | *Séquence 3* |
|---|---|---|---|---|---|
| E01 | *B20* | **B9** | *B21* | *B16* | QC |
| E02 | **B8** | *B21* | **B10** | **B12** | QC |
| E03 | **B9** | **B10** | *B22* | *B17* | QC |
| E04 | *B21* | *B22* | **B11** | **B13** | QC |
| E05 | **B10** | **B11** | *B16* | *B18* | QC |
| E06 | *B22* | *B16* | **B12** | *B19* | QC |
| E07 | **B11** | **B12** | *B17* | *B20* | QC |
| E08 | *B16* | *B17* | **B13** | **B8** | QC |
| E09 | **B12** | **B13** | *B18* | *B9* | QC |
| E10 | *B17* | *B18* | *B19* | *B21* | QC |
| E11 | **B13** | *B19* | *B20* | **B10** | QC |
| E12 | *B18* | *B20* | **B8** | *B22* | QC |
| E13 | *B19* | **B8** | **B9** | **B11** | QC |

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

Quatre autres cahiers contiennent une première séquence de type « papier crayon » et une deuxième séquence de type « travaux pratiques ». Ils sont numérotés de

« TP A » à « TP D ». « TP A » et « TP B » concernent la physique-chimie, tandis que « TP C » et « TP D » concernent les S.V.T (tableau 3).

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

Tableau 3 – Répartition des blocs dans les cahiers comprenant une séquence de « travaux pratiques »

| *Cahier TP* | *Séquence 1* | | *Séquence 2* | | *Séquence 3* |
|---|---|---|---|---|---|
| TP A | *B16* | *B20* | *TP1* | *TP2* | QC |
| TP B | *B20* | *B16* | *TP3* | *TP4* | QC |
| TP C | *B16* | *B20* | *TP5* | *TP6* | QC |
| TP D | *B20* | *B16* | *TP7* | *TP8* | QC |

Au final, pour l'évaluation CEDRE 2013, chaque cahier comprend deux séquences cognitives d'une durée d'environ une heure chacune. Elles sont complétées par une troisième séquence (questionnaire de contexte, QC), identique dans tous les cahiers, dans laquelle l'élève doit renseigner plusieurs éléments concernant l'environnement familial dans lequel il évolue, ses projets scolaires et professionnels, sa perception de la matière et de son environnement scolaire.

La partie travaux pratiques a été passée par environ 30 % des élèves (cf. la section sur l'échantillonnage). Les cahiers de ces élèves sont constitués d'une séquence « papier crayon » composée de deux blocs, d'une séquence « travaux pratiques » d'une durée d'environ une heure également et d'une troisième séquence correspondant au questionnaire de contexte.

Cette partie a été corrigée, dans la classe, par les professeurs des élèves grâce à un système de grille de correction. Il est donc important de tenir compte du fait que la correction n'est pas réalisée directement par la DEPP. Les résultats de la partie travaux pratiques ne participent donc pas à l'élaboration de l'échelle de performance ni au calcul du score des élèves. Néanmoins cette partie nous donne de bonnes indications sur les savoir-faire des élèves et il est ensuite possible de les relier aux contextes socio-scolaires.

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

**Objectifs**

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

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

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

**Principes fonctionnels**

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

## 1.4 Passation des évaluations

La passation de l'évaluation finale a eu lieu en mai 2013. Comme en 2007, cette évaluation a été précédée d'une expérimentation l'année $n-1$ de façon à tester un grand nombre d'items auprès d'un échantillon réduit d'établissements (141 collèges dont 137 répondants).

Dans chaque établissement, une personne a été désignée comme étant le coordi-

nateur, 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. Il est l'interlocuteur privilégié de la DEPP.

Entre chaque séquence, séparées par une pause, l'administrateur devait relever les cahiers, qui ne devaient pas être gardés par les élèves. Rappelons que les deux premières séquences interrogeaient les élèves sur leurs connaissances et compétences en sciences alors que la troisième séquence était une partie de « contexte » permettant d'éclairer les réponses des élèves et de nuancer certaines différences de niveaux qui peuvent apparaître (notamment entre types d'établissements fréquentés).

Le professeur de sciences de la vie et de la Terre ou de physique-chimie devait préparer la séquence « travaux pratiques ». Un manuel de passation par TP proposé était donc fourni avec l'ensemble des cahiers. Ce manuel contient la grille d'évaluation ainsi qu'une description exhaustive :
– du rôle du professeur lors de cette séquence,
– du matériel nécessaire à la préparation de la séquence,
– de l'installation de la salle et du matériel.
Les professeurs de sciences de la vie et de la Terre ou de physique-chimie de la classe ou des classes concernées ont également dû renseigner un questionnaire de contexte.

L'anonymat des élèves et des personnels a été respecté, chaque cahier étant repéré par un numéro allant de 1 au nombre total d'élèves ou d'enseignants ayant répondu dans le collège. Une fois l'évaluation terminée, les cahiers et questionnaires devaient être renvoyés dans des conditionnements prévus à cet effet, pré-affranchis et pré-étiquetés.

# 2 Sondage

## 2.1 Méthodes

### 2.1.1 Tirage équilibré de classes de 3e

De manière générale, pour le secondaire, deux options de tirage peuvent être considérées : soit un sondage par grappe en sélectionnant un échantillon de classes et tous les élèves des classes tirées au sort participant à l'évaluation ; soit un premier degré qui concerne les établissements puis un second degré où un nombre d'élèves fixe dans chaque établissement est sélectionné [2]. Les évaluations CEDRE suivent la première option tandis que l'évaluation PISA suit la seconde. Des simulations ont permis de montrer que les niveaux de précision des deux options sont très proches, dès lors que le tirage est équilibré (cf. encadré « Tirage d'établissement *versus* tirage de classes »). Le choix de sondages par grappe est motivé par la facilité de gestion. En effet, le fait de sélectionner tous les élèves d'une classe au collège permet d'éviter de mettre en place des procédures de tirage au sort d'élèves une fois les établissements tirés.

On note $U$ la population visée par une évaluation donnée, $Y$ la variable d'intérêt (typiquement le score à l'évaluation, ou bien une indicatrice de difficulté), $X$ une variable auxiliaire, c'est-à-dire connue pour l'ensemble des élèves de la population $U$. Un échantillon $S$ d'élèves est sélectionné dans la population $U$. Chaque élève $i$ a la probabilité $\pi_i$ d'être sélectionné dans l'échantillon $S$ (probabilité d'inclusion). Enfin, les poids de sondages, définis comme les inverses des probabilités d'inclusion $\pi_i$, sont notés $d_i$.

Un échantillon équilibré est un échantillon qui est représentatif de la population au regard de certaines variables auxiliaires. Cela signifie que dans un échantillon équilibré, l'estimateur du total d'une variable auxiliaire $X$ sera exactement égal au vrai total de la variable $X$ dans la population.

Cette propriété s'écrit :

$$\sum_{i \in S} \frac{X_i}{\pi_i} = \sum_{i \in U} X_i \tag{1}$$

2. Dans ce second cas, les établissements sont tirés proportionnellement à leur taille (nombre d'élèves). En effet, une fois que les établissements sont échantillonnés, un nombre fixe d'élèves est alors sélectionné quel que soit l'établissement. Par conséquent, les élèves des grands établissements ont moins de chance d'être tirés au sort que les élèves des petits établissements. Le tirage proportionnel à la taille permet ainsi de rétablir l'égalité des probabilités de tirage.

**Tirage d'établissements *versus* Tirage de classes**

Pour faciliter la logistique dans les collèges, nous réalisons un de classes de 3e, puis tous les élèves de la classe sélectionnée passent l'évaluation. On peut donc s'interroger sur la perte de la précision liée à cet effet de grappe. Pour comparer la précision entre un tirage d'établissement et un tirage de classes, nous avons réalisé des simulations à partir de la base des notes au brevet en 2009 (Garcia, Le Cam, & Rocher, à paraître).
Nous avons comparé deux stratégies d'échantillonnage. Il s'agit à chaque fois d'échantillons stratifiés à deux degrés :
– Tirage équilibré d'établissement puis tirage de 30 élèves dans chaque établissement sélectionné ;
– Tirage équilibré de classe puis sélection de tous les élèves des classes sélectionnées.
La stratification a été effectuée selon le secteur d'enseignement et dans chaque strate 2 000 élèves ont été échantillonnés.
Pour chacune des deux stratégies, 1 000 échantillons ont été tirés. Puis on calcule la moyenne des erreurs standards des notes moyennes en français, mathématiques et histoire-géographie. Le tableau ci-dessous montre que les deux stratégies de tirage ont des niveaux équivalents de précision.

**Comparaison des erreurs standards (Garcia et al., à paraître)**

| | Echantillon équilibré d'établissements | Echantillon équilibré de classes |
|---|---|---|
| Français | 0,07 | 0,07 |
| Mathématiques | 0,11 | 0,11 |
| Histoire-Géographie | 0,08 | 0,08 |

Les échantillons équilibrés ont donc comme propriété de fournir une photographie parfaite de la population, au regard des variables auxiliaires connues, ce que ne garantit pas une procédure aléatoire simple d'échantillonnage. En théorie, ils permettent également d'améliorer la précision des estimateurs s'il existe un lien entre la variable d'intérêt et les variables auxiliaires.

Le tirage équilibré est réalisé grâce au programme CUBE développé par l'INSEE et mis à disposition sous forme de macro SAS. La documentation complète est disponible sur le site Internet de l'INSEE (Rousseau & Tardieu, 2004). L'algorithme permet de choisir de manière aléatoire un échantillon parmi tous les échantillons possibles respectant les contraintes reposant sur les variables

auxiliaires. Il se déroule en deux phases : une « phase de vol » et une « phase d'atterrissage ». Durant la phase de vol, toutes les contraintes sont respectées. Elle se termine si un échantillon équilibré de manière parfaite est trouvé ou s'il n'est pas possible de trouver un échantillon en respectant toutes les contraintes. Si la phase de vol n'a pas abouti à un échantillon, la phase d'atterrissage débute. Elle consiste au relâchement des contraintes et au choix optimal de l'échantillon selon le critère choisi par l'utilisateur (ordre de priorité sur les contraintes, relâchement de la contrainte avec un coût minimal sur l'équilibrage ou garantie d'un échantillon de taille fixe).

Par ailleurs, au moment du tirage de l'échantillon, les collèges dont une classe a déjà été sélectionnée pour une autre évaluation la même année sont exclus 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 équilibré après élimination de la base des échantillons précédemment tirés »).

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

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

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

$$\forall k = 1...K, \sum_{i \in S'} w_i X_i^k = \sum_{i \in U} X_i^k \tag{2}$$

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 2.

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

**Formulation du problème et 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 $\lambda_j = \frac{\pi_j^2}{1-\pi_j^1}$

**Équilibrage**

On souhaite maintenant que l'échantillon $S_2$ soit équilibré selon certaines

variables (nombre d'élèves en retard, etc.). Soit $X$ une variable d'équilibrage, la condition s'écrit :

$$\sum_{j \in S_2} \frac{X_j}{\pi_j^2} = \sum_{j \in U} X_j$$

Pour arriver à ce résultat, le principe est de tirer $S_2$ dans $U(S_1)$ avec les probabilités d'inclusion $\lambda_j$ et avec une condition d'équilibrage sur la variable $X_j/(1-\pi_j^1)$.

Ainsi, on aura :

$$\sum_{j \in S_2} \frac{X_j}{\pi_j^2} = \sum_{j \in S_2} \frac{X_j}{\lambda_j(1-\pi_j^1)} = \sum_{j \in U(S_1)} \frac{X_j}{1-\pi_j^1}$$

Or, en espérance on a

$$E(\sum_{j \in U(S_1)} \frac{X_j}{1-\pi_j^1}) = E(\sum_{j \in U} \frac{X_j}{1-\pi_j^1} I_{j \notin S_1}) = \sum_{j \in U} X_j$$

La condition d'équilibrage initiale est donc remplie.

**Condition fondamentale**

Comme il s'agit d'une probabilité, la condition fondamentale est que $\lambda_j \in [0,1]$. Comme $\lambda_j = \frac{\pi_j^2}{1-\pi_j^1}$, la condition est en fait que

$$\pi_j^1 + \pi_j^2 \leq 1$$

Dans certains cas, par exemple des strates souvent sur-représentées comme les établissements situés dans des zones spécifiques concernant peu d'élèves (ex : Eclair), cette condition pourrait ne pas être satisfaite. Cependant, de façon concrète, la condition a toujours été respectée dans les plans de sondage réalisés.

### 2.1.3 Calcul de précision : méthode

Les résultats des évaluations sont soumis à une variabilité qui dépend notamment des erreurs d'échantillonnage. Il est possible d'estimer statistiquement ces erreurs d'échantillonnage, appelées erreurs standard.

On note $Y$ la variable d'intérêt (typiquement le score obtenu à une évaluation) et $\hat{\bar{Y}}$ l'estimateur de la moyenne de $Y$, qui constitue un estimateur essentiel sur lequel nous insistons dans la suite, bien que d'autres soient également au centre des analyses, comme ceux concernant la dispersion. La méthode retenue est cependant applicable à différents types d'estimateurs.

Nous souhaitons estimer la variance de cet estimateur, c'est-à-dire $V(\hat{\bar{Y}})$. En absence de formule théorique pour calculer $V(\hat{\bar{Y}})$, il existe plusieurs procédures permettant de l'estimer, c'est-à-dire de calculer $\hat{V}(\hat{\bar{Y}})$, l'estimateur de la variance d'échantillonnage. Il peut s'agir de méthodes de linéarisation des formules (Taylor) ou bien de méthodes empiriques (méthodes de réplication, jackknife, etc.). Ces méthodes sont bien décrites dans la littérature. Le lecteur est invité à consulter Tillé (2001) ou Ardilly (2006).

Cependant, lorsqu'un calage sur marges a été effectué, il faut en tenir compte pour le calcul de la précision. Dans ce cas, la variance de $\hat{\bar{Y}}$ est asymptotiquement équivalente à la variance des résidus de la régression de la variable d'intérêt sur les variables de calage.

En pratique, pour estimer la variance d'échantillonnage de $\hat{\bar{Y}}$, tenant compte du calage effectué, il convient alors d'appliquer la procédure suivante :

1. On effectue la régression linéaire de la variable d'intérêt sur les variables de calage, en pondérant par les poids initiaux. Les résidus $e_i$ de cette régression sont calculés.
2. Les valeurs $g_i e_i$ sont calculées, où $g_i$ représente le rapport entre les poids CALMAR ($w_i$) et les poids initiaux ($d_i$) : $g_i = \frac{w_i}{d_i}$
3. La variance d'échantillonnage de $\hat{\bar{Y}}$ est alors obtenue en calculant la variance d'échantillonnage de $g_i e_i$.

## 2.2 Echantillonnage

Le champ des évaluation CEDRE au collège est celui des élèves de 3e générale scolarisés dans des collèges publics et privés sous contrat de France métropolitaine.

La base de sondage utilisée est la base dite Scolarité construite par la DEPP. C'est une base de données individuelles anonymes contenant de nombreuses informations sur les élèves scolarisés une année scolaire donnée (date de naissance, PCS des parents, etc.). Nous disposons également d'informations sur les établissements scolaires, comme par exemple le secteur d'enseignement. Ces informations, qualifiées de variables auxiliaires, peuvent être utilisées au moment du tirage des échantillons, pour définir les variables de stratification.

### 2.2.1 Echantillon 2007

#### Modalités de sélection

Le tirage est à deux degrés. Le premier degré de sondage est composé de collèges tirés dans chaque strate selon un tirage à allocation proportionnelle. Le second degré consiste à tirer une ou deux classes dans les collèges sélectionnés.

Préalablement au tirage, les échantillons CEDRE Sciences Expérimentation 2006, CEDRE Histoire-Géographie 2006, CEDRE Mathématiques Expérimentation 2007 et l'évaluation de rentrée de 6e de septembre 2006 ont été retirés.

#### Stratification

La stratification prend en compte à la fois la taille et le secteur d'enseignement du collège :

1. Petits collèges privés (2 classes ou moins)
2. Grands collèges privés (3 classes ou plus)
3. Petits collèges public hors zep (2 classes ou moins)
4. Grands collèges public hors zep (3 classes ou plus)
5. Collèges publics en ZEP ou en REP

Dans les strates 1 et 3, toutes les classes de 3ème générale passent l'évaluation. Dans les strates 2, 4 et 5 deux classes sont sélectionnées. On vise environ 9 500 élèves.

#### Base de sondage

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

Tableau 4 – Répartition dans la base de sondage (2006-2007)

| strate | collèges | élèves |
|---|---|---|
| 1. Petits collèges privés | 615 | 22 162 |
| 2. Grands collèges privés | 1 046 | 124 946 |
| 3. Petits collèges public hors zep | 508 | 20 192 |
| 4. Grands collèges public hors zep | 3 501 | 424 560 |
| 5. Collèges publics en ZEP ou en REP | 1 030 | 106 717 |
| **Total** | **6 700** | **698 577** |

#### Échantillon

Le tableau 5 présente la répartition de l'échantillon dans les différentes strates. Au total, 394 classes ont été sélectionnées dans 200 établissements.

Tableau 5 – Répartition dans l'échantillon 2007

| strate | collèges | classes | élèves |
|---|---|---|---|
| 1. Petits collèges privés | 15 | 27 | 598 |
| 2. Grands collèges privés | 45 | 90 | 2 412 |
| 3. Petits collèges public hors zep | 15 | 27 | 606 |
| 4. Grands collèges public hors zep | 75 | 150 | 3 734 |
| 5. Collèges publics en ZEP ou en REP | 50 | 100 | 2 171 |
| **Total** | **200** | **394** | **9 521** |

### 2.2.2 Echantillon 2013

#### Modalités de sélection

Le tirage est à deux degrés. Le premier degré de sondage est composé de classes (et non de collèges) tirées dans chaque strate avec allocation proportionnelle. Le deuxième degré de sondage consiste à interroger tous les élèves de la classe sélectionnée (tirage par grappe).

Dans chacune des 3 strates, le tirage est équilibré sur les variables suivantes :
- Le nombre total d'élèves de 3e
- Le nombre d'élèves de 3e de PCS de référence « défavorisée »
- Le nombre d'élèves de 3e en retard dans la population
- Le nombre de garçons de 3e dans la population

#### Stratification

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

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

On vise environ 10 000 élèves.

#### Champ et exclusions

Pour l'année 2013, nous documentons le champ de l'évaluation qui est l'ensemble des élèves de 3e générale de collèges de France métropolitaine (tableau 6).

Tableau 6 – Exclusions pour la base de sondage

| | Etab. | Elèves |
|---|---|---|
| Etablissements accueillant des élèves de 3e | 8 370 | 805 990 |
| On retire les EREA | 8 300 | 804 623 |
| On retire les étab hors contrat | 8 141 | 802 343 |
| On retire les TOM | 8 104 | 798 239 |
| On ne garde que les collèges | 6 922 | 768 269 |
| 3e gnale en France métropolitaine | 6 689 | 709 278 |
| **Total des exclusions** | | |
| **Base CEDRE Sciences 3e** | **6 689** | **709 278** |

### Base de sondage

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

Tableau 7 – Répartition dans la base de sondage (2012-2013)

| strate | collèges | classes | élèves |
|---|---|---|---|
| 1. Public hors EP | 4 067 | 18 005 | 458 455 |
| 2. Ep | 985 | 4 370 | 97 354 |
| 4. Privé | 1 637 | 5 917 | 153 469 |
| **Total** | **6 689** | **28 292** | **709 278** |

### Échantillon

Le tableau 8 présente la répartition de l'échantillon dans les différentes strates. Au total, 399 classes ont été sélectionnées dans 389 établissements, rassemblant 10 159 élèves (après correction des effectifs pris en compte dans la base, ajoutant ainsi 166 élèves à l'échantillon prévu au départ).

Tableau 8 – Répartition dans l'échantillon 2013

| strate | collèges | classes | élèves |
|---|---|---|---|
| 1. Public hors EP | 252 | 254 | 6 563 |
| 2. EP | 58 | 62 | 1 409 |
| 3. Privé | 79 | 83 | 2 187 |
| **Total** | **389** | **399** | **10 159** |

## 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 de classes entières ou la non-réponse d'élèves dans les classes participantes. Les chiffres suivants ont été observés pour 2013. Tout d'abord, 94,5 % des classes de l'échantillon ont répondu à l'évaluation. Les 22 classes non répondantes représentent 511 élèves (tableau 9).

Tableau 9 – Non réponse des classes

| strate | N classes attendues | N classes répondantes | % de classes répondantes | N élèves non répondants |
|---|---|---|---|---|
| 1- public hors EP | 254 | 244 | 96,1% | 240 |
| 2- EP | 62 | 59 | 95,2% | 21 |
| 3- privé | 83 | 74 | 89,2% | 50 |
| **Total** | **399** | **377** | **94,5%** | **511** |

Parmi les classes ayant répondu, 89,7 % des élèves ont participé à l'évaluation (tableau 10).

Tableau 10 – Non réponse des élèves

| strate | N élèves attendus classes répondantes | N élèves répondants classes répondantes | % élèves répondants |
|---|---|---|---|
| 1- public hors EP | 6 323 | 5 669 | 89,7% |
| 2- EP | 1 388 | 1 192 | 85,9% |
| 3- privé | 1 937 | 1 793 | 92,6% |
| **Total** | **9 648** | **8 654** | **89,7%** |

Au final, 85,2 % des effectifs attendus ont participé (tableau 11).

Tableau 11 – Non réponse globale (classes + élèves)

| strate | N élèves attendus | N élèves répondants | % élèves répondants |
|---|---|---|---|
| 1- public hors EP | 6 563 | 5 669 | 86,4% |
| 2- EP | 1 409 | 1 192 | 84,6% |
| 3- privé | 2 187 | 1 793 | 82,0% |
| **Total** | **10 159** | **8 654** | **85,2%** |

### 2.3.2 Valeurs manquantes et imputation

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

### 2.3.3 Non-réponse partielle et terminale

Lorsque des non-réponses sont observées aux items, nous distinguons les cas suivants :
- La non-réponse partielle : un élève n'a pas répondu à certains items dans le cahier.
- La non-réponse terminale : un élève s'est arrêté avant la fin du cahier soit par manque de temps soit par abandon.

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

**2007**

En 2007, les cahiers étaient composés de deux séquences. La non réponse terminale a été étudiée par séquence et par cahier. Parmi les élèves ayant de la non-réponse terminale, il y en a en moyenne 5,2 pour la 1ère séquence et 7 pour la 2ème séquence.

Au final, pour 2007, on considère que :
- 19 élèves n'ont pas vu la séquence 1 dont :
  - 11 n'ont répondu à aucun item de la séquence
  - 8 ont répondu à moins de 50 % de la séquence
- 33 élèves n'ont pas vu la séquence 2 dont :
  - 22 n'ont répondu à aucun item de la séquence
  - 11 ont répondu à moins de 50 % de la séquence

Les élèves dont les deux séquences sont codées en " s " sont considérés comme de la non réponse totale. C'est le cas pour 2 élèves.

**2013**

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 2,5 pour la 1ère séquence et 2,9 pour la 2ème séquence.

Au final, on considère que :
- 187 élèves n'ont pas vu la séquence 1 dont :
  - 175 n'ont répondu à aucun item de la séquence
  - 12 ont répondu à moins de 50 % de la séquence
- 292 élèves n'ont pas vu la séquence 2 dont :
  - 269 n'ont répondu à aucun item de la séquence
  - 23 ont répondu à moins de 50 % de la séquence

Les élèves dont les deux séquences sont codées en " s " sont considérés comme de la non réponse totale. C'est le cas pour 138 élèves.

On élimine également 18 élèves qui n'ont passé que la partie TP des cahiers, ces élèves n'auront aucun item pour le calcul de score.

## 2.4 Redressement

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

Puis, deux variables de calage sont utilisées :
- la répartition selon le sexe dans la population ;
- la répartition selon le retard scolaire.

Le tableau 12 montre que le calage concerne principalement les élèves en retard, plus souvent absents à l'évaluation et donc moins nombreux dans l'échantillon que dans la population (18,7 % contre 21,1 %).

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

| | Modalité ou variable | Marge échantillon | Marge population | Pourcentage échantillon | Pourcentage population |
|---|---|---|---|---|---|
| Retard | 1 | 132 871.98 | 149 937 | 18.73 | 21.14 |
| | 2 | 576 406.02 | 559 341 | 81.27 | 78.86 |
| Sexe | 1 | 347 432.34 | 352 694 | 48.98 | 49.73 |
| | 2 | 361 845.66 | 356 584 | 51.02 | 50.27 |
| Strate | 1 | 458 455 | 458 455 | 64.64 | 64.64 |
| | 2 | 97 354 | 97 354 | 13.73 | 13.73 |
| | 3 | 153 469 | 153 469 | 21.64 | 21.64 |

## 2.5 Précision

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

Tableau 13 – Scores moyens en sciences et erreurs standard associées

| Année | Score moyen | Erreur standard |
|---|---|---|
| 2007 | 250 | 1.85 |
| 2013 | 249 | 0.92 |

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

$$\frac{\left|\hat{\bar{Y}}_{2013} - \hat{\bar{Y}}_{2007}\right|}{\sqrt{se^2_{\hat{\bar{Y}}_{2013}} + se^2_{\hat{\bar{Y}}_{2007}}}} \tag{3}$$

Avec une valeur de 0,34 (inférieure à 1,96), cela signifie que la baisse du score moyen observée entre 2007 et 2013 n'est pas statistiquement significative.

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

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

| Année | Groupe < 1 | Groupe 1 | Groupe 2 | Groupe 3 | Groupe 4 | Groupe 5 |
|---|---|---|---|---|---|---|
| 2007 | 2.0 | 13.0 | 28.9 | 28.8 | 17.3 | 10.0 |
| 2013 | 2.1 | 12.3 | 28.2 | 30.8 | 18.3 | 8.4 |

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

| Année | Groupe < 1 | Groupe 1 | Groupe 2 | Groupe 3 | Groupe 4 | Groupe 5 |
|---|---|---|---|---|---|---|
| 2007 | 0.41 | 0.82 | 0.78 | 0.80 | 0.76 | 0.65 |
| 2013 | 0.22 | 0.53 | 0.64 | 0.58 | 0.54 | 0.46 |

### *Design effect*

L'effet du plan de sondage (*Design Effect*) permet de rapporter l'erreur de mesure faite par un tirage spécifique à l'erreur de mesure qui aurait été faite en

procédant à un sondage aléatoire simple (SAS) du même nombre d'élèves. Pour la moyenne d'une variable Y et un plan de sondage complexe P, il est défini par :

$$D_{eff} = \frac{V_P(\hat{\bar{Y}})}{V_{SAS}(\hat{\bar{Y}})} \tag{4}$$

Dans le cas d'un sondage en grappes, la précision est dégradée en comparaison d'un sondage aléatoire simple. L'effet du plan de sondage est donc supérieur à 1 (tableau 16).

Tableau 16 – Effet du plan de sondage

| Année | Erreur Standard | Erreur SAS | *Design Effect* |
|---|---|---|---|
| 2007 | 1.85 | 0.56 | 10.87 |
| 2013 | 0.92 | 0.51 | 3.23 |

Cela signifie qu'en 2013, un sondage aléatoire simple avec un effectif 3 fois moins important aurait conduit au même niveau de précision. En 2007, ce rapport est encore plus élevé car le tirage était en fait à trois degrés (collèges puis classes puis tous les élèves) et non en grappes.

# 3 Analyse des items

## 3.1 Méthodologie

### 3.1.1 Approche classique

Dans un premier temps, nous posons quelques notations et nous présentons les principales statistiques descriptives utilisées pour décrire un test, issues de la « théorie classique des tests » que nous évoquons rapidement.

**Réussite et score**

On note $n$ le nombre d'élèves ayant passé une évaluation composée de $J$ items. On note $Y_i^j$ la réponse de l'élève $i$ ($i = 1, ..., n$) à l'item $j$ ($j = 1, ..., J$). Dans notre cas, les items sont dichotomiques, c'est-à-dire qu'ils ne prennent que deux modalités (la réussite ou l'échec) :

$$Y_i^j = \begin{cases} 1 & \text{si l'élève } i \text{ réussit l'item } j \\ 0 & \text{si l'élève } i \text{ échoue à l'item } j \end{cases} \tag{5}$$

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

$$p_j = \frac{1}{n} \sum_{i=1}^{n} Y_i^j \tag{6}$$

Le taux de réussite d'un item renvoie à son niveau de difficulté. C'est certainement la caractéristique la plus importante, qui permet de construire un test de niveau adapté à l'objectif de l'évaluation, en s'assurant que les différents niveaux de difficulté sont balayés.

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

$$S_i = \sum_{j=1}^{J} Y_i^j \tag{7}$$

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 [3].

Le coefficient $\alpha$ de Cronbach est un indice destiné à mesurer la fidélité de l'épreuve. Il est compris entre 0 et 1. Sa version « standardisée » s'écrit :

$$\alpha = \frac{J\bar{r}}{1 + (J-1)\bar{r}} \tag{8}$$

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 [4].

### Indices de discrimination

Des indices importants concernent le pouvoir discriminant des items. Nous présentons ici l'indice « r-bis point » ou coefficient point-bisérial qui est le coefficient de corrélation linéaire entre la variable indicatrice de réussite à l'item $Y^j$ et le score $S$.

Appelé également « corrélation item-test », il indique dans quelle mesure l'item s'inscrit dans la dimension générale. Une autre manière de l'envisager consiste à le formuler en fonction de la différence de performance constatée entre les élèves qui réussissent l'item et ceux qui l'échouent. En effet, on peut montrer que

$$r_{bis-point}(j) = corr(Y^j, S) = \frac{\bar{S}_{(j1)} - \bar{S}_{(j0)}}{\sigma_S}\sqrt{p_j(1-p_j)} \tag{9}$$

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.

3. 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.

4. 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.

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) [5]. 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. Les réponses à un item dichotomique sont définies de la manière suivante :

$$y_{ij} = \begin{cases} 0 & \text{si } z_{ij} \leq \tau_j \\ 1 & \text{si } z_{ij} > \tau_j \end{cases} \tag{10}$$

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

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

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

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

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

## 3.2 Codage des réponses aux items

### 3.2.1 Valeurs manquantes

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

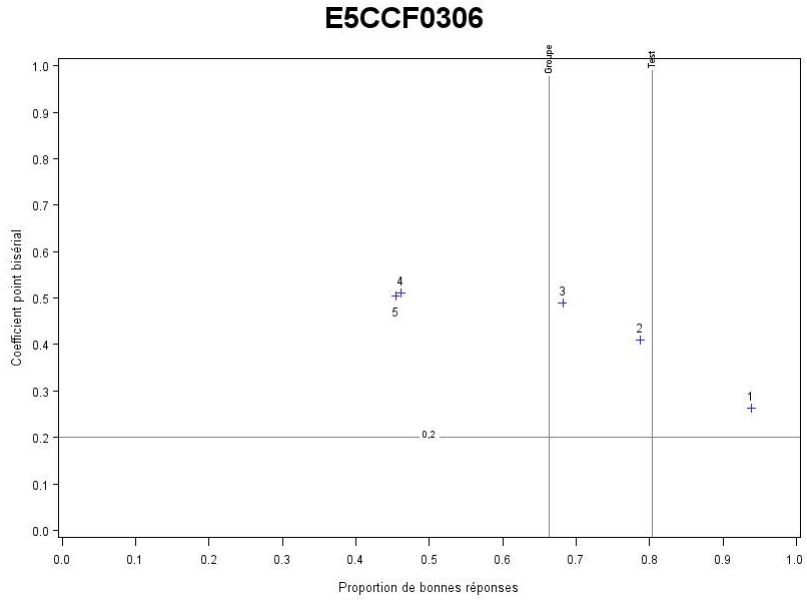
### 3.2.2 Regroupement des items

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

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

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

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




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

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

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

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

**AGATE : un outil de correction à distance des questions ouvertes**

**Objectifs**

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

**Principes fonctionnels**

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

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

## 3.3 Résultats

### 3.3.1 Pouvoir discriminant des items

Rappelons la répartition des items pour CEDRE Sciences 3e :
– 103 items communs aux évaluations 2007 et 2013 ;
– 72 items 2013 ;
– 104 items 2007 non repris en 2013.

Nous avons éliminé 33 items en raison d'un faible indice *rbis-point* :
– 13 items communs ;
– 1 items 2013 ;
– 19 items 2007.

### 3.3.2 Dimensionnalité

Le tableau 17 présente les résultats de l'analyse factorielle des items effectuée sur l'année 2013. La structure des items est fortement unidimensionnelle : le « poids » de la première dimension est très important (valeur propre de 32,9 contre 3,6 pour la deuxième dimension).

Tableau 17 – Analyse en composantes principales

| | Valeur Propre | Différence | Proportion | Proportion cumulée |
|---|---|---|---|---|
| 1 | 37.9 | 33.1 | 0.177 | 0.177 |
| 2 | 4.8 | 0.28 | 0.022 | 0.199 |
| 3 | 4.5 | 0.66 | 0.021 | 0.220 |

# 4 Modélisation

## 4.1 Méthodologie

### 4.1.1 Modèle de réponse à l'item

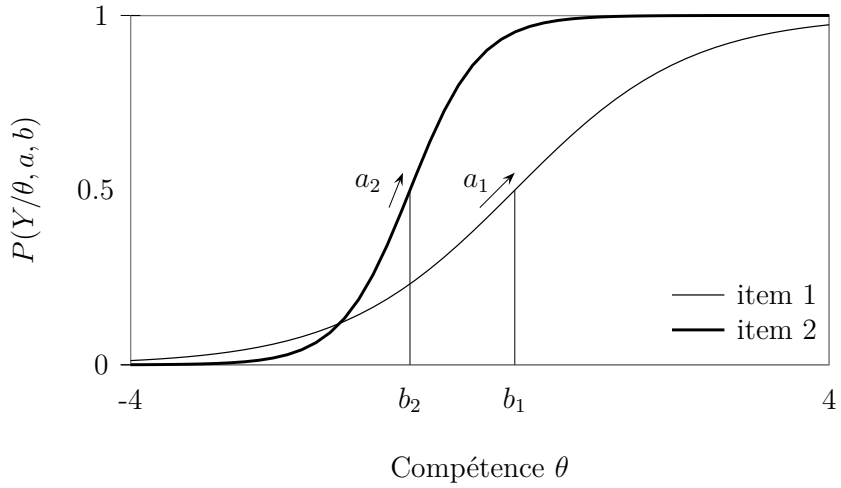
Le modèle de mesure utilisé est un modèle de réponse à l'item à deux paramètres avec une fonction de lien logistique (MRI 2PL) :

$$P_{ij} = P(Y_i^j = 1 \,|\, \theta_i, a_j, b_j) = \frac{e^{1,7a_j(\theta_i - b_j)}}{1 + e^{1,7a_j(\theta_i - b_j)}} \tag{11}$$

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

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

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




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

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

ont un intérêt pratique pour la construction de tests et la comparaison entre différents groupes d'élèves : si le modèle est bien spécifié sur un échantillon donné, les paramètres des items – en particulier leurs difficultés – peuvent être considérés comme fixes et applicables à d'autres échantillons dont il sera alors possible de déduire les paramètres relatifs aux élèves – en particulier, leur niveau de compétence.

Autre avantage : le niveau de compétence des élèves et la difficulté des items sont placés sur la même échelle, par le simple fait de la soustraction $(\theta_i - b_j)$. Cette propriété permet d'interpréter le niveau de difficulté des items par rapprochement avec le continuum de compétence. Ainsi, les élèves situés à un niveau de compétence égal à $b_j$ auront 50 % de chances de réussir l'item, ce que traduit visuellement la représentation des courbes caractéristiques des items (CCI) selon ce modèle (figure 2).

### 4.1.2 Procédures d'estimation

L'estimation est conduite en deux temps : l'estimation des paramètres des items puis l'estimation des $\theta$ en considérant les paramètres des items comme fixes. Nous donnons ici des éléments concernant ces procédures.

**Estimation des paramètres des items**

Nous reprenons les notations de l'équation (11) 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 poour l'année 2007).

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

$$L(\mathbf{y}, \xi, \theta) = \prod_{i=1}^{n} \prod_{j=1}^{J} P_{ij}^{y_{ij}} [1 - P_{ij}]^{1-y_{ij}} \tag{12}$$

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 utilisées. Elle consiste à estimer les paramètres des items en supposant que les paramètres des individus

sont issus d'une distribution fixée *a prori* (le plus souvent normale). La maximisation de vraisemblance est *marginale* dans le sens où les paramètres concernant les individus n'apparaissent plus dans la formule de vraisemblance.

Si $\theta$ est considérée comme une variable aléatoire de distribution connue, la probabilité inconditionnelle d'observer un *pattern* $\mathbf{y}_i$ donné peut s'écrire :

$$P(\mathbf{y} = \mathbf{y}_i) = \int_{-\infty}^{+\infty} P(\mathbf{y} = \mathbf{y}_i \,|\, \theta_i) g(\theta_i) d\theta_i \tag{13}$$

avec $g$ la densité de $\theta$.

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

$$L = \prod_{i=1}^{n} P(\mathbf{y} = \mathbf{y}_i) \tag{14}$$

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 (12).

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

$$B(\theta_i^{(ML)}) = \frac{-J}{2I^2} \tag{15}$$

avec

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

et

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

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

$$\frac{\partial lnL}{\partial \theta_i} + \frac{J}{2I} = 0 \tag{16}$$

### 4.1.3 Indice d'ajustement (FIT)

L'ajustement des items au modèle est étudié. Graphiquement, cela revient à comparer les courbes caractéristiques estimées avec les résultats observés (cf. figure 3). Certaines procédures proposent de comparer directement les probabilités théorique avec les proportions de réussite de groupes d'élèves. Plus généralement, nous pouvons écrire les résidus de la manière suivante :

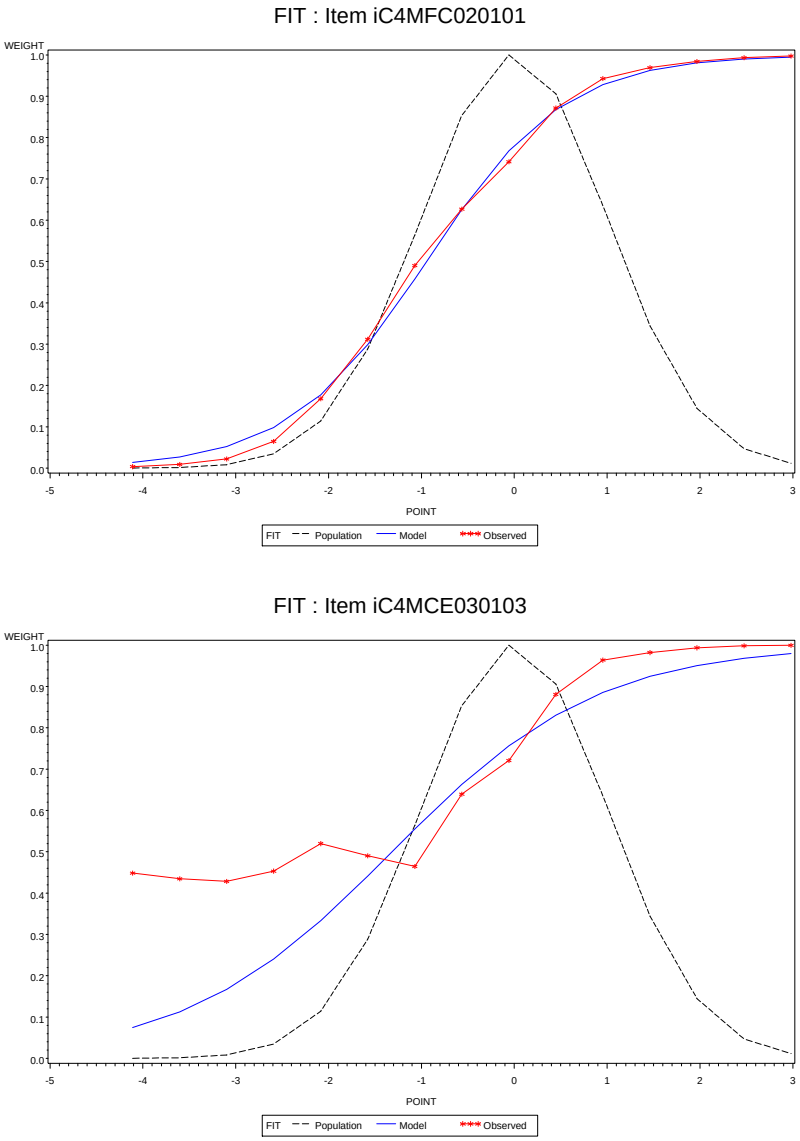
$$z_{ij} = \frac{Y_i^j - P_{ij}}{\sqrt{P_{ij}(1-P_{ij})}} \tag{17}$$

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

$$Infit_j = \frac{1}{\sum_{i=1}^{n} w_{ij}} \sum_{i=1}^{n} w_{ij} z_{ij}^2 = \frac{1}{\sum_{i=1}^{n} w_{ij}} \sum_{i=1}^{n} (Y_i^j - P_{ij})^2 \tag{18}$$

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

Figure 3 – Exemples d'ajustements (FIT)




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

### 4.1.4 Fonctionnement Différentiel d'Item (FDI)

Un fonctionnement différentiel d'item (FDI) apparaît entre des groupes d'individus dès lors qu'à niveau égal sur la variable latente mesurée, la probabilité de réussir un item donné n'est pas la même selon le groupe considéré. La question des FDI est importante car elle renvoie à la notion d'équité entre les groupes : un test ne doit pas risquer de favoriser un groupe par rapport à un autre.

Une définition formelle du FDI peut s'envisager à travers la propriété d'invariance conditionnelle : à niveau égal sur la compétence visée, la probabilité de réussir un item donné est la même quel que soit le groupe de sujets considéré. Formellement, un fonctionnement différentiel se traduit donc par :

$$P(Y \mid Z, G) \neq P(Y \mid Z) \tag{19}$$

où $Y$ est le résultat d'une mesure de la compétence visée, typiquement la réponse à un item ; Z est un indicateur du niveau de compétence des sujets ; G est un indicateur de groupes de sujets.

La probabilité de réussite, conditionnellement au niveau mesuré, est identique pour tous les groupes de sujets.

En pratique, de très nombreuses méthodes ont été proposées afin d'identifier les FDI. Ces méthodes ont chacune des avantages en matière d'investigation des différents éléments pouvant conduire à l'apparition de ces FDI (Rocher, 2013). Dans le cas des évaluations standardisées menées à la DEPP, il s'agit avant tout d'identifier les fonctionnements différentiels pouvant apparaître entre deux moments de mesure, s'agissant des items repris à l'identique. Dans ce cas, les différentes méthodes d'identification donnent des résultats relativement proches.

Une stratégie très simple, employée dans CEDRE, consiste donc à comparer les paramètres de difficulté des items repris, estimés de façon séparée pour les deux années. Si la difficulté d'un item a évolué, comparativement aux autres items, c'est le signe d'un fonctionnement différentiel, qui peut être lié par exemple à un changement de programmes ou de pratiques. Plus précisément, les paramètres des items sont estimés séparément pour les deux années, puis ajustés en tenant compte de la différence moyenne entre les deux séries de paramètres. La règle retenue pour identifier un FDI est celle d'un écart de paramètres de difficulté $\beta$ d'au moins 0,5 (cf. Rocher, 2013 pour plus de détails).

### 4.1.5 L'information du test

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

$$I_j(\theta) = (1,7a_j)^2 P_j(\theta)(1 - P_j(\theta)) \tag{20}$$

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 d'information du test est tracée pour un ensemble de valeurs de $\theta$.

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

## 4.2 Résultats

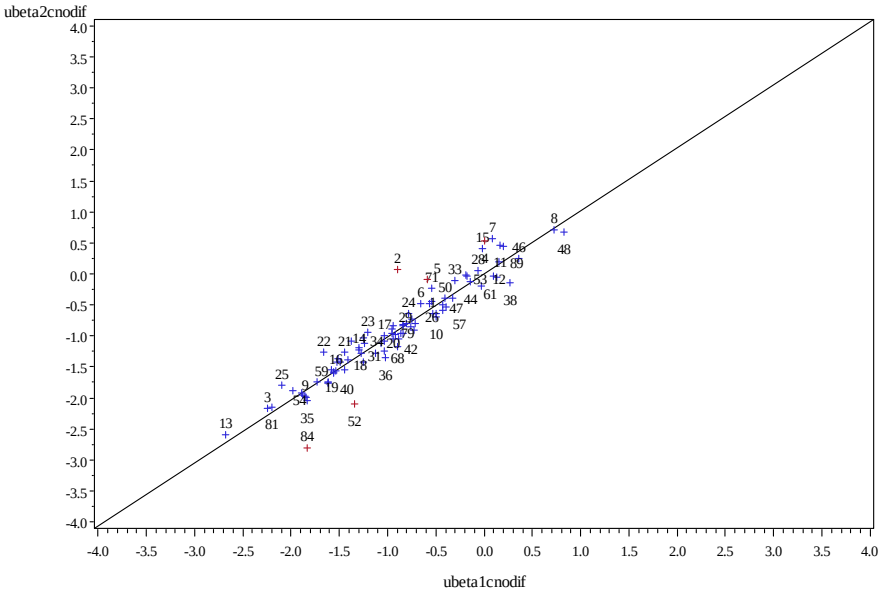
### 4.2.1 Identification des fonctionnements différentiels d'items (FDI)

L'analyse des FDI a permis de détecter 5 items : 3 items en faveur de 2007, 2 items en faveur de 2013 (figure 4). Tous ces items sont des items de physique-chimie. Ils ont été éliminés des calculs. L'évolution des programmes est susceptible de produire des FDI. Ainsi, les 3 items présentant un FDI en défaveur des élèves de 2013 sont des items de physique-chimie portant sur la combustion. Or, par le biais de changements de programmes, il se trouve que la combustion n'est plus abordée en 3e.

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

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

Figure 4 – Comparaison des paramètres de difficulté 2007-2013




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

### 4.2.3 Bilan de l'analyse des items

Au départ, il y avait :
- 103 items communs
- 72 items de 2013
- 104 items de 2007

Après suppression des items ayant un *rbis-point* inférieur à 0,2 et des items présentant un fonctionnement différentiel ou un mauvais ajustement, il reste :
- 85 items communs
- 71 items de 2013
- 85 items de 2007

## 4.3 Calcul des scores

Comme indiqué précédemment, une analyse conjointe des données (2007 et 2013) 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 fixé à 250 et leur écart-type à 50, pour l'échantillon de 2007. Le tableau 18 présente les résultats obtenus.

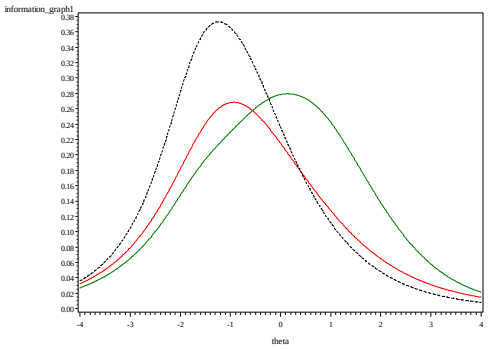
Tableau 18 – Niveaux de compétences (moyenne et écart-type)

| annee | N | Moyenne | Ecart-Type |
|---|---|---|---|
| 2007 | 7948 | 250.0 | 50.0 |
| 2013 | 8654 | 249.3 | 47.7 |

## 4.4 Courbes d'information : résultats

La figure 5 représente les courbes d'information pour les items des évaluations 2007 et 2013. La courbe rouge représente les items de 2007 non repris en 2013, la courbe verte les items nouveaux de 2013 et la courbe en pointillé les items communs aux deux évaluations. Il ressort que l'ancrage a été réalisé sur des items communs plus facile que les autres. Le choix des items d'ancrage est soumis à de nombreuses contraintes qui ne permettent pas forcément d'optimiser l'information des items retenus. Cependant, nous observons que les nouveaux items construits en 2013 ont globalement un meilleur niveau d'information, mieux centré sur le continuum de compétence (ici de moyenne 0 et d'écart-type 1).

Figure 5 – Courbe d'information du test



# 5 Construction de l'échelle

## 5.1 Méthode

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

Les scores en sciences estimés selon le modèle de réponse à l'item présenté dans la partie précédente ont été standardisés de manière à obtenir une moyenne de 250 et un écart-type de 50 pour l'année 2007. Puis, comme le montre la figure 6, la distribution des scores est « découpée » en six groupes de la manière suivante : nous déterminons le score-seuil en-deça duquel se situent 15 % des élèves (groupes 0 et 1), nous déterminons le score-seuil au-delà duquel se situent 10 % des élèves (groupe 5). Entre ces deux niveaux, l'échelle a été scindée en trois parties d'amplitudes de scores égales correspondant à trois groupes intermédiaires. Ces choix sont arbitraires et ont pour objectif de décrire plus précisément le continuum de compétence.

En effet, les modèles de réponse à l'item ont l'avantage de positionner sur la même échelle les scores des élèves et les difficultés des items. Ainsi, chaque item est associé à un des six groupes, en fonction des probabilités estimées de réussite selon les groupes. Un item est dit « maîtrisé » par un groupe dès lors que l'élève ayant le score le plus faible du groupe a au moins 50 % de chance de réussir l'item. Les élèves du groupe ont alors plus de 50 % de chance de réussir cet item.

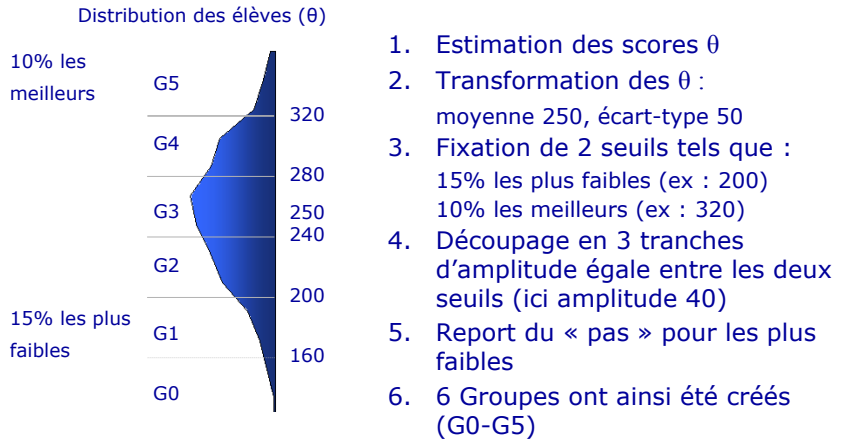
## 5.2 Caractérisation des groupes de niveaux

A partir de cette correspondance entre les items et les groupes, une description qualitative et synthétique des compétences maîtrisées par les élèves des différents groupes est proposée. Ces principaux résultats sont présentés dans une Note d'information (Bret, Garcia, & Roussel, 2014).

**Groupe $< 1$ (2,1 % des élèves)**

Le taux moyen de réussite de ces élèves est d'environ 23 $\tilde{\%}$ .
Les élèves du groupe inférieur à 1 sont capables de restituer des connaissances simples en relation avec le vécu. Ils font preuve de responsabilité face à leur santé. Ils connaissent des gestes techniques très simples.

Figure 6 – Principes de construction de l'échelle




**Groupe 1 (12,3 % des élèves)**

Le taux moyen de réussite de ces élèves est toujours inférieur à 50 % mais est presque deux fois plus élevé que le groupe inférieur à 1. Il est de 40 %. Les élèves du groupe 1 sont capables de restituer ou de mobiliser des connaissances simples, à condition qu'elles soient en relation avec leur vécu ou reprises au long du cursus du collège.
Ils savent exploiter des schémas simples (schéma électrique par exemple) et lire l'évolution d'une grandeur, d'un graphique.

**Groupe 2 (28,2 % des élèves)**

Le taux moyen de réussite de ces élèves approche les 60 %. Les élèves du groupe 2 ont des connaissances plus abstraites, à condition qu'elles les concernent sur le plan personnel (reproduction) ou bien qu'elles aient été acquises récemment. Ils peuvent extraire des informations apportées par un texte assez long, un tableau à double entrée, un graphique (avec deux courbes) et un diagramme (en bâtons ou circulaire).
C'est à partir de ce groupe que les élèves commencent à mener une démarche et à rédiger des réponses simples.
Ils établissent des relations de cause à effet, analysent des résultats simples, concrets. Ils savent conclure.
Ils sont sensibilisés aux questions liées à l'environnement.

**Groupe 3 (30,8 % des élèves)**

Le taux moyen de réussite de ces élèves est d'environ 74 %. Les élèves du groupe

3 maîtrisent le vocabulaire scientifique, dans les divers domaines rencontrés au collège. Ils exploitent, analysent un schéma (fonctionnel ou conventionnel), et le mettent en relation avec leurs connaissances. Ils croisent les informations issues de différents documents et trouvent la conclusion adéquate. Ils savent lire un graphique : détermination d'une valeur (même décimale) et d'un sens de variation, choix d'une courbe parmi plusieurs. Ils maîtrisent une chronologie négative. A ce stade, ils montrent des capacités d'abstraction : utiliser des modèles, prévoir un résultat expérimental, appliquer une relation mathématique, comprendre les conditions expérimentales à partir d'un croquis de montage.

**Groupe 4 (18,3 % des élèves)**

Le taux moyen de réussite de ces élèves atteint 85 %.
Les élèves du groupe 4 ont des connaissances très précises, même en ce qui concerne des sujets compliqués comme les transformations chimiques, la génétique.
Les élèves de ce groupe sont capables de mettre en relation un grand nombre de documents qui peuvent être complexes. Ils arrivent à traiter de nombreuses données et à les présenter de manière ordonnée et pertinente.
Ils choisissent les dispositifs expérimentaux (comportant une expérience témoin), nécessaires à la mise en évidence d'un processus. Ils sont critiques face à une expérience et sont capables de la schématiser à partir d'un modèle.
Ces élèves peuvent rédiger des réponses pour expliquer et même justifier.

**Groupe 5 (8,4 % des élèves)**

Le taux moyen de réussite de ces élèves est d'environ 94 %.
Les élèves du groupe 5 ont des connaissances très pointues. Leur raisonnement est complexe, rigoureux et pertinent.
La démarche scientifique est ancrée dans leur raisonnement : ils sont capables de reconnaître l'hypothèse testée par une manipulation donnée, ils connaissent l'utilité d'une expérience témoin.
Les élèves savent utiliser un calcul pour justifier une réponse. Le dessin scientifique est en partie maîtrisé. Ils représentent de manière ressemblante le réel, mais éprouvent encore des difficultés pour organiser leur dessin.

## 5.3 Exemples d'items

### 5.3.1 Item caractéristique du groupe <1

**Le spectacle de Kader**

Ce soir, Kader monte sur scène pour son premier concert. Le metteur en scène veut que le public voie son tee-shirt en rouge. Le chef éclairagiste illuminera la scène avec une lumière rouge. La costumière lui propose alors de choisir parmi les quatre tee-shirts dont elle dispose. Il les observe à la lumière du jour. Aucun n'est rouge. Il y en a un vert, un noir, un blanc et un bleu.

**Question 1**

Pour que le public le voie rouge, le tee-shirt que Kader doit choisir est le…

Cocher la réponse exacte.

1 ☐ blanc
2 ☐ vert
3 ☐ noir
4 ☐ bleu

C3XCP290101

Les élèves du groupe inférieur à 1 sont capables de restituer des connaissances simples en relation avec le vécu.

### 5.3.2 Item caractéristique du groupe 1

**Question 2**

La formule d'une molécule de dioxyde de carbone est …

1 ☐ $CO_2$
2 ☐ $O_2$
3 ☐ CO
4 ☐ $H_2O$

C3XCC280201

Les élèves du groupe 1 sont capables de restituer ou de mobiliser des connaissances simples, à condition qu'elles soient en relation avec leur vécu ou reprises au long du cursus du collège.

### 5.3.3 Item caractéristique du groupe 2

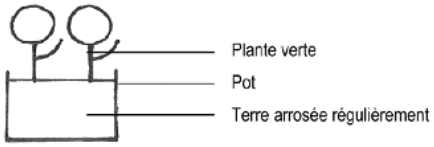
**Question 2**

Les élèves veulent tester d'autres facteurs. Ils réalisent des montages identiques avec des plantes identiques, placées dans les mêmes pots contenant la même terre. Les conditions d'arrosage et le lieu sont identiques.

Ces conditions sont représentées par le document 2.

**Document 2 : conditions d'expériences communes aux nouvelles cultures**

Le montage est placé :
- dans un endroit aéré
- à une température satisfaisante




**Document 3 : les conditions expérimentales de culture et les résultats obtenus**

| | a | b | c | d |
|---|---|---|---|---|
| |  |  |  |  |
| Sels minéraux | oui | oui | non | non |
| Lumière | oui | non | oui | non |
| Résultat : Croissance satisfaisante | oui | non | non | non |

L'étude des résultats des expériences du document 3 leur permet d'affirmer que la plante verte, en plus de l'eau, a besoin…
Cocher la réponse exacte.

1 ☐ de sels minéraux uniquement.
2 ☐ de lumière seulement.
3 ☐ de sels minéraux et de lumière.
4 ☐ de dioxyde de carbone.

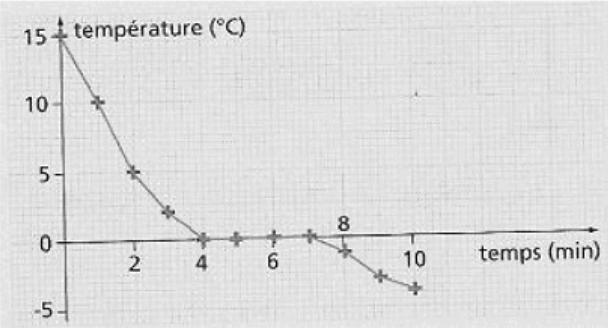
C3XDV640201

Ces élèves peuvent extraire des informations apportées par un tableau à double entrée. Ils analysent des résultats simples, concrets. Ils savent conclure.

### 5.3.4 Item caractéristique du groupe 3

**Les changements d'état de la matière**

Des élèves ont étudié un changement d'état de l'eau et ont tracé la courbe ci-contre :




**Question 1**

Quel est l'état de l'eau au début de l'expérience ?

C3XCC140101

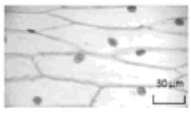
Les élèves du groupe 3 savent effectuer une lecture graphique et la mettre en relation avec leurs connaissances.

### 5.3.5 Item caractéristique du groupe 4

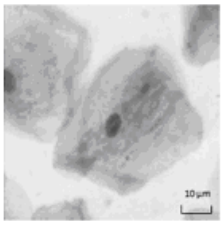
## L'Homme, l'oignon et le triton

Afin d'argumenter l'existence d'un ancêtre commun à tous les êtres vivants, on compare trois espèces à partir des données suivantes :

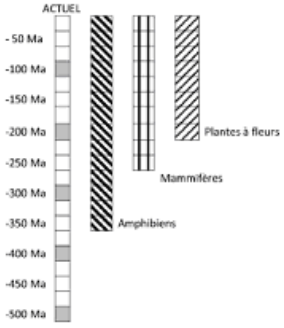
**DOCUMENT 1 : cellules de feuille d'oignon (plante à fleurs)**



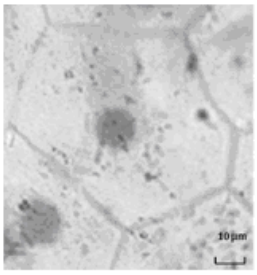

**DOCUMENT 3 : cellules de l'intérieur de la joue humaine (mammifère)**




**DOCUMENT 5 : périodes pendant lesquelles on trouve des représentants fossiles de trois groupes actuels (Ma = Millions**




**DOCUMENT 2 : cellules de la peau de triton (amphibien)**




**DOCUMENT 4 : nombre de chromosomes des cellules de diverses espèces**

| ESPECE | NOMBRE DE CHROMOSOMES |
|---|---|
| BLE | 42 |
| CHIEN | 78 |
| CHAT | 38 |
| CHEVAL | 64 |
| ESCARGOT | 24 |
| HAMSTER | 22 |
| HOMME | 46 |
| MOUCHE | 10 |
| OIGNON | 16 |
| POMME DE TERRE | 48 |
| TOMATE | 36 |
| TRITON | 24 |

**Question 1**

Construire un tableau permettant de comparer les trois espèces.

C3XEV150101

Les élèves de ce groupe sont capables de mettre en relation un grand nombre de documents qui peuvent être complexes. Ils arrivent à traiter de nombreuses données puis à les présenter de manière ordonnée et pertinente.

### 5.3.6 Item caractéristique du groupe 5

**Question 2**

Utiliser la comparaison de ces trois espèces pour argumenter en faveur de l'existence d'un ancêtre commun.

C3XDV150201

Chez les élèves de ce groupe, le raisonnement est complexe, rigoureux et pertinent.

# 6 Variables contextuelles et non cognitives

## 6.1 Variables sociodémographiques

Un certain nombre de variables sociodémographiques permettent d'enrichir l'analyse des résultats. Le score moyen des élèves est ainsi analysé en fonction du genre, du retard scolaire et quand les effectifs le permettent en fonction du secteur d'enseignement (tableau 19). Le lecteur est invité à consulter la Note d'Information pour plus de détails (Bret et al., 2014).

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

| | annee | Répartition (en %) | Score Moyen | Ecart-Type |
|---|---|---|---|---|
| Ensemble | 2007 | 100.0 | 250 | 50 |
| Ensemble | 2013 | 100.0 | 249 | **48** |
| Garçons | 2007 | 49.2 | 253 | 53 |
| Garçons | 2013 | 49.7 | 251 | 51 |
| Filles | 2007 | 50.8 | 247 | 46 |
| Filles | 2013 | 50.3 | 248 | 44 |
| Elèves en retard | 2007 | 31.3 | 225 | 41 |
| Elèves en retard | 2013 | 21.1 | 221 | 39 |
| Elèves à l'heure | 2007 | 68.7 | 261 | 50 |
| Elèves à l'heure | 2013 | 78.9 | **257** | **47** |

## 6.2 Position sociale

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 (Le Donné & Rocher, 2010).

Il n'a été possible d'établir des comparaisons qu'en termes de niveau sociale des collèges, et non au niveau individuel. En effet, en 2013, la PCS des parents est disponible pour chaque élève, mais elle ne l'était pas en 2007.

Pour chaque classe des échantillons de 2007 et 2013, la moyenne de l'indice de position socio-scolaire a été calculée et la population a ensuite été découpée en quartiles (tableau 20).

Tableau 20 – Score moyen en sciences selon l'indice de position sociale moyen de la classe en 2007 et en 2013

| Indice moyen de la classe | Année | ScoreMoyen | EcartType |
|---|---|---|---|
| 1er quartile | 2007 | 232 | 48 |
| IS : 1er quartile | 2013 | 231 | 46 |
| 2ème quartile | 2007 | 248 | 47 |
| IS : 2ème quartile | 2013 | 247 | 44 |
| 3ème quartile | 2007 | 259 | 50 |
| IS : 3ème quartile | 2013 | 252 | ***45*** |
| 4ème quartile | 2007 | 268 | 49 |
| IS : 4ème quartile | 2013 | 267 | 49 |

## 6.3 Variables conatives

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

Le tableau 21 présente en guise d'illustration les items d'une de ces dimensions, en l'occurrence le sentiment d'efficacité en sciences.

Tableau 21 – Exemple de variable conative - sentiment d'efficacité en sciences

| **Question** | **1er Axe ACP** |
|---|---|
| Je pense que j'ai un bon niveau en sciences | 0.81 |
| Je pense que je peux réussir en sciences | 0.78 |
| Je comprends bien ce que nous faisons en sciences | 0.77 |
| Les sciences, c'est trop difficile pour moi | 0.71 |

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

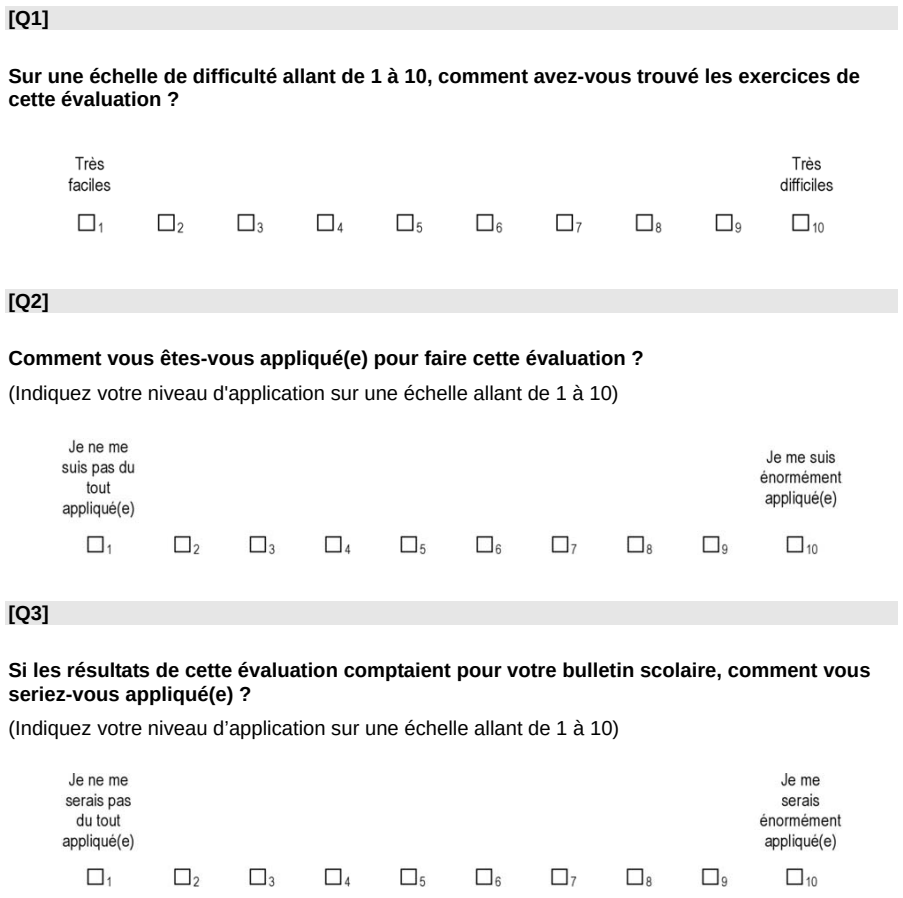
## 6.4 Motivation des élèves face à la situation d'évaluation

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. Cet instrument a été introduit dans plusieurs évaluations conduites au niveau national par la DEPP (figure 7). Les données recueillies permettent de distinguer la motivation de l'élève de la difficulté perçue du test, et ainsi de mieux appréhender le lien entre la motivation des élèves français et leur performance. L'analyse de ces données renseigne en outre sur le rôle de certaines caractéristiques, des élèves ou des évaluations elles-mêmes, dans le degré de motivation à répondre aux questions de l'évaluation (Keskpaik. & Rocher, à paraître).

Figure 7 – Instrument de mesure de la motivation au test



# 7 Annexe

## Certification AFNOR pour les évaluations CEDRE

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

### Les finalités de la certification

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

### Les enjeux pour la DEPP

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

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

### La démarche qualité

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

### Les engagements de service

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

**Les engagements de service de la DEPP**

**Des objectifs clairs et partagés**

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

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

**Des évaluations fondées sur l'expertise pédagogique**

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

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

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

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

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

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

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

**Une mesure fiable et des comparaisons temporelles pertinentes**

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

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

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

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

**Des analyses enrichies par des données de contexte**

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

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

**Transparence des méthodes et partage des résultats**

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

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

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

# Références

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

Bret, A., Garcia, E., & Roussel, L. (2014). CEDRE 2013 - sciences en fin de collège : stabilité des acquis des élèves depuis six ans. *Note d'information*, *14.28*.

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.

Garcia, E., Le Cam, M., & Rocher, T. (à paraître). Méthodes de sondage utilisées dans les programmes d'évaluation des élèves. *Education et formations*, *85-86*.

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

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

Le Donné, N., & Rocher, T. (2010). Une meilleure mesure du contexte socio-éducatif des élèves et des écoles. *Education et formations*, *Décembre*, 115.

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.

Rousseau, S., & Tardieu, F. (2004). *La macro sas cube d'échantillonnage équilibré. documentation de l'utilisateur.* Paris : INSEE.

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

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

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

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

# Liste des tableaux



# Table des figures