CEDRE

Cycle des Évaluations Disciplinaires Réalisées sur Échantillons

Rapport technique

Maîtrise de la langue 2015

École

Auteurs :
Sandra ANDREU
Etienne DALIBARD
Yann ETEVE
Saskia KESKPAIK
Marion LE CAM
Louis-Marie NINNIN
Thierry ROCHER
Ronan VOURC'H

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

Février 2018

Table des matières

Introduction

La DEPP met en place des dispositifs d'évaluation des acquis des élèves reposant sur des épreuves standardisées. Elle est également maître d'œuvre pour la France des évaluations internationales telles que PIRLS ou PISA. Ces programmes d'évaluations sont des outils d'observation des acquis des élèves pour le pilotage d'ensemble du système éducatif (Trosseille & Rocher, 2015). Les évaluations du CEDRE (Cycle d'Évaluations Disciplinaires Réalisées sur Échantillons) révèlent ainsi, en référence aux programmes scolaires, les objectifs atteints et ceux qui ne le sont pas. Ces évaluations doivent permettre d'agir au niveau national sur les programmes des disciplines, sur l'organisation des apprentissages, sur les contextes de l'enseignement, sur des populations caractérisées.

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

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

1 Cadre d'évaluation

1.1 Objectifs

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

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

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

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

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

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

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

1.2 Les compétences et connaissances visées

L'évaluation CEDRE en fin d'école en maîtrise de la langue a pour objectif de faire le point sur les connaissances et les compétences des élèves tant au niveau des savoirs que des savoir-faire. Il s'agira de mesurer l'évolution de ces connaissances et compétences entre 2003, 2009 et 2015.

Les connaissances et compétences telles qu'elles sont définies dans le socle commun de connaissances et de compétences ainsi que les programmes officiels constituent le cadre de cette évaluation.

L'élaboration des grilles de compétences pour la construction des items a pris appui sur les documents de référence que sont les programmes officiels en vigueur à partir de la rentrée scolaire 2009-2010 parus au BO spécial n° 6 du 28 août 2008 ainsi que le LPC simplifié (note de service 2012-154 du 24 septembre 2012).

Qu'est-ce que comprendre un texte ?

En dépit d'une apparence de grande simplicité, la compréhension d'un texte est une activité complexe qui met en jeu un ensemble de processus de différentes natures, dont le but ultime est la construction d'une représentation mentale cohérente.

Comprendre un discours ou un texte c'est construire une représentation mentale intégrée et cohérente de la situation décrite par ce discours ou ce texte. Les traitements mis en oeuvre pour comprendre un texte concernent donc à la fois les éléments linguistiques et les concepts et relations que ceux-ci évoquent. L'activité de compréhension se déploie en relation avec un texte et en fonction de l'objectif poursuivi par le lecteur (se distraire, rechercher une information, comprendre comment monter une maquette ou réaliser une recette, etc). Il faut donc concevoir **l'activité de compréhension comme aboutissant toujours**

à une interprétation du texte présentant une marge de liberté plus ou moins large en fonction du texte, des standards que se fixe le lecteur (ou qui lui sont imposés) et des connaissances préalables dont il dispose.

La compréhension est un phénomène dynamique qui nécessite de faire des inférences. Dans tout énoncé, il y a ce qui est littéral (dit explicitement dans le texte) et ce qui est de l'ordre de l'interprétation. **C'est un phénomène dynamique** qui n'est pas instantané mais qui se déroule dans le temps et au cours duquel le lecteur va construire une représentation de la situation décrite par le texte. **Il y a toujours un double traitement de l'information** (traitement du code et du contenu). La construction de la représentation passe par l'interprétation qui s'effectue toujours à partir de l'objectif du lecteur, de ses connaissances linguistiques et notionnelles, de ses capacités cognitives ... **La compréhension nécessite une capacité de mémoire, de mise en relation et de synthèse.**

La construction de la compréhension de tout texte exige au fur et à mesure de la lecture :
- Un apport d'informations nouvelles.
- Une mise en relation des informations nouvelles avec celles qui sont déjà disponibles (elles sont cumulables avec plus ou moins de risque).
- Une intégration de ces informations au fur et mesure de leur introduction.

L'interprétation se fait sans difficulté si on a des connaissances sur le contexte. Pour établir continuité et cohérence, il est nécessaire de faire des inférences, c'est-à-dire de compléter ce qui est explicité par ce que l'on sait déjà. **Les inférences sont indispensables à la compréhension.**

La compréhension passe par le traitement des anaphores (reprise de systèmes linguistiques pour reprendre une information). Les anaphores permettent d'assurer la continuité des personnages et des faits, cela permet dans un texte de comprendre qui fait quoi.

La compréhension passe par le **traitement des connecteurs**. Les connecteurs mettent en relation des informations facilitant ainsi l'interprétation.

Comment est définie la compétence en lecture dans les évaluations internationales ?

Dans PIRLS celle-ci est définie comme « l'aptitude à comprendre et à utiliser les formes du langage écrit que requiert la société ou qui sont importantes pour l'individu. Les jeunes lecteurs peuvent construire du sens à partir de textes très variés. Ils lisent pour apprendre, pour s'intégrer dans la société où la lecture joue un rôle essentiel et pour leur plaisir. »

En ce sens, deux objectifs principaux sont assignés à la lecture :

– Lire pour accéder aux textes littéraires.
– Lire pour acquérir et utiliser des informations.

Dans ce contexte quatre niveaux de compétences sont définis :

– **Prélever des informations explicites (Prélever) :** repérer les informations directement liées à l'objectif de la lecture ; chercher des idées précises ; chercher la définition de mots ou d'expressions ; repérer le contexte de l'histoire (époque, lieu) ; trouver l'idée principale (si elle est exprimée explicitement).
– **Faire des inférences directes (Inférer) :** déduire que tel événement a entraîné tel autre ; déduire l'élément principal d'une série d'arguments ; déterminer le référent d'un pronom ; repérer les généralisations présentées dans le texte ; décrire la relation entre deux personnages.
– **Interpréter et assimiler idées et informations (Interpréter) :** déduire le message global ou le thème d'un texte ; envisager une alternative aux actions des personnages ; comparer des informations du texte ; saisir l'atmosphère ou le ton du récit ; trouver une application concrète aux informations contenues dans le texte.
– **Examiner et évaluer le contenu, la langue et les éléments textuels (Apprécier) :** évaluer la probabilité que les événements décrits se passent réellement ; décrire la manière dont l'auteur a amené la chute ; juger de l'intégralité ou de la clarté des informations fournies dans le texte ; décrire comment le choix des adjectifs modifie le sens.

Tableau 2 – Définition des compétences (évaluation 2013)

	Éléments des programmes 2008	Éléments du LPC simplifié
Langage oral	L'élève est capable : - d'écouter le maître, - de poser des questions, - d'exprimer son point de vue, ses sentiments. Il s'entraîne à prendre la parole devant d'autres élèves pour : - reformuler, - résumer, - raconter, - décrire, - expliciter un raisonnement, - présenter des arguments. Dans des situations d'échanges variées, il apprend à tenir compte des points de vue des autres, à utiliser un vocabulaire précis appartenant au niveau,de la langue courante, à adapter ses propos en fonction de ses interlocuteurs et de ses objectifs.	Dire - s'exprimer à l'oral comme à l'écrit dans un vocabulaire approprié et précis ; - prendre la parole en respectant le niveau de langue adapté ; - répondre à une question par une phrase complète à l'oral comme à l'écrit ; - prendre part à un dialogue : prendre la parole devant les autres, écouter autrui, formuler et justifier son point de vue.
Lecture	- automatisation de la reconnaissance des mots, lecture aisée de mots irréguliers et rares, augmentation de la rapidité et de l'efficacité de la lecture silencieuse ; - compréhension des phrases ; - compréhension de textes scolaires (énoncés de problèmes, consignes, leçons et exercices des manuels) ; - compréhension de textes informatifs et documentaires ; - compréhension de textes littéraires (récits, descriptions, dialogues, poèmes). L'élève apprend à comprendre le sens d'un texte en en reformulant l'essentiel et en répondant à des questions le concernant.	Lire avec aisance un texte.

Écrire	Rédaction Les élèves apprennent : - à narrer des faits réels, - à décrire, - à expliquer une démarche, - à justifier une réponse, - à inventer des histoires, - à résumer des récits, - à écrire un poème, ... en respectant des consignes de composition et de rédaction.	Répondre à une question par une phrase complète à l'écrit. Rédiger un texte d'une quinzaine de lignes (récit, description, dialogue, texte poétique, compte-rendu) en utilisant ses connaissances en vocabulaire et en grammaire.

Tableau 3 – Niveaux de compétences évaluées (grille 2003-2009)

Intitulé	Définition	Tâche	Structure
Prélever une information Niveau 1	Montrer par son choix qu'on a prélevé une information qui existe à l'identique dans le texte.	Repérer dans le texte et choisir parmi 4 ou 5 propositions, l'information explicite demandée.	Un questionnement direct. La consigne étant rédigée à partir de la reprise d'une phrase du texte.
Prélever une information Niveau 2	Montrer par son choix qu'on a construit une nouvelle information à partir des éléments du texte, qu'il y a eu inférence.	Comprendre les informations du texte permettant de construire une réponse et de choisir parmi 4 ou 5 propositions, l'information implicite demandée.	Un questionnement direct.
Déduire une information	Montrer par son choix qu'on a construit une nouvelle information à partir des éléments du texte, qu'il y a eu inférence de type déductif.	Déduire une information à partir de 2 propositions tenues pour vraies.	« C » est vrai si et seulement si « A et B » sont vrais. On propose aux élèves : « A », « B », « A et B », « non A et non B ».
Analyser un document	Tirer les éléments essentiels qui aident à la compréhension de cette situation.	Montrer par son choix qu'on a sélectionné les informations permettant une compréhension fine du texte.	5 ou 6 affirmations tirées du texte sont proposées. 2 ou 3, seulement, aident à la compréhension du texte.
Synthétiser un document Niveau 1	Trouver la thématique.	Montrer par son choix qu'on a repéré de quoi parle le texte.	Trouver un titre, un thème.
Synthétiser un document Niveau 2	Trouver un résumé.	Montrer par son choix qu'on a repéré la structure et l'enchaînement du propos, le but du texte.	choix parmi quatre résumés
Lexique		Poser une question sur un champ lexical du texte.	Chasser l'intrus parmi une liste de 5 termes.
Outils de la langue	la maîtrise des connecteurs, de la reprise anaphorique, de l'organisation chronologique du propos, de l'accord dans le groupe nominal et/ou le groupe verbal, de la concordance des temps et des homophones.		

Tableau 4 – Niveaux de compétences évaluées (grille 2013)

Intitulé 2003-2009	**Définition**	**Intitulé 2013**	**Description**	**Modèle théorique de référence**	**GEODE (codes)**
Prélever une information Niveau 1	Montrer par son choix qu'on a prélevé une information qui existe à l'identique dans le texte.	Prélever une information explicite (Retrouver/ localiser une information)	Questions littérales : La réponse est prélevée directement dans le texte. Elle peut nécessiter la compréhension de paraphrases	Construction modèle de situation : Localiser/ retrouver	PR
Prélever une information Niveau 2	Montrer par son choix qu'on a construit une nouvelle information à partir des éléments du texte, qu'il y a eu inférence.	Inférence de niveau 1 (Inférer/ généraliser)	Cohérence locale (ou base de texte ou encore inférences textuelles) : regroupe les questions nécessitant la compréhension d'une causalité implicite, d'une anaphore, d'un cadre spatio-temporel implicite, l'inférence du sens d'un mot non connu.	Construction modèle de situation : Inférer/ généraliser	IN
Déduire une information	Montrer par son choix qu'on a construit une nouvelle information à partir des éléments du texte, qu'il y a eu inférence de type déductif.	Inférence de niveau 2 (Inférer/ généraliser)	Cohérence globale (ou modèle de situation) : questions nécessitant une inférence de connaissances (question sur les sentiments des personnages) ou la compréhension du thème et/ou des idées principales.	Construction modèle de situation : Inférer/ généraliser	IF
Synthétiser un document	Trouver la thématique Choisir un résumé	Synthétiser (Organiser/ résumer) (Intégrer/ synthétiser)	Organiser les idées principales à l'aide d'organisateurs graphiques, choisir un résumé, ordonner/souligner les idées principales...	Construction modèle de situation : Organiser/ résumer	SY

Tableau 5 – Niveaux de compétences évaluées (grille 2013, métacognition)

Intitulé 2003-2009	**Définition**	**Intitulé 2014**	**Description**	**Modèle théorique de référence**	**GEODE (codes)**
Métacognition					
Analyser un document.	Tirer les éléments essentiels qui aident à la compréhension de cette situation.	Métacognition Expliquer/ raisonner.	Guider/contrôler sa compréhension, faire des liens entre ses propres connaissances et le texte, résoudre les incohérences (ou difficultés), expliquer, raisonner ; comprendre les motivations, objectifs, relations logiques. Choisir /donner une justification. Pourquoi ? Comment est-ce arrivé ? Quelle est la cause de X ? comparer X et Y ... Expliquer une réponse (ou son choix).	Habiletés de compréhension appliquées. Expliquer/ raisonner.	MC
		Métacognition Évaluer/ critiquer.	Dans le cadre d'un texte unique : repérer si l'information lue permet ou non de répondre à certaines questions (apporte l'information pertinente/à un pb). Savoir repérer des raisonnements erronés, contradictoires, l'objectif de l'auteur, son point de vue.	Habiletés de compréhension appliquées. Évaluer/ critiquer.	

1.3 Construction du test

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

1.3.1 Elaboration des items

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

Un équilibre de proportion entre les items considérés comme étant de difficulté « facile », « moyenne » ou « difficile » est recherché. Pour les quatre domaines évalués, la compréhension de l'écrit, la compréhension de l'oral, l'étude de la langue et la production d'écrit, certains items sont identiques à ceux proposés en 2009 et 2003 afin d'assurer une comparabilité de qualité.

Deux types de formats de questions sont utilisés : les questions fermées (QCM, QCM-images, série, série-images) et les questions ouvertes appelant une réponse écrite (réponse courte - un chiffre, un nombre - ou réponse longue - production en autonomie de l'élève). Un entraînement est prévu au début de chaque cahier afin de familiariser les élèves avec le type de question rencontré.

Les réponses des formats QCM, QCM-images sont saisies de manière automatisée à la fin de la passation. Les réponses des formats série et série-images sont saisies de manière automatisée et donnent lieu à un regroupement ultérieur de leurs propositions. Dans le cas de ces séries, des seuils statistiques ont été établis pour valider les réponses des élèves. Les réponses des formats « réponse libre de l'élève » sont corrigées par des experts via une interface Internet. Ce dispositif de correction à distance s'appuie sur le logiciel AGATE (cf. partie « Analyse des items »).

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

Objectifs

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

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

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

Principes fonctionnels

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

1.3.2 Constitution des cahiers

L'évaluation CEDRE Maîtrise de la langue 2015 se présente sous la forme de QCM et de questions ouvertes. Elle est composée d'items repris à l'identique par rapport à 2003 et 2009 :

– 61 items passés en 2009 et 2015 ;
– 31 items passés en 2003, 2009 et 2015.

Ces items dits « d'ancrage » représentent près de 43% de l'évaluation. Elle est également composée de 122 items nouveaux qui ont fait l'objet d'une expérimentation en 2014.

Dans le cadre d'une évaluation sur support papier, le test se compose d'un ensemble de cahiers, constitués de blocs, qui sont eux-mêmes composés d'unités (ensemble d'items). Pour cette opération, 22 blocs sont constitués et répartis au sein de 20 cahiers :

- 13 cahiers constitués de blocs de compréhension de l'écrit et d'étude de la langue ;
- 6 cahiers constitués de blocs de compréhension orale, compréhension écrite, étude de langue et production d'écrit ;
- 1 cahier constitué de blocs de compréhension de l'écrit, d'une dictée et de deux situations de compréhension de l'écrit proposées en 1987, 1997 et 2007 et issues de l'opération Lire Ecrire Compter.

Tableau 6 – Exemple de répartition des blocs dans les cahiers

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

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 telles que chaque bloc devant se retrouver un même nombre de fois au total et chaque association de blocs doit figurer au moins une fois dans un cahier. Ce dispositif, couramment utilisé dans les évaluations bilans, notamment les évaluations internationales, permet d'estimer la probabilité de réussite de chaque élève à chaque item sans que chaque élève ait passé l'ensemble des items.

Au final, pour l'évaluation CEDRE 2015, chaque cahier comprend quatre séquences cognitives de 30 minutes chacune. Elles sont complétées par une cinquième séquence de 30 minutes (questionnaire de contexte), identique dans tous les cahiers (cf. partie 6 du rapport).

1.4 Passation des évaluations

La passation de l'évaluation finale a eu lieu en mai 2015. Comme en 2009, cette évaluation a été précédée d'une expérimentation l'année $n-1$ de façon à tester un grand nombre d'items auprès d'un échantillon réduit d'établissements.

Dans chaque école, le directeur ou la directrice a été désigné comme étant l'administrateur du test, son rôle étant de veiller au strict respect de la procédure à suivre pour que l'évaluation soit passée dans les mêmes conditions quel que soit l'établissement.

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

Les professeurs des écoles des classes concernées ont également dû renseigner un questionnaire de contexte. L'anonymat des élèves et des personnels a été respecté, chaque cahier étant repéré par un numéro.

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

2 Sondage

2.1 Méthodes

2.1.1 Sondage par grappes stratifié

Dans le premier degré, nous ne disposons pas des informations auxiliaires présentes dans les bases de sondage de la DEPP, telle que la PCS des parents par exemple. Il n'est donc pas possible de réaliser un tirage équilibré comme c'est le cas pour les évaluations CEDRE en 3e.

Le tirage consiste donc simplement en un sondage par grappes stratifié. La stratification porte généralement sur la zone de scolarisation et tous les élèves de CM2 des écoles sélectionnés participent. Le choix de sondages par grappes est motivé par la facilité de gestion. En effet, le fait de sélectionner tous les élèves d'une école permet d'éviter de mettre en place des procédures de tirage au sort d'élèves une fois les écoles tirées.

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

2.1.2 Redressement de la non réponse : calage sur marges

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

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

lage :

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

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

Tirage après élimination de la base des échantillons précédemment tirés

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

Notations

Un échantillon S_1 a été tiré. Il est connu et les probabilités d'inclusion des établissements π_j^1 sont également connues. On souhaite alors tirer un échantillon S_2 dans la population U avec les probabilités π_j^2, mais sans aucun recouvrement avec l'échantillon S_1. On va donc tirer l'échantillon S_2 dans la population $U(S_1)$, c'est-à-dire la population U privée des établissements de l'échantillon S_1 qui appartiennent à U. Notons d'emblée que S_1 n'a pas nécessairement été tiré dans U, mais potentiellement dans une autre population, plus large ou plus réduite ; cela n'affecte en rien la formulation envisagée ici. Notons également que l'indice j est utilisé ici : il concerne les établissements et non les élèves, représentés par l'indice i.

Il s'agit donc de procéder à un tirage conditionnel. On note π_j^{2/S_1} les probabilités d'inclusion conditionnelles des établissements dans le second échantillon S_2, sachant que le premier échantillon est connu. Ces probabilités

conditionnelles peuvent s'écrire :

$$\pi_j^{2/S_1} = \begin{cases} \lambda_j & \text{si } j \notin S_1 \\ 0 & \text{si } j \in S_1 \end{cases} \text{, avec } \lambda_j \in [0,1]$$

On a $\pi_j^2 = E(\pi_j^{2/S_1}) = \lambda_j(1-\pi_j^1)$ d'où $\lambda_j = \frac{\pi_j^2}{1-\pi_j^1}$

Condition fondamentale

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

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

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

2.1.3 Calcul de précision : méthode

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

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

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

Cependant, lorsqu'un calage sur marges a été effectué, il faut en tenir compte

pour le calcul de la précision. Dans ce cas, la variance de $\hat{\bar{Y}}$ est asymptotiquement équivalente à la variance des résidus de la régression de la variable d'intérêt sur les variables de calage.

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

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

2.2 Echantillonnage

Champ

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

Sont donc exclus du champ :
– Les TOM.
– Les écoles hors contrat.
– Les écoles à l'étranger.
– Les écoles spécialisées.
– Les écoles de moins de 6 élèves de CM2.
– Les DOM.
La base de sondage est relativement pauvre en informations dans le premier degré. Nous disposons cependant d'informations sur les établissements scolaires, comme le secteur d'enseignement.

Stratification

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

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

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

Modalités de sélection

On effectue un tirage d'écoles, stratié selon trois strates (variable strate2). Ensuite, tous les élèves de CM2 des écoles sélectionnées sont interrogés. C'est un sondage par grappes. Les écoles des échantillons de TIMSS CM1 et de l'expérimentation PIRLS CM1 ont été préalablement retirées de la base de tirage pour qu'il n'y ait pas de recouvrement. L'échantillon se compose de trois sous-échantillons indépendants. Le premier sous-échantillon est retiré de la base de sondage pour le tirage du deuxième sous- échantillon, lui-même retiré de la base de sondage pour le tirage du troisième. Les trois sous-échantillons sont tirés avec la macro CUBE.

Echantillon 2015

L'échantillon se compose de trois sous-échantillons indépendants. Le premier sous-échantillon est retiré de la base de sondage pour le tirage du deuxième sous- échantillon, lui-même retiré de la base de sondage pour le tirage du troisième.

Pour le 1er échantillon (13 cahiers tournants avec 13 blocs), on vise 8 000 élèves avec une surreprésentation de l'éducation prioritaire : 3 000 élèves dans chacune des strates 1 et 2, et 2 000 élèves dans la strate 3, ce qui correspond à 120 écoles dans la strate 1, 93 écoles dans la strate 2 et 73 écoles dans la strate 3. Pour le 2ème échantillon (6 cahiers de production d'écrit et compréhension orale ainsi que 2 blocs parmi les 13 du 1er échantillon) on vise 2 400 élèves avec allocation proportionnelle, soit 69 écoles dans la strate 1, 10 dans la strate 2 et 14 dans la strate 3. Pour le 3ème échantillon (1 cahier avec 4 textes repris de 1987-2007 et une dictée ainsi que 2 blocs parmi les 13 du 1er échantillon) on vise 3 600 élèves avec allocation proportionnelle, soit 104 écoles dans la strate 1, 15 écoles dans la strate 2 et 20 écoles dans la strate 3).

Base de sondage

Le tableau 7 présente les exclusions dans la population ciblée.

Le tableau 8 présente la répartition de la population ciblée selon le secteur d'enseignement.

Tableau 7 – Exclusions pour la base de sondage - CEDRE Maîtrise de la langue CM2 2015

	Écoles	**Élèves**	**Écoles exclues**	**Élèves exclus**
Ecoles accueillant des Élèves de CM2	32 822	812 262		
On retire les TOM	32 483	803 666	339	8 596
On retire les Écoles hors contrat	32 204	801 244	279	2 422
On retire les Écoles spécialisées	32 185	800 725	19	519
On retire les petites Écoles (<6 CM2)	29 677	791 702	2 508	9 023
On retire les DOM	28 783	754 230	894	37 472
Total exclusions CEDRE			**4 039**	**58 032**
Base de sondage CEDRE CM2	**28 783**	**754 230**		
On retire PIRLS FT et TIMSS4			195	7 496
Base de tirage CEDRE MDL-1 CM2	**28 588**	**746 734**		
On retire Cedre MDL-1			286	8 010
Base de tirage CEDRE MDL-2 CM2	**28 302**	**738 724**		
On retire Cedre MDL-2			93	2 428
Base de tirage CEDRE MDL-3 CM2	**28 209**	**736 296**		

Tableau 8 – Répartition base de sondage - CEDRE Maîtrise de la langue CM2 2015

	Nb d'écoles	Pct d'écoles	Nb d'élèves	Pct d'élèves
1. Public hors EP	21 592	75,0%	541 779	71,8%
2. EP	3 030	10,5%	98 431	13,1%
3. Privé	4 161	14,5%	114 020	15,1%
Total	**28 783**	**100,0%**	**754 230**	**100,0%**

Échantillon

Le tableau 9 présente la répartition de l'échantillon selon le secteur d'enseignement. Au total, 518 écoles ont été sélectionnées.

Tableau 9 – Répartition dans l'échantillon - CEDRE Maîtrise de la langue CM2 2015

	MDL1		MDL2		MDL3		**Total**	
	Écoles	Élèves	Écoles	Élèves	Écoles	Élèves	**Écoles**	**Élèves**
1. Public hors EP	120	3 018	69	1 721	104	2 592	**293**	**7 331**
2. EP	93	3 017	10	333	15	506	**118**	**3 856**
3. Privé	73	1 975	14	374	20	577	**107**	**2 926**
Total	**286**	**8 010**	**93**	**2 428**	**139**	**3 675**	**518**	**14 113**

2.3 Etat des lieux de la non-réponse

2.3.1 Non-réponse totale

Parmi la non-réponse totale, nous distinguons selon la non-réponse d'écoles entières ou la non-réponse d'élèves dans les écoles participantes. Les chiffres suivants ont été observés pour 2015. Tout d'abord, 95,5 % des écoles de l'échantillon ont répondu à l'évaluation (tableau 10).

Au final, 92,7 % des effectifs attendus ont participé (tableau 11).

Tableau 10 – Non-réponse des écoles - CEDRE Maîtrise de la langue CM2 2015

strate	Nb écoles attendues	Nb écoles répondantes	% de écoles répondantes
1- public hors EP	120	115	95,8%
2- EP	93	89	95,7%
3- privé	73	69	94,5%
Total	**286**	**273**	**95,5%**

Tableau 11 – Non-réponse globale - CEDRE Maîtrise de la langue CM2 2015

strate	Nb élèves attendus	Nb élèves répondants	% élèves répondants
1- public hors EP	3 018	2 856	94,6%
2- EP	3 017	2 762	91,5%
3- privé	1 975	1 810	91,6%
Total	**8 010**	**7 428**	**92,7%**

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").

Pour CEDRE MDL 2015, les cahiers élèves sont composés de quatre 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 4,9 pour la 1ère séquence, 5,5 pour la 2ème séquence, 3,2 pour la troisième séquence et 5,5 pour la qutrième séquence.

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

Au final, on considère que :
- 265 élèves n'ont pas vu la séquence 1 dont :
  - 211 n'ont répondu à aucun item de la séquence
  - 54 ont répondu à moins de 50 % de la séquence
- 241 élèves n'ont pas vu la séquence 2 dont :
  - 206 n'ont répondu à aucun item de la séquence
  - 35 ont répondu à moins de 50 % de la séquence
- 305 élèves n'ont pas vu la séquence 3 dont :
  - 258 n'ont répondu à aucun item de la séquence
  - 47 ont répondu à moins de 50 % de la séquence
- 370 élèves n'ont pas vu la séquence 4 dont :
  - 338 n'ont répondu à aucun item de la séquence
  - 32 ont répondu à moins de 50 % de la séquence

Les élèves dont les quatre séquences sont codées en « s » sont considérés comme de la non réponse totale. C'est le cas pour 32 élèves.

2.4 Redressement

Pour tenir compte de la non réponse, l'échantillon a été redressé à l'aide d'un calage sur marge. Préalablement au calage, on effectue tout d'abord une post-stratification. Puis, deux variables de calage sont utilisées :
– la répartition selon le sexe dans la population ;
– la répartition selon le retard scolaire.
Le tableau 12 montre que l'ampleur du calage est très réduite.

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

	Modalité ou variable	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population
Retard	1	84 403.05	78 800	11.19	10.45
	2	669 826.87	675 430	88.81	89.55
Sexe	1	378 699.81	384 869	50.21	51.03
	2	375 530.12	369 361	49.79	48.97
Strate	1	541 778.98	541 779	71.83	71.83
	2	98 431.00	98 431	13.05	13.05
	3	114 019.95	114 020	15.12	15.12

2.5 Précision

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

Tableau 13 – Scores moyens et erreurs standard associées - CEDRE Maîtrise de la langue CM2

Année	Score moyen	Erreur standard
2003	250	1.37
2009	251.0	1.45
2015	250.7	1.20

Pour savoir par exemple si l'évolution entre 2009 et 2015 est significative , il faut calculer la valeur suivante :

$$\frac{\left|\hat{\bar{Y}}_{2015} - \hat{\bar{Y}}_{2009}\right|}{\sqrt{se^2_{\hat{\bar{Y}}_{2015}} + se^2_{\hat{\bar{Y}}_{2009}}}} \tag{2}$$

Avec une valeur de 0,16 (inférieure à 1,96), cela signifie que la baisse du score moyen observée entre 2009 et 2015 n'est pas statistiquement significative, tout comme les évolutions de scores observées entre 2003 et 2009 (0,50), et entre 2003 et 2015 (0,38).

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 - CEDRE Maîtrise de la langue CM2

Année	Groupe < 1	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5
2003	2.7	12.3	26.3	29.4	19.3	10.0
2009	2.6	11.3	25.6	31.7	19.5	9.4
2015	1.0	10.0	29.1	33.1	19.7	7.2

Tableau 15 – Erreurs standards des répartitions en % dans les groupes de niveaux - CEDRE Maîtrise de la langue CM2

Année	Groupe < 1	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5
2003	0.3	0.7	0.7	0.6	0.6	0.6
2009	0.3	0.7	0.7	0.8	0.9	0.6
2015	0.1	0.6	0.9	0.7	0.7	0.6

Design effect

L'effet du plan de sondage (*Design Effect*) permet de rapporter l'erreur de mesure faite par un tirage spécifique à l'erreur de mesure qui aurait été faite en procédant à un sondage aléatoire simple (SAS) du même nombre d'élèves. Pour la moyenne d'une variable Y et un plan de sondage complexe P :

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

Dans le cas d'un sondage en grappes, la précision est dégradée en comparaison d'un sondage aléatoire simple (tableau 16). Cela signifie qu'en 2015, un sondage aléatoire simple avec un effectif 2,4 fois moins important aurait conduit au même niveau de précision.

Tableau 16 – Effet du plan de sondage - CEDRE Maîtrise de la langue CM2 2015

Année	Erreur Standard	Erreur SAS	*Design Effect*
2003	1.37	0.64	2.15
2009	1.45	0.69	2.11
2015	1.20	0.50	2.41

3 Analyse des items

3.1 Méthodologie

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

3.1.1 Approche classique

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

Réussite et score

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

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

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

$$p_j = \frac{1}{n} \sum_{i=1}^{n} Y_i^j \tag{5}$$

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

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

$$S_i = \sum_{j=1}^{J} Y_i^j \tag{6}$$

La théorie classique des tests a précisément pour objet d'étude le score S_i obtenu par un élève à un test. Elle postule notamment que ce score observé résulte de la somme d'un score « vrai » inobservé et d'une erreur de mesure. Un certain

nombre d'hypothèses portent alors sur le terme d'erreur (pour plus d'informations, cf. par exemple Laveault et Gregoire, 2002).

Fidélité

Dans le cadre de la théorie classique des tests, la fidélité (*reliability*) est définie comme la corrélation entre le score observé et le score vrai : le test est fidèle, lorsque l'erreur de mesure est réduite. Une manière d'estimer cette erreur de mesure consiste par exemple à calculer les corrélations entre les différents sous-scores possibles : plus ces corrélations sont élevées, plus le test est dit fidèle [1].

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

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

où $\bar{r}$ est la moyenne des corrélations inter-items.

De ce point de vue, cet indicateur renseigne sur la consistance interne du test. En pratique, une valeur supérieure à 0,8 témoigne d'une bonne fidélité [2].

Indices de discrimination

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

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

1. Notons au passage que la naissance des analyses factorielles est en lien avec ce sujet : Charles Spearman cherchait précisément à dégager un facteur général à partir de l'analyse des corrélations entre des scores obtenus à différents tests.

2. La littérature indique plutôt un seuil de 0,70 (Peterson, 1994). Cependant, comme le montre la formule ci-dessus, le coefficient α est lié au nombre d'items, qui est important dans les évaluations conduites par la DEPP afin de couvrir les nombreux éléments des programmes scolaires. Des facteurs de correction existent néanmoins et permettent de comparer des tests de longueur différentes.

En effet, on peut montrer que

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

où $\bar{S}_{(j1)}$ est le score moyen sur l'ensemble de l'évaluation des élèves ayant réussi l'item j, $\bar{S}_{(j0)}$ celui des élèves l'ayant échoué et σ_S est l'écart-type des scores.

C'est donc bien un indice de discrimination, entre les élèves qui réussissent et ceux qui échouent à l'item. En pratique, on préfère s'appuyer sur les $r_{bis-point}$ corrigés, c'est à dire calculés par rapport au score à l'évaluation privée de l'item considéré. Une valeur inférieure à 0,2 indique un item peu discriminant (Laveault et Grégoire, 2002).

3.1.2 Analyse factorielle des items

L'analyse factorielle permet d'étudier la structure des données et, plus particulièrement, la structure des corrélations entre les variables observées (ou manifestes) [3]. Il s'agit d'identifier les différentes dimensions sous-jacentes aux réussites observées et surtout d'évaluer le poids de la dimension principale, dans la mesure où c'est une optique unidimensionnelle qui sera envisagée lors de la modélisation.

Dans le cas où les items sont dichotomiques, la matrice des corrélations entre items est en fait la matrice des coefficients ϕ, qui sont bornés selon les taux de réussite aux items (Rocher, 1999). Une analyse factorielle basée sur cette matrice peut donc montrer quelques faiblesses : des facteurs « artefactuels » sont susceptibles d'apparaître, en lien avec le niveau de difficulté des items et non avec les dimensions auxquelles ils se rapportent. De plus, d'un point de vue théorique, certaines hypothèses utiles pour l'estimation, comme la normalité des variables, ne sont pas envisageables.

L'optique retenue est alors de se ramener à un modèle linéaire : les variables observées catégorielles sont considérées comme la manifestation de variables latentes continues.

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

Les réponses à un item dichotomique sont définies de la manière suivante :

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

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

L'analyse factorielle des items consiste donc en une analyse factorielle linéaire sur les variables continues Z_j. Deux modèles sont donc considérés. D'une part, une variable latente continue et conditionnant la réponse à l'item est fonction linéaire de facteurs communs et d'un facteur spécifique. D'autre part, un modèle de seuil représente la relation non linéaire entre la variable latente et la réponse à l'item. Ce procédé permet de se ramener à une analyse factorielle linéaire, à la différence que les variables Z_j ne sont pas connues. Il s'agit donc d'estimer la matrice de corrélation de ces variables, sous certaines hypothèses.

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

3.2 Codage des réponses aux items

3.2.1 Valeurs manquantes

Trois types de valeurs manquantes sont distinguées :
- Valeurs manquantes structurelles : l'élève n'a pas vu l'item. C'est le cas pour les cahiers tournants, où les élèves ne voient pas tous les items. Dans ce cas, on considère l'item comme *non administré*, l'absence de réponse n'est alors pas considérée comme une erreur.
- Absence de réponse : l'élève a vu l'item mais n'y a pas répondu. L'absence de réponse est alors considérée comme une erreur de la part de l'élève.

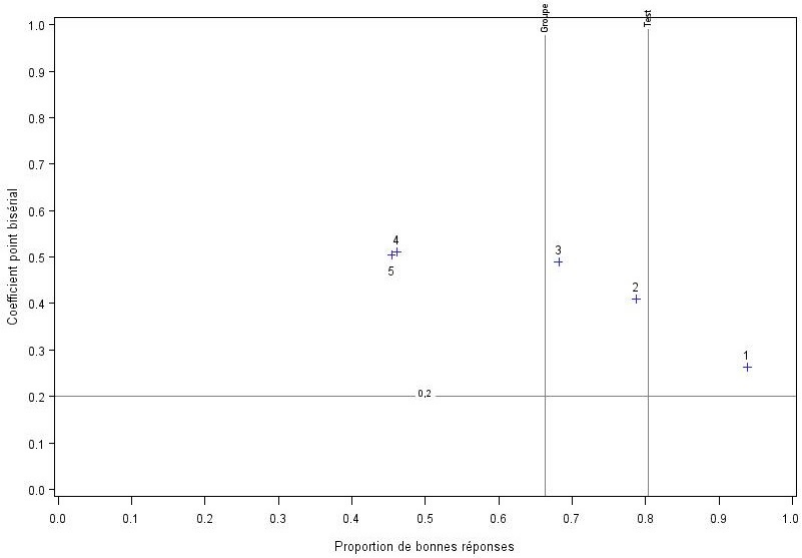
– Non-réponse terminale : l'élève s'est arrêté au cours de l'épreuve, potentiellement en raison d'un manque de temps. Des choix sont effectués pour déterminer le traitement de ces valeurs. Nous considérons que si un élève a passé moins de 50 % d'une séquence, il n'a pas vu la séquence, les valeurs manquantes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

3.2.2 Regroupement des items

Les séries d'items comportant seulement deux réponses, comme les Vrai/Faux, font l'objet d'un traitement spécifique. Les items de ce type sont regroupés pour former un seul item à réponse binaire (réussite ou échec). En effet, la plus forte potentialité de réponse au hasard et l'inter-dépendance des items fragilisent leur utilisation individuelle.

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

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



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

3.2.3 Traitement des données et correction des questions ouvertes

Tous les cahiers recueillis dans le cadre de cette opération ont été scannés par une société extérieure. Les réponses aux questions à choix multiples 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 précise et d'autre part à corriger collectivement à blanc un nombre défini d'imagettes pour s'assurer de la compréhension et de la bonne mise en oeuvre des consignes. Puis, pendant la télécorrection, il arbitre les litiges lors des doubles-corrections. Le correcteur corrige les items en portant un codage de réussite/erreur sur chaque item. En cas de doute, il peut se référer à son superviseur de groupe. Une messagerie interne complète le dispositif et permet un échange de point de vue entre les différents acteurs.

3.3 Résultats

3.3.1 Pouvoir discriminant des items

Le calcul les indices de discrimination nous a amené à éliminer 10 items dont les indices *rbis-point* étaient trop faibles : 4 items de 2003, 4 items de 2015, et 2 items communs à 2009 et 2015.

3.3.2 Dimensionnalité

Le tableau 17 présente les résultats de l'analyse factorielle des items effectuée sur l'année 2015.

Tableau 17 – ACP (CEDRE Maîtrise de la langue CM2 2015)

	Valeur Propre	Difference	Proportion	Proportion cumulee
1	57.6	51.2	0.28	0.28
2	6.4	2.2	0.03	0.31
3	4.2	0.21	0.02	0.33

La structure des items est fortement unidimensionnelle : le « poids » de la première dimension est très important (valeur propre de 57,6 contre 6,4 pour la deuxième dimension).

4 Modélisation

4.1 Méthodologie

4.1.1 Modèle de réponse à l'item

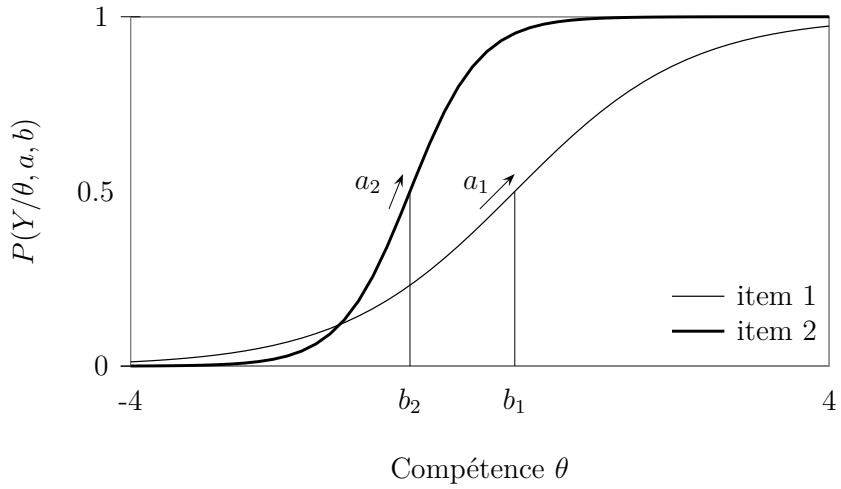
Le modèle de mesure utilisé est un modèle de réponse à l'item à deux paramètres avec une fonction de lien logistique (MRI 2PL) :

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

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

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

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



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

L'avantage de ce type de modélisation, c'est de séparer deux concepts-clé, à savoir la difficulté de l'item et le niveau de compétence de l'élève. Les MRI ont un intérêt pratique pour la construction de tests et la comparaison entre différents groupes d'élèves : si le modèle est bien spécifié sur un échantillon donné, les paramètres des items – en particulier leurs difficultés – peuvent être considérés comme fixes et applicables à d'autres échantillons dont il sera alors possible de déduire les paramètres relatifs aux élèves – en particulier, leur niveau de compétence. Pour une présentation générale, le lecteur est invité à consulter Rocher (2015).

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

4.1.2 Procédures d'estimation

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

Estimation des paramètres des items

Nous reprenons les notations de l'équation (10) qui formule la probabilité P_{ij} d'un élève i de répondre correctement à un item j dans le cadre d'un modèle de réponse à l'item, avec les items sont dichotomiques.

Notons tout d'abord que les modèles présentés ne sont pas identifiables. En effet, les transformations $\theta_i^* = A\theta_i + B$, $b_j^* = Ab_j + B$ et $a_j^* = a_j/A$ avec A et B deux constantes $(A > 0)$, conduisent aux mêmes valeurs des probabilités. Dans CEDRE, nous levons l'indétermination en standardisant la distribution des θ pour les données du premier cycle (en l'occurrence, moyenne de 250 et écart-type de 50 pour l'année 2003).

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

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

s'écrit :

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

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

La procédure MML (*Marginal Maximum Likelihood*) est utilisée. Elle consiste à estimer les paramètres des items en supposant que les paramètres des individus sont issus d'une distribution fixée *a prori* (le plus souvent normale). La maximisation de vraisemblance est *marginale* dans le sens où les paramètres concernant les individus n'apparaissent plus dans la formule de vraisemblance.

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

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

avec g la densité de θ.

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

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

Cependant, l'annulation des dérivées de L par rapport aux a_j et aux b_j conduit à résoudre un système d'équations relativement complexe et à procéder à des calculs d'intégrales qui peuvent s'avérer très coûteux en termes de temps de calcul.

La résolution de ces équations est classiquement réalisée grâce à l'algorithme EM (*Expectation-Maximization*) impliquant des approximations d'intégrales par points de quadrature. L'algorithme EM est théoriquement adapté dans le cas de valeurs manquantes. Le principe général est de calculer l'espérance conditionnelle de la vraisemblance des données complètes (incluant les valeurs manquantes) avec les valeurs des paramètres estimées à l'étape précédente, puis de maximiser cette espérance conditionnelle pour trouver les nouvelles valeurs des paramètres. Le calcul de l'espérance conditionnelle nécessite cependant de connaître (ou de supposer) la loi jointe des données complètes. Une version modifiée de l'algorithme considère dans notre cas le paramètre θ lui-même comme une donnée manquante. Pour plus de détails, le lecteur est invité à consulter Rocher (2013).

En outre, ce cadre d'estimation permet aisément de traiter des valeurs manquantes structurelles, par exemple dans le cas de cahiers tournants ou bien dans le cas de reprise partielle d'une évaluation.

Estimation des niveaux de compétence

Une fois les paramètres des items estimés, ils sont considérés comme fixes et il est possible d'estimer les θ_i, par exemple *via* la maximisation de la vraisemblance donnée par l'équation (11).

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

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

avec

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

et

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

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

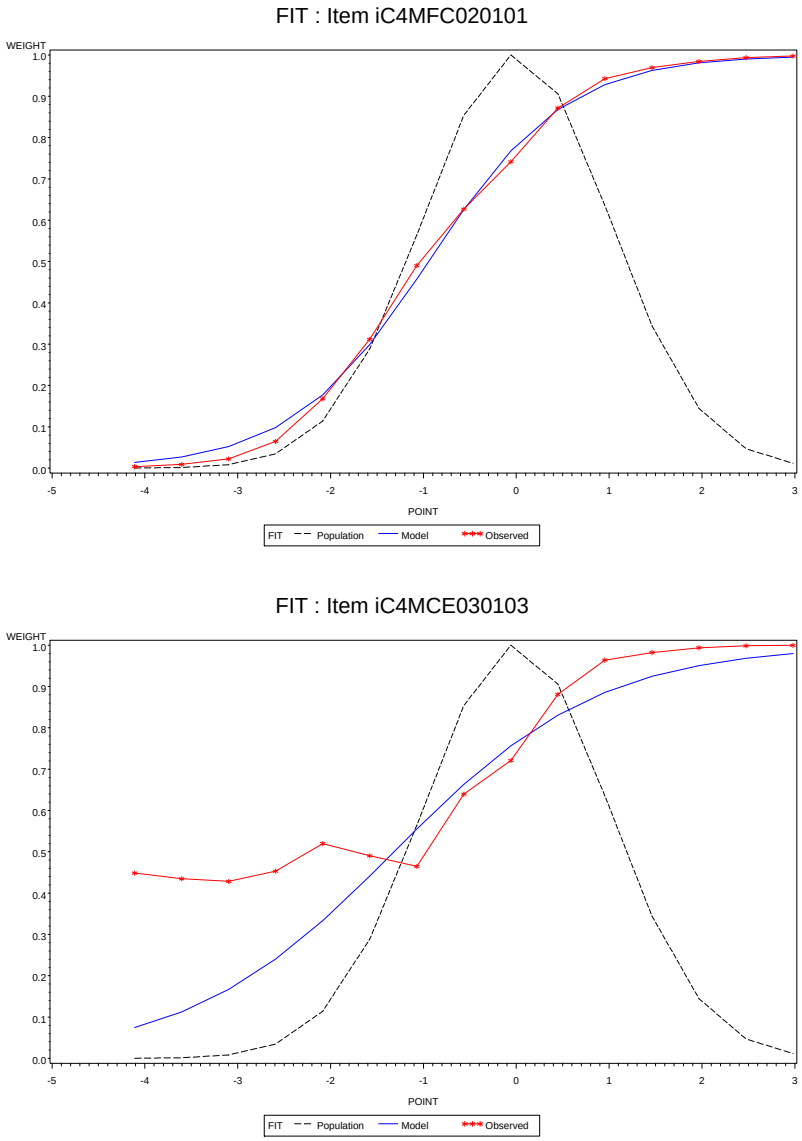
$$\frac{\partial lnL}{\partial \theta_i} + \frac{J}{2I} = 0 \tag{15}$$

4.1.3 Indice d'ajustement (FIT)

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

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

Figure 3 – Exemples d'ajustements (FIT)



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

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

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

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

4.1.4 Fonctionnement Différentiel d'Item (FDI)

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

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

$$P(Y \mid Z, G) \neq P(Y \mid Z) \tag{18}$$

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

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

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

Une stratégie très simple, employée dans CEDRE, consiste donc à comparer les paramètres de difficulté des items repris, estimés de façon séparée pour les deux

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

4.1.5 L'information du test

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

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

avec $P_j(\theta)$, la probabilité de réussite à l'item pour individu de compétence θ.

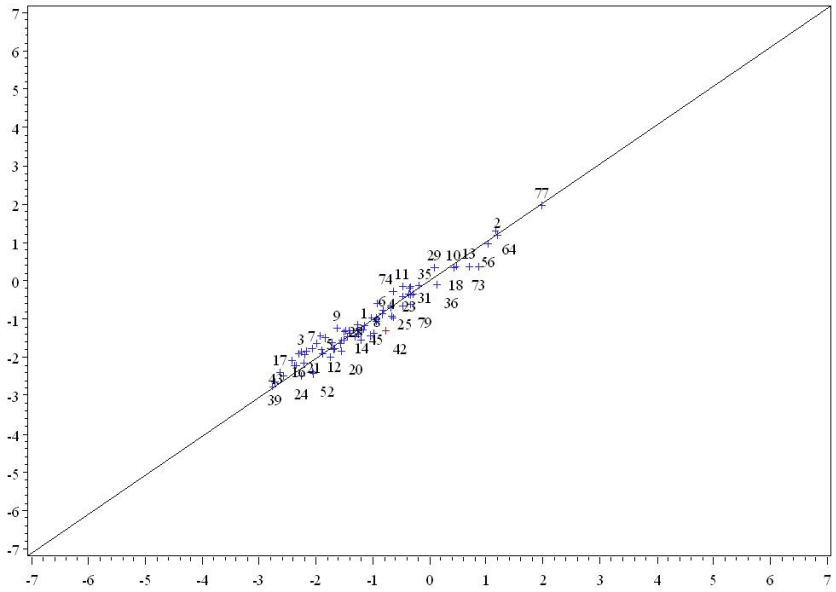
L'information moyenne du test pour un élève de compétence θ est la somme de l'information apporté par chaque item pour θ. La courbe d'information du test est tracée pour un ensemble de valeurs de θ. L'erreur de mesure étant inversement proportionnelle à l'information, cette courbe d'information permet de visualiser la précision avec laquelle le niveau de compétence θ des élèves est estimé.

4.2 Résultats

4.2.1 Identification des fonctionnements différentiels d'items (FDI)

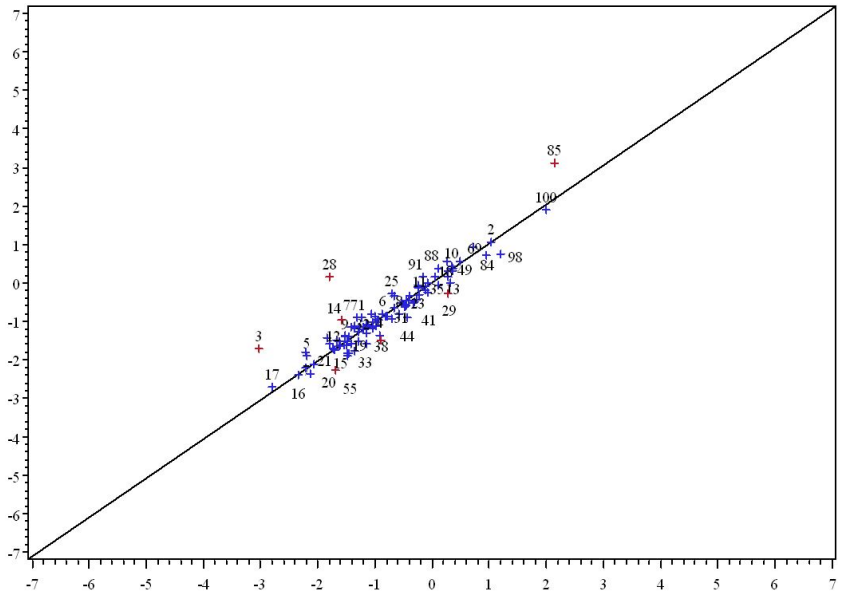
L'analyse des FDI 2003-2009 a permis de détecter 1 items en faveur de 2009 (figure 4). L'analyse des FDI 2009-2015 a permis de détecter 7 items : 4 en faveur de 2009 et 3 en faveur de 2015 (figure 5). Ces items ont été éliminés des calculs.

Figure 4 – Comparaison des paramètres de difficulté 2003-2009 (CEDRE Maîtrise de la langue CM2)



Note de lecture : Les points sont les items. En abscisse figure la valeur des paramètres de difficulté estimés en 2003, et en ordonnée la la valeur des paramètres de difficulté estimés et ajustés pour l'année 2009. Les items présentant un FDI apparaissent en rouge.

Figure 5 – Comparaison des paramètres de difficulté 2009-2015 (CEDRE Maîtrise de la langue CM2)



Note de lecture : Les points sont les items. En abscisse figure la valeur des paramètres de difficulté estimés en 2009, et en ordonnée la la valeur des paramètres de difficulté estimés et ajustés pour l'année 2015. Les items présentant un FDI apparaissent en rouge.

4.2.2 Identification des items présentant un mauvais ajustement (FIT)

L'analyse des ajustements (FIT) a conduit à supprimer 4 items : 2 items communs aux trois années, 1 item de 2009 et 1 item de 2015.

4.2.3 Bilan de l'analyse des items

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

- 37 items communs aux trois années
- 43 items communs à 2003 et 2009
- 64 items communs à 2009 et 2015
- 83 items de 2003
- 16 items de 2009
- 132 items de 2015

Cela représente 375 items passés par les élèves en tout, dont 233 en 2015.

Après suppression des items présentant un mauvais Rbis, un fonctionnement différentiel ou un mauvais ajustement, il reste :
– 35 items communs aux trois années
– 40 items communs à 2003 et 2009
– 57 items communs à 2009 et 2015
– 79 items de 2003
– 15 items de 2009
– 127 items de 2015
353 items sont donc conservés dans l'analyse, dont 219 utilisés dans l'évaluation 2015.

4.3 Calcul des scores

Comme indiqué précédemment, une analyse conjointe des données (2003 ,2009 et 2015) a permis d'estimer les paramètres des items, puis les niveaux de compétences θ des élèves. Afin de lever l'indétermination du modèle, la moyenne des θ a été fixé à 250 et leur écart-type à 50, pour l'échantillon de 2003. Le tableau 18 présente les résultats obtenus.

Tableau 18 – Niveaux de compétences CEDRE Maîtrise de la langue École (moyennes et écarts-type)

annee	Nb élèves	Moyenne	Ecart-Type
2003	6 110	250.0	50.0
2009	4 942	251.0	48.4
2015	7 428	250.7	43.0

5 Construction de l'échelle

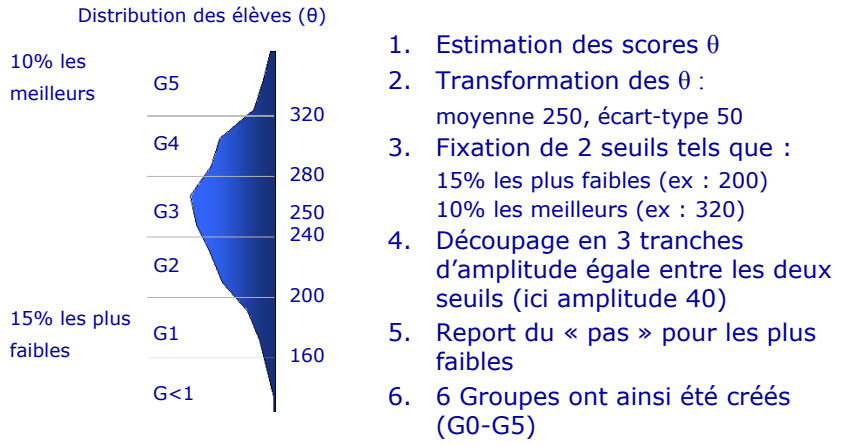
5.1 Méthode

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 Maîtrise de la langue 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 2003. Puis, comme le montre la figure 6, la distribution des scores est « découpée » en six groupes de la manière suivante : nous déterminons le score-seuil en-deça duquel se situent 15 % des élèves (groupes < 1 et 1), nous déterminons le score-seuil au-delà duquel se situent 10 % des élèves (groupe 5). Entre ces deux niveaux, l'échelle a été scindée en trois parties d'amplitudes de scores égales correspondant à trois groupes intermédiaires. Ces choix sont arbitraires et ont pour objectif de décrire plus précisément le continuum de compétence.

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

Figure 6 – Principes de construction de l'échelle



5.2 Caractérisation des groupes de niveaux

A partir de cette correspondance entre les items et les groupes, une description qualitative et synthétique des compétences maîtrisées par les élèves des différents groupes est proposée. Ces principaux résultats sont présentés dans une Note d'information (Andreu, Dalibard, & Eteve, 2016).

Groupe < 1 (1 % des élèves)

Des élèves en grande difficulté de lecture.

Leur compréhension reste locale. Ils ne sont pas en mesure de dresser un portrait global du texte. Ces élèves sont capables de prélèvement direct lorsque l'information est placée au début du texte. Ils sont capables de produire des inférences lorsque les informations sont contenues dans un texte et une image, et renvoient à des connaissances personnelles proches du quotidien. Ils savent identifier une famille de mots et le sens d'une expression familière.

Groupe 1 (10 % des élèves)

De faibles « compreneurs » et des lecteurs en difficulté générale de lecture. Leur compréhension reste locale. Ils ne sont pas en mesure de dresser un portrait global du texte. Ils peuvent mettre en lien deux informations spatialement proches l'une de l'autre pour identifier un lien de causalité. Ils repèrent des connecteurs logiques pertinents. Ils commencent à savoir catégoriser les types d'écrit (la règle du jeu). Ils associent un champ sémantique à un champ lexical et inversement.

Groupe 2 (29,1 % des élèves)

Des élèves capables de comprendre des écrits plus longs que ceux des groupes inférieurs et d'accéder à un premier niveau de synthèse pour des écrits courts.

Ils sont capables de prélever des informations et faire des inférences dans des écrits plus longs. Ils maîtrisent des compétences relevant principalement du cycle 2 : identifier une reprise anaphorique, replacer des événements dans l'ordre chronologique de l'histoire, associer un texte à une illustration, choisir un titre ou un résumé (pour un écrit court). En vocabulaire, ils savent retrouver le sens d'un mot ou d'une expression en contexte. Ils maîtrisent l'accord du verbe avec son sujet lorsque celui-ci est à la troisième personne du singulier. Ils connaissent l'orthographe lexicale d'usage.

Groupe 3 (33,1 % des élèves)

Des lecteurs capables d'accéder à la compréhension globale d'un écrit long et d'appréhender les pensées des personnages.

Ils sont en capacité de catégoriser des écrits ou d'identifier le genre d'un texte. Ils peuvent choisir un résumé long ou un titre pour un texte littéraire conséquent. Ils ont une compréhension globale d'un texte, ils interprètent les sentiments des personnages et appréhendent l'ambiance générale. Ils identifient les relations logiques qui constituent la trame narrative et repèrent le mot qui permet de comprendre une situation. Ils peuvent traiter une expression de sens figuré, trouver un synonyme, classer un mot dans une liste en respectant l'ordre alphabétique. Ils respectent l'accord en genre et en nombre dans le groupe nominal. Ils prennent appui sur l'orthographe grammaticale pour identifier le genre et le nombre de personnages. Ils ont acquis une maîtrise satisfaisante des compétences et connaissances exigibles en fin d'école.

Groupe 4 (19,7 % des élèves)

Des lecteurs qui déploient des stratégies et atteignent une représentation globale du texte.

Ces élèves comprennent les intentions de l'auteur et prélèvent des informations dispersées dans les textes pour formuler des réponses construites à partir d'inférences multiples. Ils identifient l'élément déclencheur du récit. Ils sont capables de généraliser le thème d'un texte pour en extraire une morale ou encore de dégager le thème commun à deux textes. Ils respectent l'accord sujet verbe lorsque le verbe est mis en apposition. Ils maîtrisent une très grande partie des compétences et connaissances en maîtrise de la langue exigibles en fin d'école.

Groupe 5 (7,2 % des élèves)

Des lecteurs critiques qui ont une compréhension de l'évolution de la trame narrative et peuvent identifier des registres de langue.

Ces élèves sont capables de saisir l'atmosphère et le ton d'un texte pour comprendre les intentions des personnages. Ils effectuent des inférences sur la personnalité des protagonistes en convoquant des connaissances personnelles. Ils savent interpréter l'évolution des sentiments d'un personnage au fil d'un texte. En vocabulaire, ils savent notamment identifier un synonyme en passant du registre de langue soutenue au registre courant. Ces élèves ont une maîtrise parfaite des compétences et connaissances exigibles en maîtrise de la langue à la fin d'école.

5.3 Exemples d'items

5.3.1 Item caractéristique du groupe < 1

Le support de cet item regroupe quatre textes et quatre illustrations issues d'une anthologie poétique. Les illustrations sont de type gravure, genre pictural caricature, en noir et blanc, très éloignées d'une représentation scientifique des insectes. Les poèmes comportent entre huit et onze lignes chacun. Les poèmes A et C sont écrit dans un registre de langue plutôt humoristique, alors que le poème B décrit une situation proche de la réalité et le D s'apparente à un exercice de style sur un mode onirique. Le niveau lexical est complexe.

Figure 7 – Exemple groupe < à 1

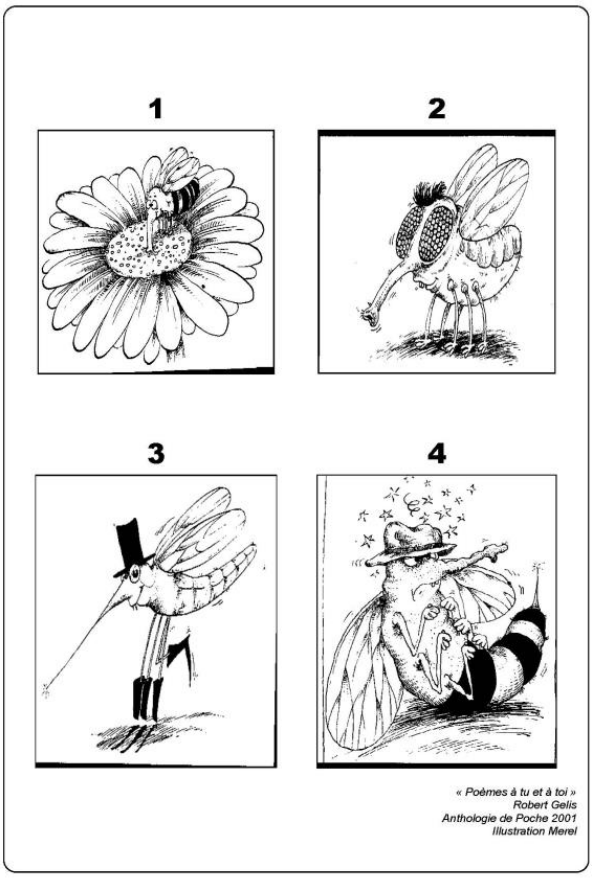
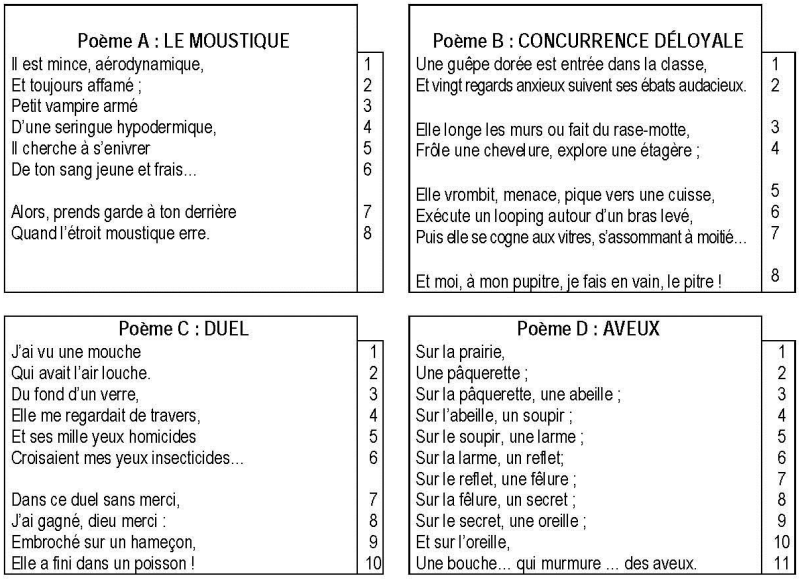


Figure 8 – Exemple groupe < à 1



Poème A : LE MOUSTIQUE

Il est mince, aérodynamique, 1
Et toujours affamé ; 2
Petit vampire armé 3
D'une seringue hypodermique, 4
Il cherche à s'enivrer 5
De ton sang jeune et frais... 6

Alors, prends garde à ton derrière 7
Quand l'étroit moustique erre. 8

Poème B : CONCURRENCE DÉLOYALE

Une guêpe dorée est entrée dans la classe, 1
Et vingt regards anxieux suivent ses ébats audacieux. 2

Elle longe les murs ou fait du rase-motte, 3
Frôle une chevelure, explore une étagère ; 4

Elle vrombit, menace, pique vers une cuisse, 5
Exécute un looping autour d'un bras levé, 6
Puis elle se cogne aux vitres, s'assommant à moitié... 7

Et moi, à mon pupitre, je fais en vain, le pitre ! 8

Poème C : DUEL

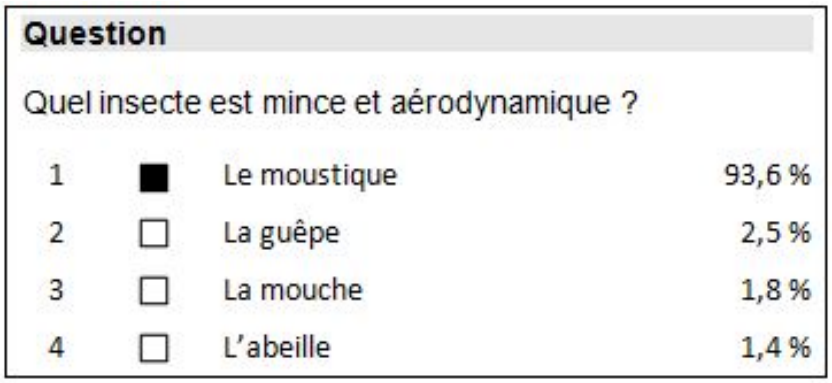
J'ai vu une mouche 1
Qui avait l'air louche. 2
Du fond d'un verre, 3
Elle me regardait de travers, 4
Et ses mille yeux homicides 5
Croisaient mes yeux insecticides... 6

Dans ce duel sans merci, 7
J'ai gagné, dieu merci : 8
Embroché sur un hameçon, 9
Elle a fini dans un poisson ! 10

Poème D : AVEUX

Sur la prairie, 1
Une pâquerette ; 2
Sur la pâquerette, une abeille ; 3
Sur l'abeille, un soupir ; 4
Sur le soupir, une larme ; 5
Sur la larme, un reflet; 6
Sur le reflet, une fêlure ; 7
Sur la fêlure, un secret ; 8
Sur le secret, une oreille ; 9
Et sur l'oreille, 10
Une bouche... qui murmure ... des aveux. 11

Figure 9 – Exemple groupe < à 1



Question

Quel insecte est mince et aérodynamique ?

1	■	Le moustique	93,6 %
2	□	La guêpe	2,5 %
3	□	La mouche	1,8 %
4	□	L'abeille	1,4 %

Cet exercice est réussi à plus de 93 % dans son ensemble. Ce fort taux de réussite s'explique par le fait que les informations à prélever sont faciles d'accès et spatialement proches (titre et première ligne). Pour répondre à la question, il est nécessaire de lire le premier vers et de le référer au titre du poème (le moustique). Le caractère non scientifique des dessins d'illustrations des insectes est cependant distracteur, n'apportant quasiment aucune aide aux élèves.

5.3.2 Item caractéristiques du groupe 1

Le support de cet item est un texte littéraire, non illustré. Le contexte est courant et le texte ne présente pas difficulté de lexique particulière.

Figure 10 – Exemple groupe 1

Lorsqu' elle allait au marché, ma mère me laissait au passage dans la salle de classe de mon père, qui apprenait à lire à des gamins de six ou sept ans. Je restais assis, bien sage, au premier rang et j'admirais la toute-puissance paternelle. Il tenait à la main une baguette de bambou : elle lui servait à montrer les lettres et les mots qu'il écrivait au tableau noir.
Un beau matin, ma mère me déposa à ma place, et sortit sans mot dire, pendant qu'il écrivait magnifiquement sur le tableau : « La maman a puni son petit garçon qui n'était pas sage. »
Tandis qu'il arrondissait un admirable point final, je criai : « Non ! Ce n'est pas vrai ! »
Mon père se retourna soudain, me regarda stupéfait, et s'écria : « Qu'est-ce que tu dis ? »
- Maman ne m'a pas puni ! Tu n'as pas bien écrit !...

D'après La gloire de mon père
De Marcel Pagnol

Figure 11 – Exemple groupe 1

Question

Le métier du père est ...

1	☐	écrivain.	2,8 %
2	■	maître d'école.	94 %
3	☐	surveillant.	1,1 %
4	☐	gardien d'école.	0,8 %

Cet item est réussi à 94 %. La question posée nécessite que l'élève réalise une inférence globale, mettant en lien des éléments textuels et des connaissances personnelles. Pour ce faire, l'élève doit prélever des informations dans les deux premières lignes. L'accès aisé, la proximité des informations et la familiarité du contexte facilitent l'inférence demandée.

5.3.3 Item caractéristique du groupe 2

Le support est un texte narratif de dix-huit lignes, comportant un discours rapporté. Le contexte est familier. Le lexique est de difficulté moyenne (quelques termes relevant d'un niveau de langue soutenu).

Figure 12 – Exemple groupe 2

« Ça m'étonnerait fort ! commenta Sarah.
- Complètement impossible ! ajouta Steven.
- Tu n'es qu'un vantard, on ne te croit pas ! conclut Loïc.
- Ah, c'est comme ça ? Vous croyez que j'ai tout inventé ? Eh bien regardez ! »
Geoffrey posa ses mains au-dessus de la tasse à café, ferma les yeux et se concentra en murmurant des formules bizarres :
« Que les forces de l'étrange me viennent en aide ! Que l'énergie du mystère suprême soulève cette tasse ! »
Geoffrey transpirait à grosses gouttes, ses mains étaient prises de tremblements incontrôlables, comme s'il était plongé dans une transe magique.
Et là sous les yeux ébahis de ses trois amis, la tasse se souleva de quelques millimètres ...
Ils échangèrent des regards incrédules, ne sachant plus que penser.
Le père de Geoffrey rompit le charme. Il fit irruption dans la chambre, et ouvrant les rideaux, déclara :
« Mais qu'est-ce que vous faites dans l'obscurité ? Éteignez cette bougie, vous allez mettre le feu à la maison ! Dis, Geoffrey, tu n'aurais pas vu ma bobine de fil de pêche ? Je la cherche depuis une demi-heure ! »

Figure 13 – Exemple groupe 2

Question

Parmi ces quatre propositions, quel titre choisirais-tu pour ce texte ?

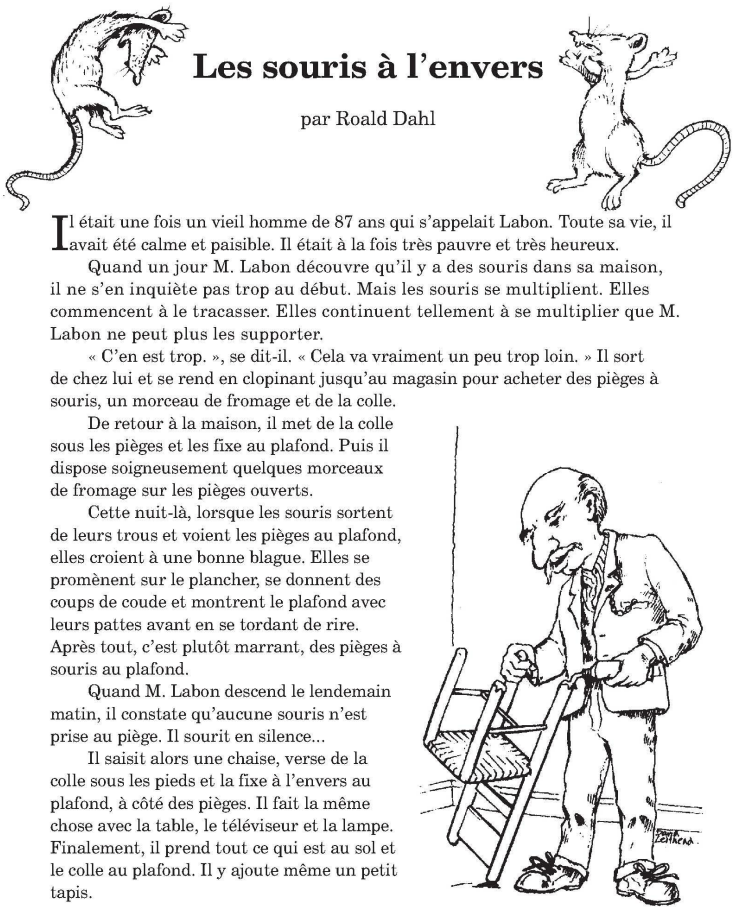
1 ☐	Un agréable goûter chez des amis	7 %
2 ☐	Le mensonge de Steven	4,6 %
3 ■	Le tour de magie dévoilé	85,7 %
4 ☐	Un incendie accidentel	1,8 %

Cet item est réussi à près de 86 %. L'élève doit ici choisir un titre. Pour ce faire, il doit prélever des informations dans les deux premières lignes. L'accès aisé, la proximité des informations et la familiarité du contexte facilitent la réalistaion de cette compétence de synthèse.

5.3.4 Item caractéristique du groupe 3

Le support est un texte narratif de Roald Dahl, comportant quatre illustrations et d'une longueur de cinquante-deux lignes. Le ton est humoristique et le langage est celui de la vie courante.

Figure 14 – Exemple groupe 3



Les souris à l'envers

par Roald Dahl

Il était une fois un vieil homme de 87 ans qui s'appelait Labon. Toute sa vie, il avait été calme et paisible. Il était à la fois très pauvre et très heureux.

Quand un jour M. Labon découvre qu'il y a des souris dans sa maison, il ne s'en inquiète pas trop au début. Mais les souris se multiplient. Elles commencent à le tracasser. Elles continuent tellement à se multiplier que M. Labon ne peut plus les supporter.

« C'en est trop. », se dit-il. « Cela va vraiment un peu trop loin. » Il sort de chez lui et se rend en clopinant jusqu'au magasin pour acheter des pièges à souris, un morceau de fromage et de la colle.

De retour à la maison, il met de la colle sous les pièges et les fixe au plafond. Puis il dispose soigneusement quelques morceaux de fromage sur les pièges ouverts.

Cette nuit-là, lorsque les souris sortent de leurs trous et voient les pièges au plafond, elles croient à une bonne blague. Elles se promènent sur le plancher, se donnent des coups de coude et montrent le plafond avec leurs pattes avant en se tordant de rire. Après tout, c'est plutôt marrant, des pièges à souris au plafond.

Quand M. Labon descend le lendemain matin, il constate qu'aucune souris n'est prise au piège. Il sourit en silence...

Il saisit alors une chaise, verse de la colle sous les pieds et la fixe à l'envers au plafond, à côté des pièges. Il fait la même chose avec la table, le téléviseur et la lampe. Finalement, il prend tout ce qui est au sol et le colle au plafond. Il y ajoute même un petit tapis.

Cette nuit-là, les souris sortent de leurs trous en ricanant et en faisant des plaisanteries sur ce qu'elles ont vu la veille. Mais cette fois, quand elles regardent au plafond, elles arrêtent de rire brusquement.

« Hé ! Regardez ! Voilà que le sol est là-haut ! », s'écrie l'une d'elles.

« Incroyable ! Alors, nous devons être au plafond ! », s'exclame une autre.

« Je commence à me sentir un peu étourdie », dit une autre.

« Le sang me descend à la tête », se plaint une quatrième.

« C'est horrible ! », dit une très vieille souris aux longues moustaches. « C'est vraiment horrible ! Il faut faire quelque chose tout de suite ! »

« Je vais m'évanouir si je dois me tenir une seconde de plus sur la tête ! », crie une jeune souris.

« Moi aussi ! »

« Je n'en peux plus ! »

« Au secours ! Vite, que quelqu'un fasse quelque chose ! »

Elles devenaient hystériques. « Je sais ce que nous allons faire », dit la très vieille souris. « Nous allons toutes nous tenir sur la tête et alors nous serons dans le bon sens. »

Docilement, elles se placent toutes sur la tête et au bout d'un long moment, le sang coulant vers leur cerveau, elles s'évanouissent l'une après l'autre.

Quand M. Labon descend le lendemain matin, le sol est couvert de souris. Il les ramasse rapidement et les met dans un panier.

Voici ce qu'il faut retenir de cette histoire : chaque fois que le monde semble à l'envers, mieux vaut rester les pieds sur terre.

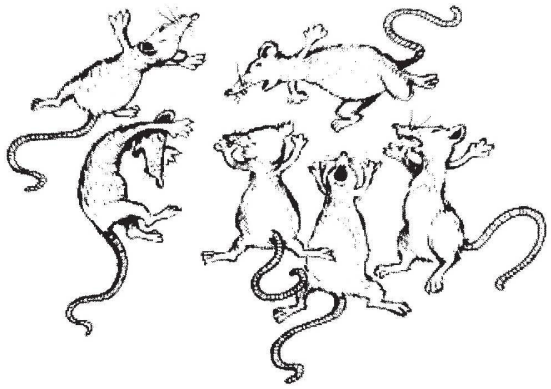
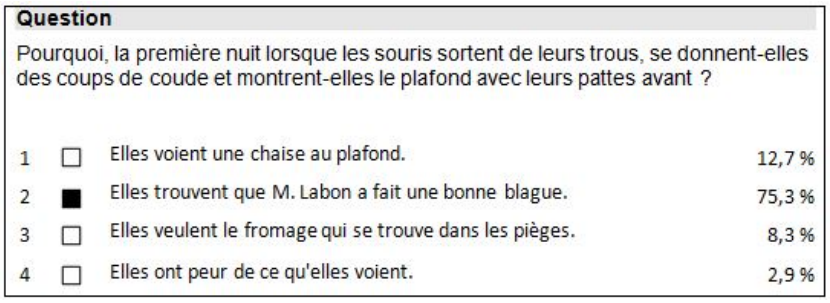


Figure 15 – Exemple groupe 3



Cet item est réussi à plus de 75 %.

Il demande à l'élève de réaliser une inférence globale qui consiste à appréhender l'ambiance générale d'un texte.

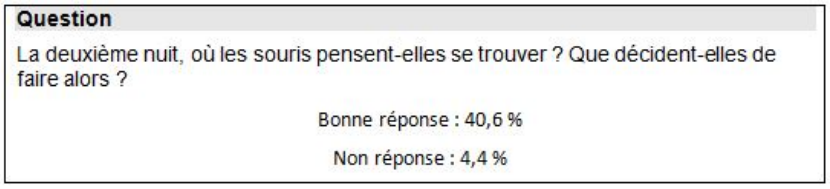
Il doit analyser le 3ème paragraphe et mettre en lien l'information donnée par le texte et l'étrangeté de la situation (gravité). Il doit également savoir que se donner des coups de coude est une attitude de moquerie complice. Les deux premières modalités de réponse sont acceptables, ce qui peut contribuer à expliquer que le taux de réussite n'est pas plus important.

La présence du mot « blague » dans le texte et dans la réponse attendue permet d'emporter la décision de choisir la deuxième modalité de réponse.

5.3.5 Item caractéristique du groupe 4

Le support est le même (texte de Roald Dahl).

Figure 16 – Exemple groupe 4



La question est ouverte ; elle implique une réponse construite. Le taux de réussite approche les 41 %. Deux éléments de réponse sont attendus : un lieu (le plafond) et une action (se tenir sur la tête pour être dans le bon sens). L'élève doit retrouver les deux phrases du texte qui permettent de répondre aux deux questions. Il lui faut également comprendre que l'expression « cette nuit-là » correspond à « la deuxième nuit » évoquée dans la question. Il doit aussi appréhender qui parle dans le texte pour répondre.

Figure 17 – Extrait du guide de codage

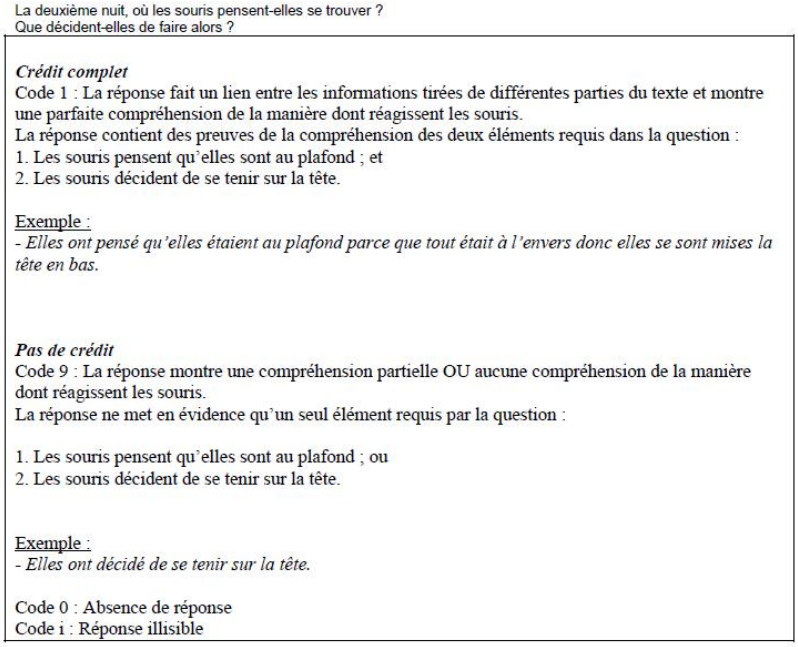


Figure 18 – Exemple de réponse d'élève

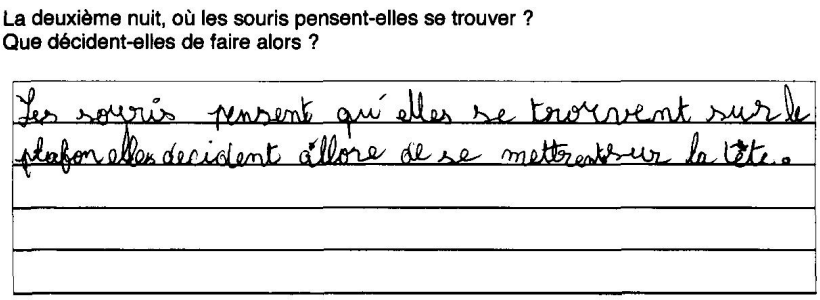


Figure 19 – Exemple de réponse d'élève

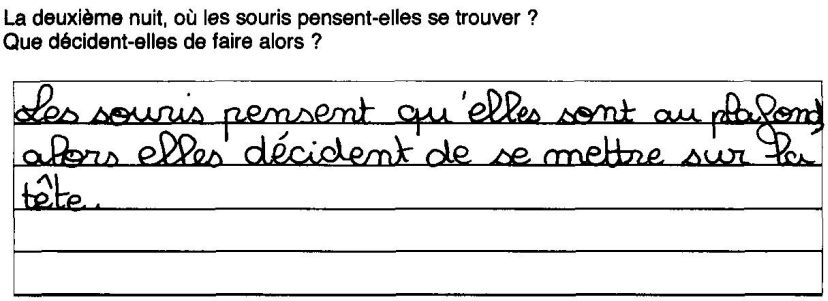
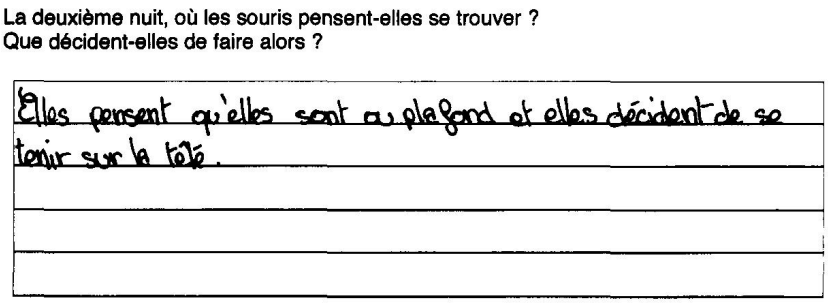


Figure 20 – Exemple de réponse d'élève



5.3.6 Items caractéristiques du groupe 5

Le support est le même (texte de Roald Dahl).

Figure 21 – Exemple groupe 5

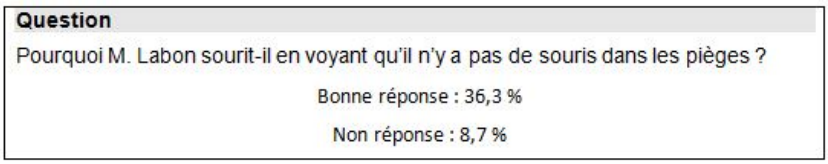
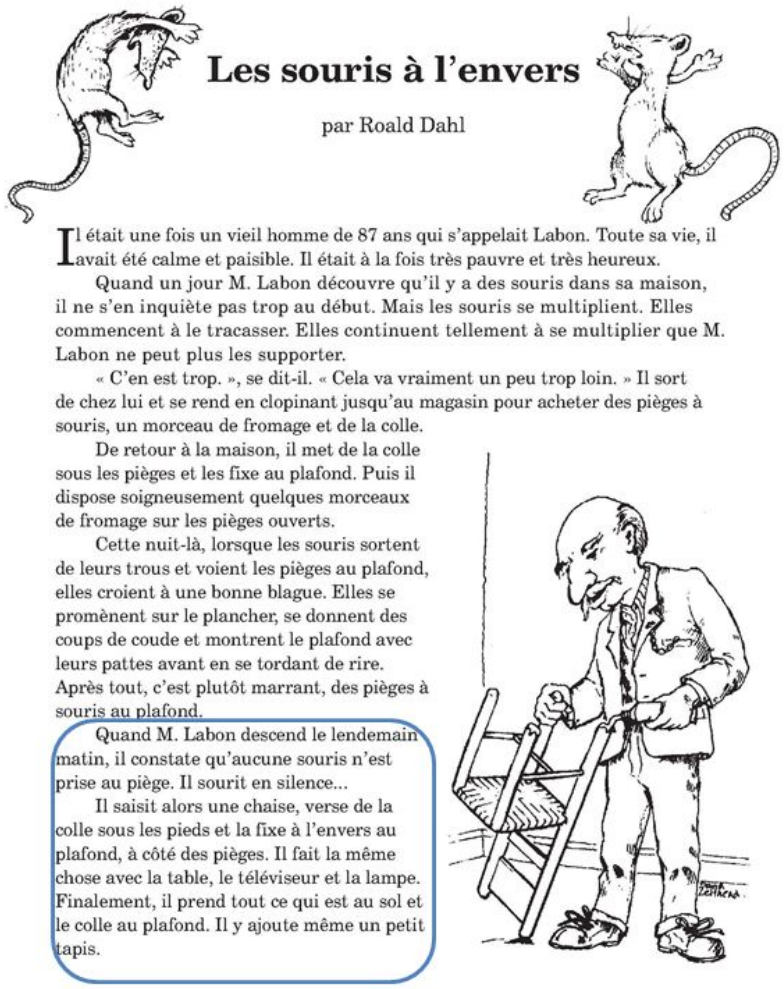


Figure 22 – Exemple groupe 5



Les souris à l'envers

par Roald Dahl

Il était une fois un vieil homme de 87 ans qui s'appelait Labon. Toute sa vie, il avait été calme et paisible. Il était à la fois très pauvre et très heureux.

Quand un jour M. Labon découvre qu'il y a des souris dans sa maison, il ne s'en inquiète pas trop au début. Mais les souris se multiplient. Elles commencent à le tracasser. Elles continuent tellement à se multiplier que M. Labon ne peut plus les supporter.

« C'en est trop. », se dit-il. « Cela va vraiment un peu trop loin. » Il sort de chez lui et se rend en clopinant jusqu'au magasin pour acheter des pièges à souris, un morceau de fromage et de la colle.

De retour à la maison, il met de la colle sous les pièges et les fixe au plafond. Puis il dispose soigneusement quelques morceaux de fromage sur les pièges ouverts.

Cette nuit-là, lorsque les souris sortent de leurs trous et voient les pièges au plafond, elles croient à une bonne blague. Elles se promènent sur le plancher, se donnent des coups de coude et montrent le plafond avec leurs pattes avant en se tordant de rire. Après tout, c'est plutôt marrant, des pièges à souris au plafond.

Quand M. Labon descend le lendemain matin, il constate qu'aucune souris n'est prise au piège. Il sourit en silence...

Il saisit alors une chaise, verse de la colle sous les pieds et la fixe à l'envers au plafond, à côté des pièges. Il fait la même chose avec la table, le téléviseur et la lampe. Finalement, il prend tout ce qui est au sol et le colle au plafond. Il y ajoute même un petit tapis.

Le taux de réussite à cette question à réponse construite est de 36,3 %. Ici, l'élève doit réaliser des inférences multiples pour analyser et comprendre les sentiments, les intentions (non abouties) et les pensées des personnages. L'une des difficultés consiste à percevoir que le plan de M. Labon se déroule en deux temps.

Figure 23 – Extrait du guide de codage

Pourquoi M. Labon sourit-il en voyant qu'il n'y a pas de souris dans les pièges ?

Crédit complet
Code 1 : La réponse fournit une interprétation appropriée de la réaction de M. Labon dans le contexte général de l'histoire.
La réponse démontre que l'élève a compris que M. Labon n'était pas surpris de trouver les pièges vides. Elle peut décrire les intentions de M. Labon de mettre en œuvre un plan plus élaboré afin d'attraper les souris. OU la réponse peut montrer que l'élève a compris que M. Labon ne cherchait qu'à tromper les souris, pas à les attraper dès la première nuit.

Exemples :
- Il avait un plan pour tromper les souris et s'en débarrasser.
- Il savait qu'elles n'iraient pas chercher le fromage la première nuit.

Pas de crédit
Code 9 : Autres réponses
Code 0 : Absence de réponse
Code i : Réponse illisible

Figure 24 – Exemple de réponse d'élève

Pourquoi M. Labon sourit-il en voyant qu'il n'y a pas de souris dans les pièges ?

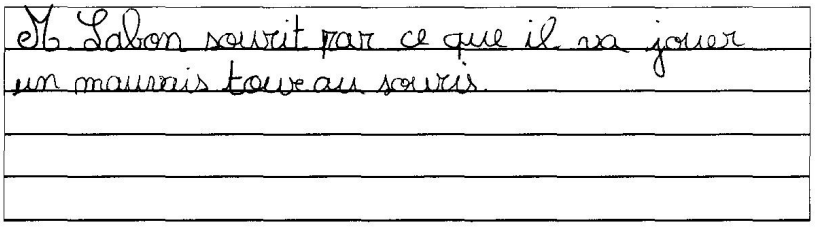


Figure 25 – Exemple de réponse d'élève

Pourquoi M. Labon sourit-il en voyant qu'il n'y a pas de souris dans les pièges ?

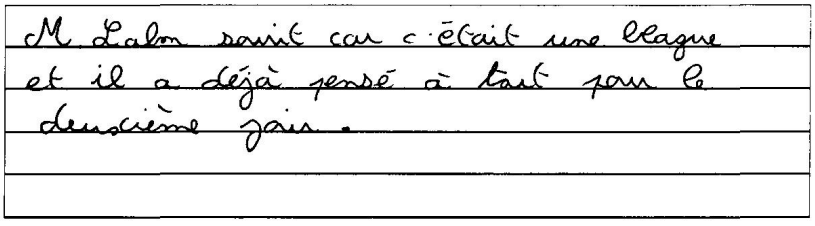
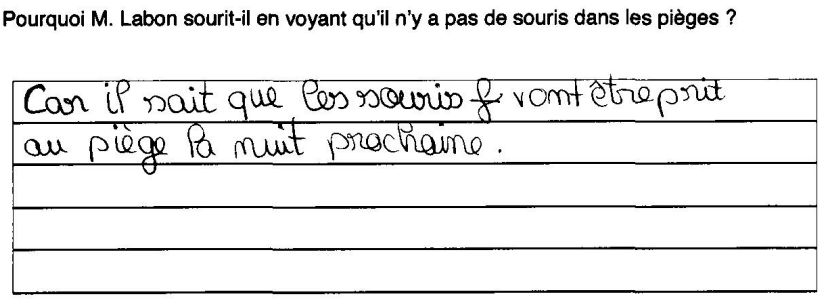


Figure 26 – Exemple de réponse d'élève



6 Variables contextuelles et non cognitives

6.1 Variables sociodémographiques et indice de position sociale

Un certain nombre de variables sociodémographiques permettent d'enrichir l'analyse des résultats. Le score moyen des élèves est ainsi analysé en fonction du genre, du retard scolaire et quand les effectifs le permettent en fonction du secteur d'enseignement. Le lecteur est invité à consulter la Note d'Information pour plus de détails (Andreu et al., 2016).

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

Il n'a été possible d'établir des comparaisons qu'en termes de niveau social des écoles, et non au niveau individuel. En effet, en 2015, la PCS des parents est disponible pour chaque élève, mais elle ne l'était pas dans les cycles antérieurs. Pour chaque établissement des échantillons de 2003, 2009 et 2015, la moyenne de l'indice de position socio-scolaire a été calculée et la population a ensuite été découpée en quatre groupes selon les quartiles (tableau 19).

Tableau 19 – Score moyen selon l'indice de position sociale moyen de l'établissement (CEDRE maîtrise de la langue école 2003, 2009 et 2015)

Indice moyen école	Année	Répartition (%)	Score moyen	Écart type
1er quart	2003	24,8	231	49
1er quart	2009	23,8	**238**	48
1er quart	2015	24,6	235	**42**
2e quart	2003	25,1	247	49
2e quart	2009	25,6	249	**46**
2e quart	2015	25,2	250	43
3e quart	2003	24,9	256	48
3e quart	2009	24,4	259	48
3e quart	2015	24,9	252	**41**
4e quart	2003	25,1	266	47
4e quart	2009	26,2	**258**	49
4e quart	2015	25,3	265	**41**

Note de lecture : en 2015, l'écart type des élèves appartenant au quart des écoles les plus défavorisées (1er quart) baisse de 7 points par rapport à 2003. Les évolutions significatives sont indiquées en gras.

6.2 Élaboration des questionnaires de contexte

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

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

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

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

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

6.3 Construction des scores factoriels et des indicateurs

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

Le tableau 20 présente en guise d'illustration les items d'une de ces dimensions, en l'occurrence la perception de soi.

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

Tableau 20 – Exemple de variable conative - perception de soi (CEDRE maîtrise de la langue école 2015)

Question	**1er Axe ACP**
J'ai l'impression que je suis très bon(ne) dans mon travail scolaire	0,77
Je suis assez satisfait(e) de moi	0,73
Je fais très bien mon travail en classe	0,71
J'aime bien le genre de personne que je suis	0,71
Je suis très content(e) d'être comme je suis	0,71
En général, j'aime bien la façon dont je mène ma vie	0,69
Je me sens aussi intelligent(e) que ceux de mon âge	0,62
Je suis capable de me rappeler facilement des choses	0,58

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. Plus la valeur de l'indicateur est élevé, et plus grande est « l'adhésion » de l'élève à la dimension correspondante.

6.4 Motivation des élèves face à la situation d'évaluation

Les évaluations standardisées des élèves, telles que CEDRE ou PISA, renvoient à des enjeux politiques croissants, alors qu'elles restent à faible enjeu pour les élèves participants. Dans le système éducatif français, où la notation tient une place prépondérante, la question de la motivation des élèves face à ces évaluations mérite d'être posée.

Un instrument pour mesurer la motivation a été adapté à partir du « thermomètre d'effort » proposé dans PISA (Keskpaik. & Rocher, 2015). Cet instrument (cf. figure 27) a été introduit dans plusieurs évaluations conduites au niveau national par la DEPP, y compris dans CEDRE maîtrise de la langue. Les données recueillies permettent de distinguer la motivation de l'élève de la difficulté perçue du test, et ainsi de mieux appréhender le lien entre la motivation des élèves français et leur performance. L'analyse de ces données renseigne en outre sur le rôle de certaines caractéristiques, des élèves ou des évaluations elles-mêmes, dans le degré de motivation à répondre aux questions de l'évaluation.

Le tableau 21 présente les grands résultats de cet instrument.

Tableau 21 – Résultats de l'instrument de mesure de la motivation au test (CEDRE maîtrise de la langue école 2015)

		%
Comment as-tu trouvé les exercices de cette évaluation ?	Très faciles	12,6
	Faciles	67,9
	Difficiles	18,3
	Très difficiles	1,2
Je me suis bien appliqué(e) pour faire cette évaluation	Pas du tout d'accord	1,6
	Pas d'accord	7,2
	D'accord	57,5
	Tout à fait d'accord	33,7
Je me suis autant appliqué(e) pour faire cette évaluation que le travail quotidien de classe	Pas du tout d'accord	5,4
	Pas d'accord	16,3
	D'accord	40,3
	Tout à fait d'accord	38,0

Figure 27 – Instrument de mesure de la motivation au test

[Q1]

Comment as-tu trouvé les exercices de cette évaluation ?

1☐ Très faciles
2☐ Faciles
3☐ Difficiles
4☐ Très difficiles

[Q2]

Es-tu d'accord avec ces affirmations ?

(Coche une case par ligne)

	Pas du tout d'accord	Pas d'accord	D'accord	Tout à fait d'accord
Je me suis bien appliqué(e) pour faire cette évaluation	☐ 1	☐ 2	☐ 3	☐ 4
Je me suis autant appliqué(e) à faire cette évaluation que le travail quotidien de classe	☐ 1	☐ 2	☐ 3	☐ 4

7 Annexe

Certification AFNOR pour les évaluations CEDRE

La DEPP est engagée dans un processus de certification. Elle a obtenu en mars 2015 la certification pour les évaluations CEDRE.

Les finalités de la certification

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

Les enjeux pour la DEPP

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

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

La démarche qualité

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

Les engagements de service

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

Les engagements de service de la DEPP

Des objectifs clairs et partagés

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

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

Des évaluations fondées sur l'expertise pédagogique

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

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

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

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

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

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

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

Une mesure fiable et des comparaisons temporelles pertinentes

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

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

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

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

Des analyses enrichies par des données de contexte

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

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

Transparence des méthodes et partage des résultats

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

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

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

Références

Andreu, S., Dalibard, E., & Eteve, Y. (2016). CEDRE 2013 - 2009 - 2015 - maîtrise de la langue en fin d'école : l'écart se creuse entre filles et garçons. *Note d'information*, *19*.

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

Christine, M., & Rocher, T. (2012, janvier). Construction d'échantillons astreints à des conditions de recouvrement par rapport à un échantillon antérieur et à des conditions d'équilibrage par rapport à des variables courantes : aspects théoriques et mise en œuvre dans le cadre du renouvellement des échantillons des enquêtes d'évaluation des élèves. In *Journées de méthodologie statistique.* Paris.

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

Keskpaik., S., & Rocher, T. (2015). La motivation des élèves français face à des évaluations à faibles enjeux. comment la mesurer ? son impact sur les réponses. *Education et formations*, *85-86*, 119-139.

Rocher, T. (1999). *Psychométrie et théorie des sondages* (Mémoire de Master non publié). Université Paris VI.

Rocher, T. (2013). *Mesure des compétences : les méthodes se valent-elles ? questions de psychométrie dans le cadre de l'évaluation de la compréhension de l'écrit* (Thèse de doctorat non publiée). Université Paris-Ouest.

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

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

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

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

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

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

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

Liste des tableaux

Table des figures