CEDRE

Cycle des Évaluations Disciplinaires Réalisées sur Échantillons

Rapport technique

Anglais 2016

Collège

Auteurs :

Sylvie BEUZON
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

Octobre 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 échantillon (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 au regard des objectifs fixés par les programmes officiels. 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 régulièrement, 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 rendues publiques car, devant être en grande partie reprises lors des cycles d'évaluation suivants, elles ne doivent pas servir d'exercices dans les classes.

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

1.2 Les compétences et connaissances visées

1.2.1 Constition des épreuves

En anglais, l'évaluation a été proposée dans quatre activités langagières : la compréhension de l'oral, la compréhension de l'écrit, l'expression écrite et, pour

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

la première fois en 2016, l'expression orale en continu.

L'évaluation CEDRE a été élaborée à partir des objectifs fixés par les programmes officiels entrés en vigueur à la rentrée 2006 (BO n°6 du 25 août 2005), programmes adossés au Cadre européen commun de référence pour les langues (CECRL). Les situations d'évaluation relèvent pour la plupart d'entre elles du niveau A2 *(intermédiaire)*. En effet, dans le cadre des programmes de 2005, pour la validation du socle commun de connaissances et de compétences, c'est ce niveau A2 qui est requis ; la majorité des items proposés relèvent donc de ce niveau de compétence. Néanmoins, afin d'apprécier au mieux les différentes performances des élèves, des items de niveau A1 *(découverte)* et d'autres d'un niveau tendant vers B1 *(indépendant)* ont également été proposés.

1.2.2 Objectifs d'évaluation et supports

Dans chacune des activités langagières retenues, on a évalué les connaissances et compétences des élèves. Pour ce faire, on a retenu plusieurs grands domaines de savoirs et de savoir-faire, gradués en fonction de la complexité croissante des opérations mentales nécessaires pour les mettre en oeuvre.

En **compréhension de l'oral**, on a vérifié que les élèves sont capables, dans un message sonore, de repérer des informations explicites (lexique de la vie quotidienne, éléments culturels simples, repères temporels et spatiaux) et de construire du sens en inférant à partir de l'explicite, en identifiant l'implicite et en synthétisant l'information grâce à une mise en relation de divers types d'indices (linguistiques, paralinguistiques, culturels). Pour évaluer la compétence de compréhension de l'oral (« écouter et comprendre »), des supports variés balayant un large éventail de situations ont été proposés : extraits de blogs, témoignages, messages téléphoniques, bulletins météo, annonces publiques, dialogues. Pour la première fois, des supports vidéo ont également été proposés.

Comme en compréhension de l'oral, on a mesuré en **compréhension de l'écrit**

les aptitudes des élèves à reconnaître dans un support des expressions mémorisées et un lexique de la vie quotidienne, à identifier l'information pertinente (repères culturels, thème, informations explicites, repères temporels et spatiaux) et à construire le sens en identifiant l'information implicite, en inférant le sens d'une expression et en synthétisant. Un ensemble de différents types d'écrits (articles de presse, lettres, courriers et messages électroniques, sites Internet, brochures et prospectus) a été proposé. Au total, 25 supports écrits (lecture d'un ou plusieurs documents) ont été utilisés, dont 13 accompagnés d'iconographie.

En **expression écrite**, on a vérifié que les élèves sont capables d'écrire des mots isolés, des énoncés simples et brefs, des phrases reliées par des connecteurs simples et des textes articulés et nuancés. A partir de situations contextualisées (rédaction de courriels, de lettres, cartes postales et messages de différents formats) et de supports iconographiques, les élèves étaient dans certains cas guidés pour rédiger ; dans d'autres, il leur était demandé une production plus autonome.

Enfin, en **expression orale en continu**, on a évalué les aptitudes à décrire et justifier un choix à partir d'un scénario proposé ainsi que des supports iconographiques variés, déclencheurs de parole et à fort ancrage culturel. Pour des raisons de faisabilité, seul un sous-échantillon d'élèves (3 élèves par classe) a été évalué dans cette activité langagière. L'élaboration des grilles de compétences (tableaux 2, 3, 4 et 5) en vue de la construction des items s'est fondée sur les documents de référence : les programmes officiels (BO n°6 du 25 août 2005 entré en vigueur en 2006).

Tableau 2 – Définition des compétences en compréhension de l'oral (évaluation 2016)

Compréhension de l'oral **Objectifs d'évaluation**	
Repérer l'information explicite	Repérer le lexique de la vie quotidienne Repérer des éléments culturels simples Identifier des repères temporels et spatiaux
Construire le sens	Mettre en réseau des informations explicites Inférer à partir d'éléments explicites Identifier l'information implicite Synthétiser à partir de la mise en relation d'indices

Tableau 3 – Définition des compétences en compréhension de l'écrit (évaluation 2016)

Compréhension de l'écrit **Objectifs d'évaluation**	
Identifier l'information pertinente	Reconnaître des expressions figées Reconnaître un lexique de la vie quotidienne Connaître des repères culturels Identifier le(s) thème(s) Repérer l'information explicite Identifier des repères spatiaux / temporels Identifier l'élément qui justifie une affirmation Identifier la situation
Construire le sens	Identifier l'information implicite Induire/déduire le sens d'une expression ou d'une phrase Retrouver l'ordre logique ou chronologique d'un texte Synthétiser

Tableau 4 – Définition des compétences en expression écrite (évaluation 2016)

Expression écrite **Objectifs d'évaluation**	
Ecrire des mots	Ecrire des mots ou expressions isolés
Ecrire des phrases	Ecrire des énoncés simples et brefs : - sur soi-même et des personnages imaginaires, où ils vivent et ce qu'ils font - sur sa famille, ses conditions de vie, son collège - sur les aspects quotidiens de son environnement, par exemple les gens, les lieux... Ecrire des phrases simples reliées par des connecteurs simples tels que « et », « mais » et « parce que »
Ecrire des textes	Ecrire des textes articulés et nuancés Faire une description brève et élémentaire d'un événement, d'activités passées et d'expériences personnelles

Tableau 5 – Définition des compétences en expression orale en continu (évaluation 2016)

Expression orale en continu **Objectifs d'évaluation**
Décrire Justifier un choix

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 chargé d'études et l'inspection pédagogique régionale. Un item proposé par un concepteur, pédagogue de terrain ayant une bonne connaissance des pratiques de classe, fait l'objet d'une discussion jusqu'à aboutir à un consensus, validé au final par le chargé d'études et l'inspection. L'item fait alors l'objet d'un cobayage, c'est-à-dire d'une passation auprès d'une ou plusieurs classes pour estimer sa difficulté et recueillir les réactions des élèves.

Un équilibre de proportion entre les items considérés comme étant "faciles", "moyennement faciles" ou "difficiles" est recherché (correspondant pour les langues vivantes étrangères, aux niveaux A1, A2 et tendant vers le niveau B1 du Cadre Européen). Deux formats de questions sont utilisés : questions à choix multiples (QCM) et questions ouvertes appelant une réponse écrite construite. Les questions dites ouvertes appellent des réponses sous forme de productions écrites. Elles supposent la mise en place d'un dispositif de correction experte à distance pour l'épreuve finale, nécessitant la formation technique des correcteurs et l'élaboration d'un cahier des charges strict de corrections, afin d'éviter toute subjectivité ou la validation de réponses trop imprécises ou succinctes. Une réponse rédigée à une question ouverte peut faire l'objet de plusieurs items qui couvrent les différentes compétences nécessaires pour répondre.

Les réponses au format QCM sont saisies de manière automatisée et les questions ouvertes corrigées par des experts via une interface Internet. Certaines questions, notamment celles constituant un ensemble de vrai/faux, sont regroupées afin qu'un item à deux modalités de réponse ne pèse pas autant qu'une question à quatre ou cinq propositions. Dans le cas de ces séries, des seuils statistiques sont établis pour valider les réponses des élèves. Pour l'évaluation de l'expression orale en continu, la formation d'un groupe de correcteurs et la rédaction d'un cahier des charges strict a également été requis pour là encore éviter toute subjectivité ou la validation de réponses trop imprécises ou succinctes.

Plusieurs items peuvent être regroupés dans "une situation". Cependant, ils restent indépendants les uns des autres. Les items au format QCM occupent la plus large part de l'évaluation-bilan. Une application ad hoc nommée GEODE est utilisée en interne pour faciliter la création des items, ainsi que leur édition, leur stockage et la gestion des évaluations (cf. encadré ci-dessous).

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

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 l'évaluation CEDRE anglais collège 2016, 13 blocs de compréhension de l'écrit et d'expression écrite sont constitués et répartis au sein de 13 cahiers. Pour la compréhension de l'oral, 6 blocs sont constitués et répartis dans 6 autres cahiers.

L'évaluation CEDRE anglais collège 2016 est constituée d'items de 2010 (représentant 50 % du nombre total d'items de l'évaluation) repris à l'identique afin d'assurer la comparabilité dans le temps et de nouveaux items qui ont fait l'objet d'une expérimentation en 2015.

Afin de pouvoir évaluer un nombre important d'items sans allonger le temps de passation pour l'élève, CEDRE utilise la méthodologie des cahiers tournants. Les items sont ainsi répartis dans des blocs d'une durée de 12 minutes et les blocs sont ensuite distribués dans les cahiers tout en respectant certaines contraintes : chaque bloc doit se retrouver un même nombre de fois au total et chaque association de blocs doit figurer au moins une fois dans un cahier. Ce dispositif, couramment utilisé dans les évaluations bilans, notamment les évaluations internationales, permet d'estimer la probabilité de réussite de chaque élève à chaque item sans que chaque élève ait à passer l'ensemble de ceux-ci.

Au final, pour l'évaluation CEDRE anglais collège 2016, chaque cahier comprend deux séquences de 50 minutes au cours desquelles l'élève est tout d'abord évalué en compréhension de l'oral (12 minutes), puis en compréhension de l'écrit et expression écrite (24 minutes). Les séances se terminent par un questionnaire de contexte (une première partie en fin de séance 1 et une deuxième partie en fin de séance 2), identique dans tous les cahiers, dans lequel l'élève doit répondre à des questions concernant notamment l'environnement familial dans lequel il évolue, ses projets scolaires et professionnels, sa perception de la matière et de son environnement scolaire.

1.4 Passation des évaluations

La passation de l'évaluation finale a eu lieu en mai 2016. Comme en 2010, cette évaluation a été précédée d'une expérimentation l'année N-1 de façon à tester un grand nombre d'items auprès d'un échantillon réduit d'établissements.

Dans chaque établissement, une personne a été désignée comme étant l'administrateur du test, son rôle étant de veiller au strict respect de la procédure à suivre afin que l'évaluation soit passée dans les meilleures conditions quel que soit l'établissement ; la collecte de l'information s'est faite par questionnaires " papier-crayon".

L'anonymat des élèves et des personnels est respecté, chaque cahier étant repéré par un numéro. Une fois l'évaluation terminée, les cahiers et questionnaires sont renvoyés dans des conditionnements prévus à cet effet, préaffranchis et pré-étiquetés. Aucun travail de correction n'a été demandé aux établissements.

2 Sondage

2.1 Méthodes

2.1.1 Tirage équilibré de classes de 3e

De manière générale, pour le secondaire, deux options de tirage peuvent être considérées : soit un sondage par grappe en sélectionnant un échantillon de classes et tous les élèves des classes tirées au sort participent à l'évaluation ; soit un premier degré qui concerne les établissements puis un second degré où un nombre d'élèves fixe dans chaque établissement est sélectionné [1]. Les évaluations CEDRE suivent la première option tandis que l'évaluation PISA suit la seconde. Des simulations ont permis de montrer que les niveaux de précision des deux options sont très proches, dès lors que le tirage est équilibré (cf. encadré « Tirage d'établissement *versus* tirage de classes »). Le choix de sondages par grappe est motivé par la facilité de gestion. En effet, le fait de sélectionner tous les élèves d'une classe au collège permet d'éviter de mettre en place des procédures de tirage au sort d'élèves une fois les établissements tirés.

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

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

Cette propriété s'écrit :

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

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

Tirage d'établissements *versus* Tirage de classes

Pour faciliter la logistique dans les collèges, nous réalisons un tirage de classes de 3e, puis tous les élèves de la classe sélectionnée passent l'évaluation. On peut donc s'interroger sur la perte de la précision liée à cet effet de grappe.
Pour comparer la précision entre un tirage d'établissement et un tirage de classes, nous avons réalisé des simulations à partir de la base des notes au brevet en 2009 (Garcia, Le Cam, & Rocher, 2015).
Nous avons comparé deux stratégies d'échantillonnage. Il s'agit à chaque fois d'échantillons stratifiés à deux degrés :

- Tirage équilibré d'établissement puis tirage de 30 élèves dans chaque établissement sélectionné ;
- Tirage équilibré de classe puis sélection de tous les élèves des classes sélectionnées.

La stratification a été effectuée selon le secteur d'enseignement et dans chaque strate 2 000 élèves ont été échantillonnés.
Pour chacune des deux stratégies, 1 000 échantillons ont été tirés. Puis on calcule la moyenne des erreurs standards des notes moyennes en français, mathématiques et histoire-géographie. Le tableau ci-dessous montre que les deux stratégies de tirage ont des niveaux équivalents de précision.

Comparaison des erreurs standards (Garcia et al., 2015)

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

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

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

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

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

2.1.2 Redressement de la non réponse : calage sur marges

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

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

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

Ils sont obtenus par minimisation de l'expression $\sum_{i \in S'} d_i G(\frac{w_i}{d_i})$ où G désigne une fonction de distance, sous les contraintes définies dans l'équation 2.

Tirage équilibré après élimination de la base des échantillons précédemment tirés

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

Formulation du problème et notations

Un échantillon S_1 a été tiré. Il est connu et les probabilités d'inclusion des établissements 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 $\lambda_j = \frac{\pi_j^2}{1-\pi_j^1}$

Équilibrage

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

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

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

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

Ainsi, on aura :

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

Or, en espérance on a

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

La condition d'équilibrage initiale est donc remplie.

Condition fondamentale

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

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

Dans certains cas, par exemple des strates souvent sur-représentées comme les établissements situés dans des zones spécifiques concernant peu d'élèves (ex : 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 évaluation CEDRE au collège est celui des élèves de 3e générale scolarisés dans des collèges publics et privés sous contrat de France métropolitaine.

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

Stratification

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

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

Modalités de sélection

Le tirage est à deux degrés. Le premier degré de sondage est composé de classes (et non de collèges) tirées dans chaque strate avec allocation proportionelle. Le deuxième degré de sondage consiste à interroger tous les élèves de la classe sélectionnée (tirage par grappe).La macro CUBE de l'INSEE est utilisée pour garantir des échantillons équilibrés sur la base de sondage selon certaines variables

Dans chacune des 3 strates, le tirage est équilibré sur les variables suivantes :

— Le nombre total d'élèves de 3e
— L'indice de position sociale (Rocher, 2016)
— Le nombre d'élèves de 3e en retard dans la population
— Le nombre de garçons de 3e dans la population

Echantillon 2016

L'échantillon vise 4 000 élèves répartis proportionnellement selon les trois strates.

Base de sondage

Le tableau 6 présente les exclusions dans la population ciblée.

Tableau 6 – Exclusions pour la base de sondage - CEDRE 2016 Anglais Collège

	Établissements	**Elèves**
Etab. accueillant des élèves de 3e	8 419	844 891
On retire les TOM	8 382	840 616
On retire les étab hors contrat	8 224	838 315
On retire les EREA	8 154	836 976
On retire les UPE2A	8 142	834 347
On retire les ULIS	8 126	832 052
On retire les DOM	7 870	792 243
On ne garde que les collèges	6 694	762 609
On ne garde que les 3ème générales	6 692	738 636
On garde les élèves ANG LV1 des classes >= 10	6 670	702 563
On retire Socle 3ème et Cedre All.	5 938	617 236
Base de tirage CEDRE Ang. 3e	**5 938**	**617 236**

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

Tableau 7 – Répartition dans la base de sondage - CEDRE 2016 Anglais Collège

Strate	**Établissements**	**Élèves**
1. Public hors EP	4 097	453 065
2. EP	961	96 199
3. Privé	1 612	153 299
Total	**6 670**	**702 563**

Échantillon

Le tableau 8 présente la répartition de l'échantillon selon le secteur d'enseignement. Au total, 164 écoles ont été sélectionnées.

Tableau 8 – Répartition dans l'échantillon - CEDRE 2016 Anglais Collège

Strate	**Établissements**	**Élèves**
1. Public hors EP	105	2 597
2. EP	25	557
3. Privé	34	874
Total	**164**	**4 028**

2.3 État des lieux de la non-réponse

2.3.1 Non-réponse totale

Parmi la non-réponse totale, nous distinguons la non-réponse des établissements de la non-réponse des élèves des établissements participants. Les chiffres suivants ont été observés pour 2016.

97.6 % des établissements de l'échantillon ont répondu à l'évaluation (tableau 9). 88.2 % des éffectifs attendus ont participé (tableau 10).

Tableau 9 – Non-réponse des établissements - CEDRE 2016 Anglais Collège

Strate	Nb établissements attendus	Nb établissements répondants	% d'établissements répondants
1. Public hors EP	105	102	97.1 %
2. EP	25	25	100 %
3. Privé	34	33	97.1 %
Total	**164**	**160**	**97.6 %**

Tableau 10 – Non-réponse des élèves - CEDRE 2016 Anglais Collège

Strate	Nb élèves attendus	Nb élèves répondants	% d'élèves répondants
1. Public hors EP	2 597	2 307	88.8 %
2. EP	557	478	85.8 %
3. Privé	874	767	87.8 %
Total	**4 028**	**3 552**	**88.2 %**

2.3.2 Valeurs manquantes et imputation

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

2.3.3 Non-réponse partielle et terminale

Lorsque des non-réponses sont observées aux items, nous distinguons les cas suivants :

— La non-réponse partielle : un élève n'a pas répondu à certains items dans le cahier.
— La non-réponse terminale : un élève s'est arrêté avant la fin du cahier soit par manque de temps soit par abandon.

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

2.3.3.a Compréhension de l'écrit et expression écrite

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

— 2.5 items pour la séquence 1
— 2.6 items pour la séquence 2

On considère que :

— 179 élèves n'ont pas vu la séquence 1, dont :
  — 160 n'ont répondu à aucun items de la séquence
  — 19 ont répondu à moins de 50 % de la séquence
— 211 élèves n'ont pas vu la séquence 2, dont :
  — 189 n'ont répondu à aucun items de la séquence
  — 22 ont répondu à moins de 50 % de la séquence

2.3.3.b Compréhension de l'oral

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

— 2.4 items pour la séquence 1
— 2.6 items pour la séquence 2

On considère que :

— 183 élèves n'ont pas vu la séquence 1, dont :
  — 181 n'ont répondu à aucun items de la séquence
  — 2 ont répondu à moins de 50 % de la séquence
— 176 élèves n'ont pas vu la séquence 2, dont :
  — 174 n'ont répondu à aucun items de la séquence
  — 2 ont répondu à moins de 50 % de la séquence

Les élèves dont toutes les séquences sont codées en "s" sont classés en non réponse totale. C'est le cas pour 42 élèves.

2.4 Redressement

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

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

Tableau 11 – Comparaison entre les marges de l'échantillon et les marges dans la population : Compréhension de l'écrit et expression écrite - CEDRE 2016 Anglais Collège

Modalité	Variable	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population
Retard	1	105 157.28	118 106	14.97	16.81
	2	597 405.72	584 457	85.03	83.19
Sexe	1	353 211.8	351 164	50.27	49.98
	2	349 351.2	351 399	49.73	50.02
Strate	1	453 064.95	453 065	64.49	64.49
	2	96 199	96 199	13.69	13.69

Tableau 12 – Comparaison entre les marges de l'échantillon et les marges dans la population : Compréhension de l'oral - CEDRE 2016 Anglais Collège

Modalité	Variable	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population
Retard	1	104 479.97	118 106	14.87	16.81
	2	598 083.11	584 457	85.13	83.19
Sexe	1	352 521.27	351 164	50.18	49.98
	2	350 041.81	351 399	49.82	50.02
Strate	1	453 065.05	453 065	64.49	64.49
	2	96 199	96 199	13.69	13.69

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 2016 Anglais Collège

	Année	Score moyen	Erreur standard
Compréhension de l'écrit	2004	250	1.64
	2010	252.2	1.99
	2016	278.3	2.4
Expression écrite	2004	250	1.67
	2010	238.3	1.57
	2016	235.4	1.68
Compréhension de l'oral	2004	250	1.97
	2010	240.1	1.82
	2016	255.9	2.12

Pour savoir par exemple si l'évolution entre 2010 et 2016 est significative , il faut calculer la valeur suivante :

$$\frac{\left|\hat{\bar{Y}}_{2016} - \hat{\bar{Y}}_{2010}\right|}{\sqrt{se^2_{\hat{\bar{Y}}_{2016}} + se^2_{\hat{\bar{Y}}_{2010}}}} \tag{3}$$

Compréhension de l'écrit : entre 2010 et 2016, on obtient une valeur de 8.35 (supérieure à 1.96). Cela signifie que l'évolution du score moyen est statistiquement significative.
Expression écrite : l'évolution du score entre 2010 et 2016 n'est pas significative (1.23).
Compréhension de l'oral : l'évolution du score entre 2010 et 2016 est significative (5.64).

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

Tableau 14 – Répartitions en % dans les groupes de niveaux - CEDRE 2016 Anglais Collège

	Année	Groupe <1	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5
Compréhension de l'écrit	2004	2.9	12.1	31.2	28.2	15.6	10
	2010	7	14.4	26.2	23.2	14.7	14.6
	2016	6	9.7	16.2	20.5	17.2	30.5
Expression écrite	2004	5.5	9.5	20.5	30.5	24	10
	2010	8.2	14	21	32.3	18.6	5.8
	2016	9.9	16.9	26.1	23.3	14.4	9.4
Compréhension de l'oral	2004	1.6	13.4	32.1	28.4	14.5	10
	2010	4.8	14.8	34.2	28.4	11.5	6.2
	2016	2.4	12.2	28	26.9	15.6	14.9

Tableau 15 – Erreurs standards des répartitions en % dans les groupes de niveaux - CEDRE 2016 Anglais Collège

	Année	Groupe <1	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5
Compréhension de l'écrit	2004	0.3	0.8	1	0.8	0.8	0.8
	2010	0.5	0.8	0.9	0.8	0.7	1
	2016	0.5	0.6	0.8	0.8	0.6	1.4
Expression écrite	2004	0.5	0.6	0.8	0.8	0.9	0.8
	2010	0.6	0.7	0.7	0.9	1	0.6
	2016	0.6	0.7	0.9	0.8	0.8	0.9
Compréhension de l'oral	2004	0.3	0.9	1.1	0.9	0.7	0.9
	2010	0.6	0.8	1.1	1	0.8	0.7
	2016	0.3	0.9	1.2	0.9	0.8	1.2

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 :

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

Tableau 16 – Effet du plan de sondage - CEDRE 2016 Anglais Collège

Année	Erreur Standard	Erreur SAS	*Design Effect*
2004	1.64	0.69	2.39
2010	1.99	0.96	2.08
2016	2.4	1.24	1.94

Dans le cas d'un sondage en grappes, la précision est dégradée en comparaison d'un sondage aléatoire simple. Cela signifie qu'en 2016, un sondage aléatoire simple avec un effectif 1.94 fois moins important aurait conduit au même niveau de précision.

3 Analyse des items

3.1 Méthodologie

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

3.1.1 Approche classique

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

Réussite et score

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

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

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

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

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

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

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

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

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

Fidélité

Dans le cadre de la théorie classique des tests, la fidélité (*reliability*) est définie comme la corrélation entre le score observé et le score vrai : le test est fidèle, lorsque l'erreur de mesure est réduite. Une manière d'estimer cette erreur de mesure consiste par exemple à calculer les corrélations entre les différents sous-scores possibles : plus ces corrélations sont élevées, plus le test est dit fidèle [2].

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 :

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

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

De ce point de vue, cet indicateur renseigne sur la consistance interne du test. En pratique, une valeur supérieure à 0,8 témoigne d'une bonne fidélité [3].

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.

2. 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.

3. 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

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

où $\bar{S}_{(j1)}$ est le score moyen sur l'ensemble de l'évaluation des élèves ayant réussi l'item j, $\bar{S}_{(j0)}$ celui des élèves l'ayant échoué et 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) [4]. 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.

4. 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 :

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

La réponse y_{ij} de l'élève i à l'item j est incorrecte tant que la variable latente Z_j reste en deça d'un certain seuil 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.

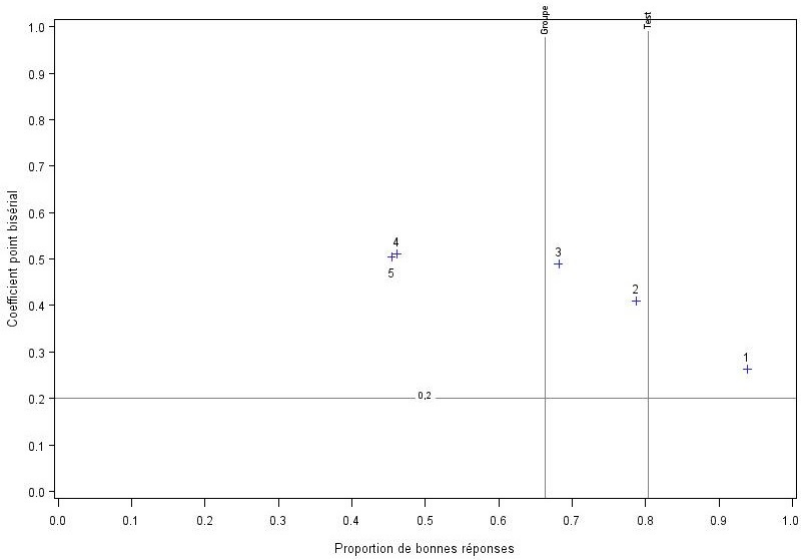
— Non-réponse terminale : l'élève s'est arrêté au cours de l'épreuve, potentiellement en raison d'un manque de temps. Des choix sont effectués pour déterminer le traitement de ces valeurs. Nous considérons que si un élève a passé moins de 50 % d'une séquence, il n'a pas vu la séquence, les valeurs manquantes 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 précises et d'autre part à corriger collectivement à blanc un nombre défini d'imagettes pour s'assurer de la compréhension et de la bonne mise en oeuvre des consignes. Puis, pendant la télécorrection, il arbitre les litiges lors des doubles-corrections. Le correcteur corrige les items en portant un codage de réussite/erreur sur chaque item. En cas de doute, il peut se référer à son superviseur de groupe. Une messagerie interne complète le dispositif et permet un échange de point de vue entre les différents acteurs.

3.3 Résultats

3.3.1 Pouvoir discriminant des items

Compréhension de l'écrit

Le calcul des indices de discrimination conduit à éliminer 5 items dont les indices *rbis-point* sont trop faibles :

- 2 items de 2004
- 1 item commun à 2004 et 2010
- 2 items communs à 2010 et 2016

Expression écrite

Le calcul des indices de discrimination conduit à éliminer 8 items dont les indices *rbis-point* sont trop faibles :

- 6 items de 2004
- 1 item commun à 2004 et 2010
- 1 item de 2010

Compréhension de l'oral

Le calcul des indices de discrimination conduit à éliminer 13 items dont les indices *rbis-point* sont trop faibles :

- 6 items de 2004
- 1 item de 2010
- 6 items de 2016

4 Modélisation

4.1 Méthodologie

4.1.1 Modèle de réponse à l'item

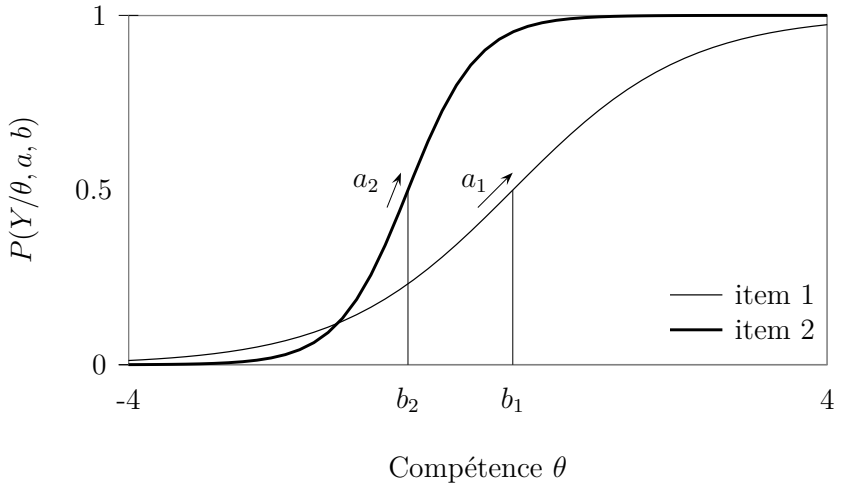
Le modèle de mesure utilisé est un modèle de réponse à l'item à deux paramètres avec une fonction de lien logistique (MRI 2PL) :

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

où la probabilité P_{ij} que l'élève i réussisse l'item j est fonction du niveau de compétence 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 (11) qui formule la probabilité P_{ij} d'un élève i de répondre correctement à un item j dans le cadre d'un modèle de réponse à l'item, avec les items sont dichotomiques.

Notons tout d'abord que les modèles présentés ne sont pas identifiables. En effet, les transformations $\theta_i^* = A\theta_i + B$, $b_j^* = Ab_j + B$ et $a_j^* = a_j/A$ avec A et B deux constantes $(A > 0)$, conduisent aux mêmes valeurs des probabilités. Dans CEDRE, nous levons l'indétermination en standardisant la distribution des θ pour les données du premier cycle (en l'occurrence, moyenne de 250 et écart-type de 50 pour l'année 2004).

Sous l'hypothèse d'indépendance locale des items [5], la fonction de vraisemblance s'écrit :

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

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

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

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

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

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

avec g la densité de θ.

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

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

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

La résolution de ces équations est classiquement réalisée grâce à l'algorithme EM (*Expectation-Maximization*) impliquant des approximations d'intégrales par points de quadrature. L'algorithme EM est théoriquement adapté dans le cas de valeurs manquantes. Le principe général est de calculer l'espérance conditionnelle de la vraisemblance des données complètes (incluant les valeurs manquantes) avec les valeurs des paramètres estimées à l'étape précédente, puis de maximiser cette espérance conditionnelle pour trouver les nouvelles valeurs des paramètres. Le calcul de l'espérance conditionnelle nécessite cependant de connaître (ou de supposer) la loi jointe des données complètes. Une version modifiée de l'algorithme considère dans notre cas le paramètre 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 (12).

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

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

avec

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

et

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

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

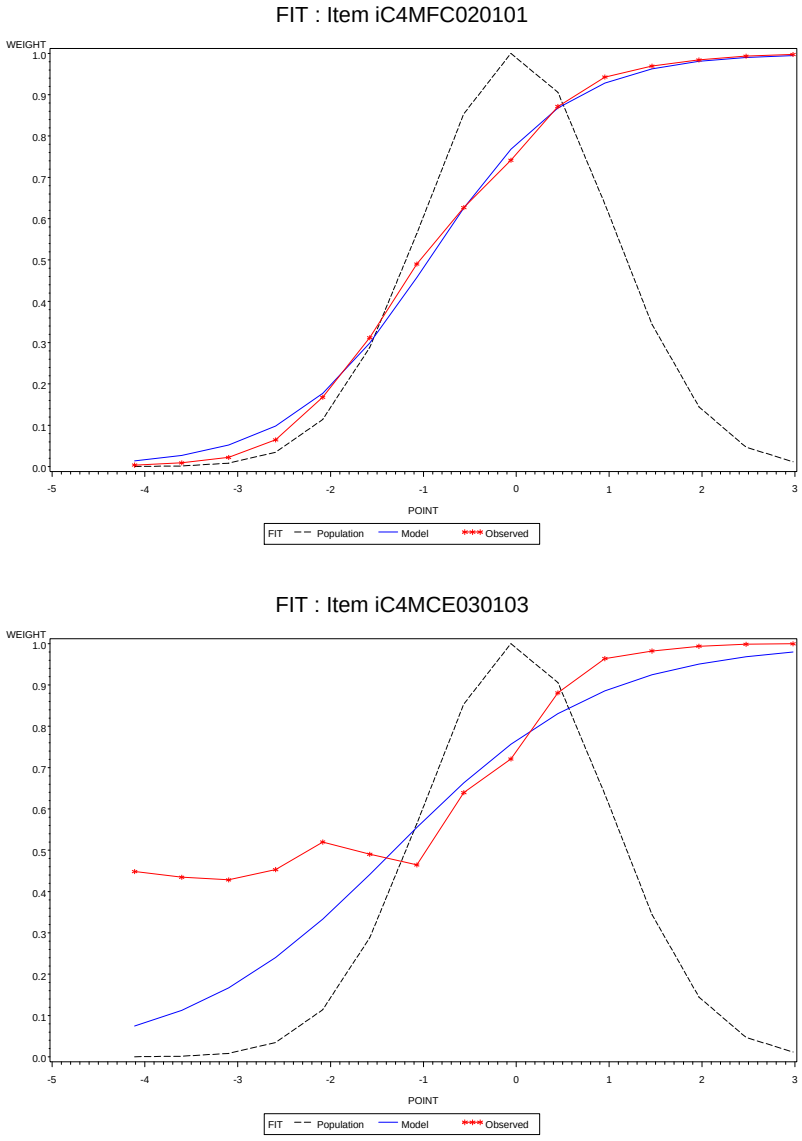
$$\frac{\partial lnL}{\partial \theta_i} + \frac{J}{2I} = 0 \tag{16}$$

4.1.3 Indice d'ajustement (FIT)

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

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

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 :

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

avec le poids $w_{ij} = P_{ij}(1 - P_{ij})$. Une transformation de cet indice est utilisé de manière à obtenir une statistique suivant approximativement 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 :

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

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

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 :

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

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

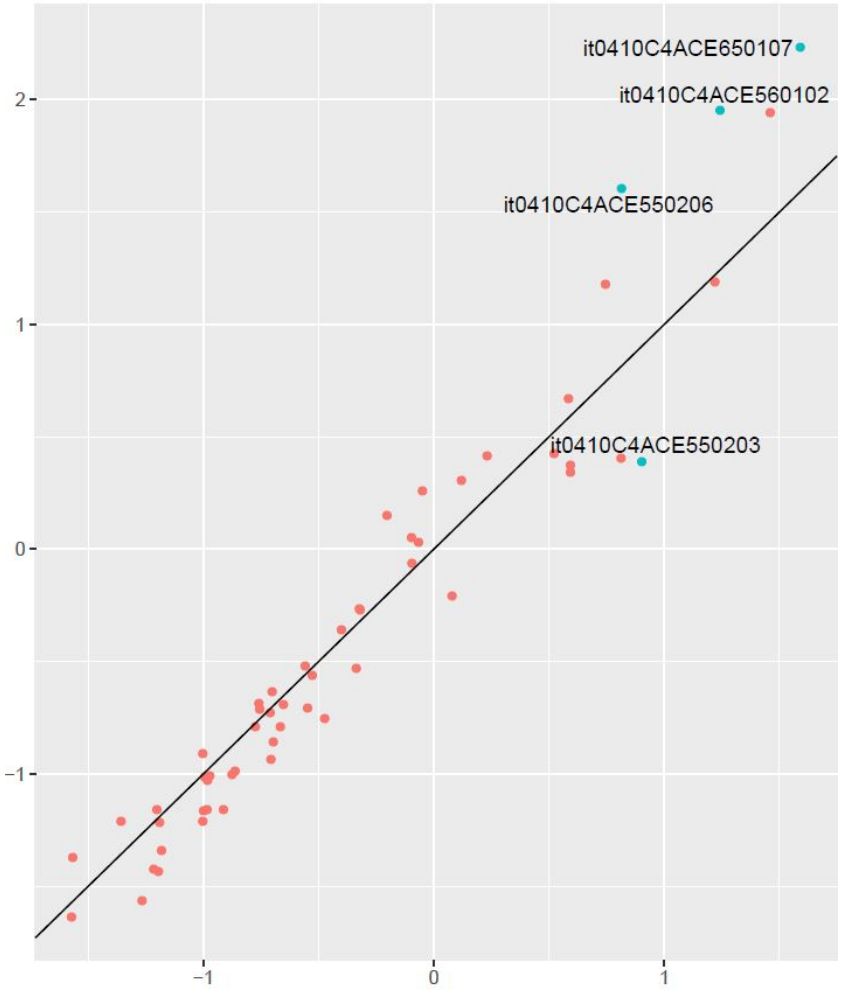
4.2.1 Identification des fonctionnements différentiels d'items (FDI)

Compréhension de l'écrit
6 items ont été éliminés des calculs :
- 4 items pour 2004-2010
- 2 items pour 2010-2016

Expression écrite
6 items ont été éliminés des calculs :
- 1 item pour 2004-2010
- 5 items pour 2010-2016

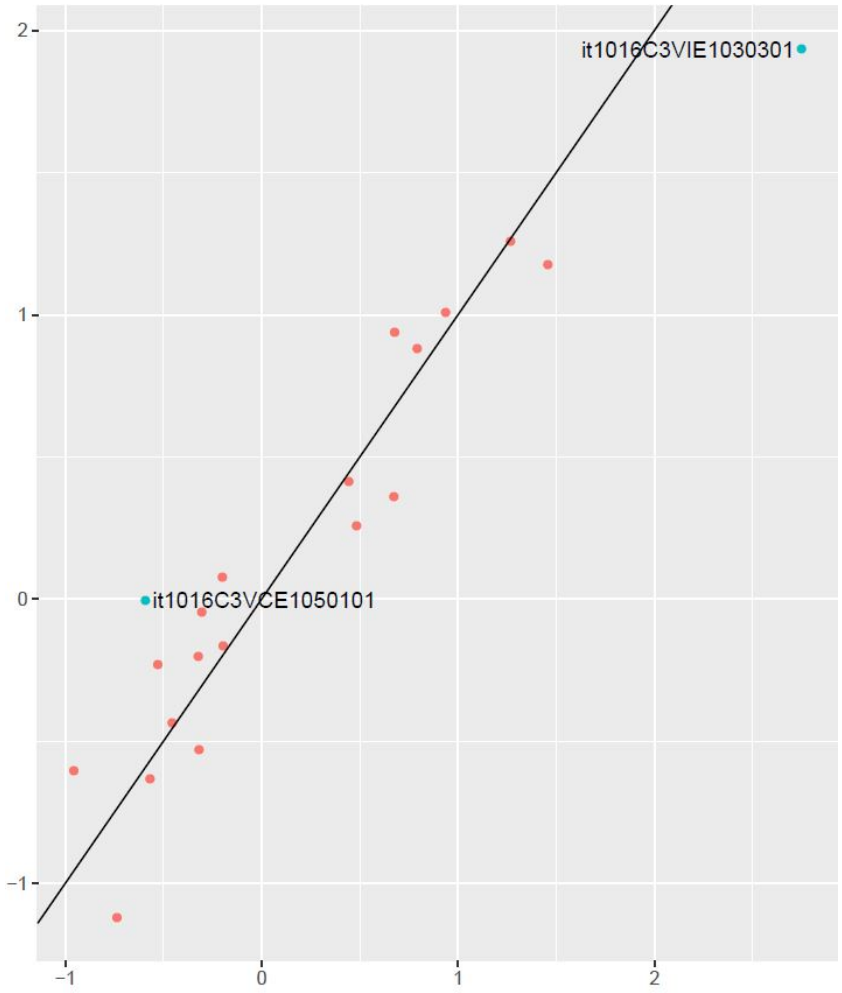
Compréhension de l'oral
10 items ont été éliminés des calculs :
- 4 items pour 2004-2010
- 6 items pour 2010-2016

Figure 4 – Comparaison des paramètres de difficulté 2004-2010 : Compréhension de l'écrit - (CEDRE Anglais 2016 Collège)



Note de lecture : Les points sont les items. En abscisse figure la valeur des paramètres de difficulté estimés en 2004, et en ordonnée la la valeur des paramètres de difficulté estimés et ajustés pour l'année 2010. Les items présentant un FDI apparaissent en bleu.

Figure 5 – Comparaison des paramètres de difficulté 2010-2016 : Compréhension de l'écrit - (CEDRE Anglais 2016 Collège)



Note de lecture : Les points sont les items. En abscisse figure la valeur des paramètres de difficulté estimés en 2010, et en ordonnée la la valeur des paramètres de difficulté estimés et ajustés pour l'année 2016. Les items présentant un FDI apparaissent en bleu.

4.2.2 Identification des items présentant un mauvais ajustement (FIT)

Aucun item présentant un mauvais ajustement n'a été détecté.

4.2.3 Bilan de l'analyse des items

Compréhension de l'écrit

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

- 106 items de 2004
- 4 items de 2010
- 14 items de 2016
- 57 items d'ancrage 2004-2010
- 22 items d'ancrage 2010-2016

Cela représente 203 items passés par les élèves en tout, dont 36 en 2016.

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

- 104 items de 2004
- 4 items de 2010
- 14 items de 2016
- 52 items d'ancrage 2004-2010
- 18 items d'ancrage 2010-2016

192 items sont donc conservés dans l'analyse, dont 32 utilisés dans l'évaluation 2016.

Expression écrite

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

- 140 items de 2004
- 89 items de 2010
- 13 items de 2016
- 13 items d'ancrage 2004-2010
- 16 items d'ancrage 2010-2016

Cela représente 271 items passés par les élèves en tout, dont 29 en 2016.

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

- 134 items de 2004
- 88 items de 2010
- 13 items de 2016
- 11 items d'ancrage 2004-2010
- 11 items d'ancrage 2010-2016

257 items sont donc conservés dans l'analyse, dont 24 utilisés dans l'évaluation 2016.

Compréhension de l'oral

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

- 39 items de 2004

— 4 items de 2010
— 40 items de 2016
— 40 items d'ancrage 2004-2010
— 37 items d'ancrage 2010-2016

Cela représente 160 items passés par les élèves en tout, dont 77 en 2016.

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

— 33 items de 2004
— 3 items de 2010
— 34 items de 2016
— 36 items d'ancrage 2004-2010
— 31 items d'ancrage 2010-2016

137 items sont donc conservés dans l'analyse, dont 65 utilisés dans l'évaluation 2016.

4.3 Calcul des scores

Comme indiqué précédemment, une analyse conjointe des données des 3 années a permis d'estimer les paramètres des items, puis les niveaux de compétences θ des élèves. Afin de lever l'indétermination du modèle, la moyenne des θ a été fixé à 250 et leur écart-type à 50, pour l'échantillon de 2004. Le tableau 17 présente les résultats obtenus.

Tableau 17 – Niveaux de compétences (moyennes des scores et écarts-types) - CEDRE 2016 Anglais Collège

	Année	Score moyen	Écart-type
Compréhension de l'écrit	2004	250	50
	2010	252.2	60.9
	2016	278.3	74.1
Expression écrite	2004	250	50
	2010	238.3	48.7
	2016	235.4	49.7
Compréhension de l'oral	2004	250	50
	2010	240.1	50.1
	2016	255.9	54.8

5 Construction de l'échelle

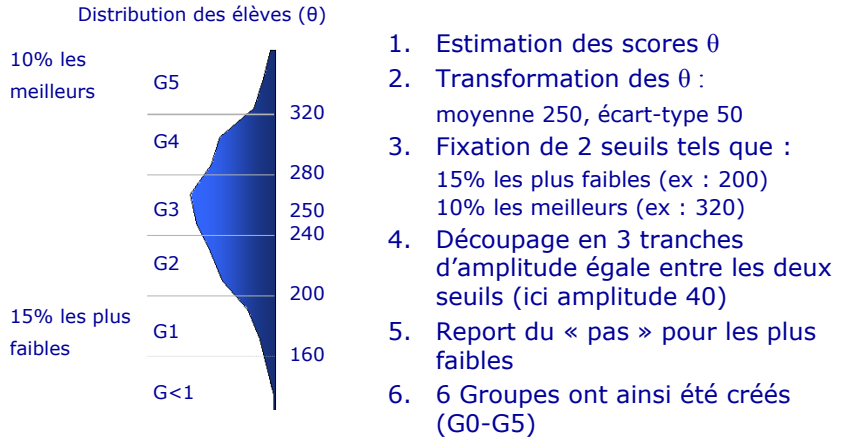
5.1 Méthode

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 anglais 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 2004. Puis, comme le montre la figure 6, la distribution des scores est « découpée » en six groupes de la manière suivante : nous déterminons le score-seuil en-deça duquel se situent 15 % des élèves (groupes < 1 et 1), nous déterminons le score-seuil au-delà duquel se situent 10 % des élèves (groupe 5). Entre ces deux niveaux, l'échelle a été scindée en trois parties d'amplitudes de scores égales correspondant à trois groupes intermédiaires. Ces choix sont arbitraires et ont pour objectif de décrire plus précisément le continuum de compétence.

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

Figure 6 – Principes de construction de l'échelle



5.2 Caractérisation des groupes de niveaux

A partir de cette correspondance entre les items et les groupes, une description qualitative et synthétique des compétences maitrisées par les élèves des différents groupes est proposée.

5.2.1 Compréhension de l'oral

Groupe < 1 (2,4 % des élèves)

Bien que capables de répondre ponctuellement à quelques questions, les élèves ne maîtrisent quasiment aucune des compétences attendues en fin de troisième. Ils savent reconnaître des données chiffrées simples et ils ont quelques acquis lexicaux simples. Même si ces derniers représentent un facteur capital d'ancrage, ils ne constituent néanmoins pas un socle suffisant permettant d'accéder au sens.

Groupe 1 (12,2 % des élèves)

Dans un message oral, les élèves savent repérer des expressions et un lexique très courant de la vie quotidienne et/ou un lexique transparent. Ils maîtrisent une partie de la " langue de fonctionnement " (anglais de classe) et comprennent des consignes simples à l'oral. Ils peuvent repérer quelques éléments simples comme des nombres. Leurs acquis lexicaux sont encore peu étendus.

Groupe 2 (28 % des élèves)

Les élèves de ce groupe peuvent comprendre quelques points d'une conversation sur un sujet familier ou le sujet principal d'une conversation. Ils savent repérer une information explicite en s'appuyant sur des éléments lexicaux simples et dans des contextes plus variés.

Groupe 3 (26,9 % des élèves)

Dans un message oral, ces élèves savent repérer un lexique plus étendu et plus riche même quand le débit est plus rapide. Ils peuvent comprendre les éléments clés d'une situation et ils commencent à pouvoir inférer à partir d'éléments explicites, en mettant en relation certaines informations.

Groupe 4 (15,6 % des élèves)

Dans un message oral, les élèves de ce groupe savent mettre en relation des informations explicites pour accéder à du sens même si le contexte est moins familier. Ils savent garder en mémoire les expressions et les mots porteurs de sens pour opérer des déductions, même quand le débit est plus rapide. Ils peuvent construire l'information en mettant en relation des indices pour accéder au sens de façon satisfaisante. Et ils commencent à accéder au message implicite.

Groupe 5 (14,9 % des élèves)

Les élèves du groupe 5 ont une bonne maîtrise des compétences même complexes : lls savent exploiter les informations présentes dans un document sonore et ont suffisamment d'acquis pour pouvoir relever des indices et les traiter de manière pertinente afin d'accéder au sens. Ils peuvent traiter à la fois l'explicite et accéder à un certain degré d'implicite.

5.2.2 Compréhension de l'écrit

Groupe $<$ 1 (6 % des élèves)

Bien que capables de répondre ponctuellement à quelques questions, les élèves ne maîtrisent quasiment aucune des compétences attendues en fin de troisième.

Groupe 1 (9,7 % des élèves)

Les élèves ont une connaissance limitée du lexique de base concernant leur environnement proche. Ils sont capables de reconnaître le genre de certains documents en s'appuyant sur un lexique transparent ou sur une information explicite très facilement identifiable.

Groupe 2 (16,2 % des élèves)

Les élèves ont une certaine connaissance du lexique simple relatif à leur environnement proche (la description physique, la maison, les animaux, la nourriture). Ils sont capables de retrouver une information explicite facilement identifiable.

Groupe 3 (20,5 % des élèves)

Les élèves de ce groupe discriminent assez bien les informations nécessaires pour accomplir une tâche donnée dans des supports variés, relatifs à la vie quotidienne (menus de restaurant, horaires de train, sites internet, articles de journaux) et ils maîtrisent les compétences de repérage, d'identification d'éléments informatifs explicites. Ils commencent à accéder à des informations implicites et à mettre en relation des indices prélevés dans un support écrit.

Groupe 4 (17,2 % des élèves)

Les élèves ont un degré de connaissance certain du lexique courant, des expressions figées et des repères culturels minimaux. Ils repèrent aisément l'information explicite, qu'elle soit facilement identifiable ou non. Ils peuvent construire l'information en mettant en relation les indices pour accéder au sens de façon satisfaisante et ils accèdent au message implicite de manière assez satisfaisante.

Groupe 5 (30,5 % des élèves)

Les élèves du groupe 5 ont une bonne maîtrise des compétences même complexes : Ils savent exploiter les informations recueillies dans un document et ont suffisamment d'acquis pour pouvoir relever des indices et les traiter de manière pertinente pour accéder au sens. Ils peuvent traiter à la fois l'explicite et l'implicite et ils sont capables de synthétiser.

5.3 Exemples d'items

5.3.1 Compréhension de l'oral

5.3.1.a Item caractéristique du groupe < 1

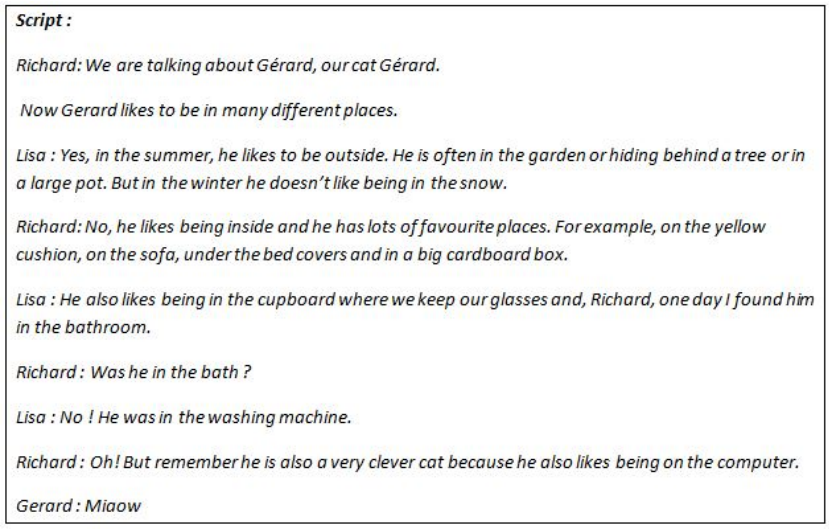
Cette situation évalue la maîtrise de la compétence " Repérer l'information explicite " et la tâche relève du niveau A1 du Cadre européen commun de référence pour les langues. Un document sonore simple, au débit lent, dont la thématique est proche de l'environnement immédiat des élèves et de leur sphère personnelle (les animaux domestiques, en l'occurrence ici un chat) est proposé aux élèves. Ils doivent comprendre les informations explicites et indiquer si les affirmations énoncées sont vraies ou fausses : le chat Gérard est souvent dans le jardin en été ; il aime jouer dans la neige l'hiver ; il aime s'assoir sur le canapé, ... Cette situation a été réussie par une très large majorité d'élèves, les taux globaux de réussite se situant entre 69,8 % et 95,5 %. L'item 1 *"In the summer he is often in the garden"*, réussi à 95,5 %, est accessible aux élèves du groupe <1, qui ont pu s'appuyer sur les énoncés *"in the summer, [...] He is often in the garden"*, facilement identifiables puisqu'identiques à l'oral et à l'écrit (dans la consigne), pour trouver la bonne réponse. Par comparaison, l'item 4 *"He hates the bed covers"*, réussi à 69,8 %, plus difficile d'un point de vue lexical, n'était pas à la portée du groupe <1 (avec des chances de réussite pour seulement 0,2 % des élèves) mais du groupe 3.

Figure 7 – Exemple d'item réussi par le groupe < à 1

Listen to Lisa and Richard talking about their cat Gerard.
First read the statements.
Now listen and tick RIGHT or WRONG.
>> *Listen again and check your answers.*

	RIGHT	WRONG
In the summer, he is often in the garden.	☐ 1	☐ 2
In the winter, he enjoys playing in the snow.	☐ 1	☐ 2
He likes sitting on the sofa.	☐ 1	☐ 2
He hates the bed covers.	☐ 1	☐ 2
He hates computers.	☐ 1	☐ 2

Figure 8 – Script associé à l'item



Script :

Richard: We are talking about Gérard, our cat Gérard.

Now Gerard likes to be in many different places.

Lisa : Yes, in the summer, he likes to be outside. He is often in the garden or hiding behind a tree or in a large pot. But in the winter he doesn't like being in the snow.

Richard: No, he likes being inside and he has lots of favourite places. For example, on the yellow cushion, on the sofa, under the bed covers and in a big cardboard box.

Lisa : He also likes being in the cupboard where we keep our glasses and, Richard, one day I found him in the bathroom.

Richard : Was he in the bath ?

Lisa : No ! He was in the washing machine.

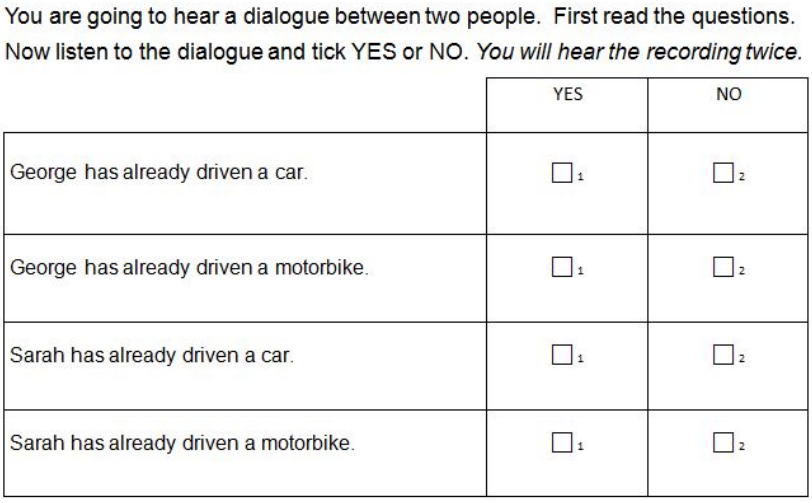
Richard : Oh! But remember he is also a very clever cat because he also likes being on the computer.

Gerard : Miaow

5.3.1.b Item caractéristique du groupe 1

Cette situation, qui évalue elle aussi la maîtrise de la compétence " Repérer l'information explicite ", est un peu différente dans la mesure où elle est plus éloignée de la sphère quotidienne des élèves de 3ème : il y est en effet question d'adultes ayant une conversation au sujet de la conduite d'une voiture, notamment. Le niveau lexical et de langue d'une manière générale reste très simple et accessible, ce qui explique sans doute les forts taux de réussite, se situant entre 90 et 94,5 %. Seuls les items 3 et 4 étaient cependant à la portée des élèves du groupe 1 ; les deux premiers items étant accessibles aux seuls élèves du groupe 2. Dans cette conversation, les personnages échangent de manière spontanée, utilisant des réponses brèves (*"yes","no", "I have"*) sur un rythme assez rapide, ce qui différencie cette situation de la précédente, présentée pour le groupe <1.

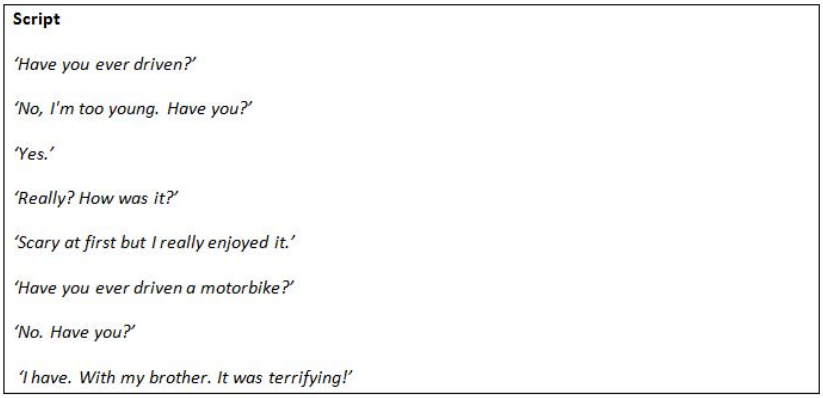
Figure 9 – Exemple d'item réussi par le groupe 1



You are going to hear a dialogue between two people. First read the questions.
Now listen to the dialogue and tick YES or NO. *You will hear the recording twice.*

	YES	NO
George has already driven a car.	☐ 1	☐ 2
George has already driven a motorbike.	☐ 1	☐ 2
Sarah has already driven a car.	☐ 1	☐ 2
Sarah has already driven a motorbike.	☐ 1	☐ 2

Figure 10 – Script associé à l'item



Script

'Have you ever driven?'

'No, I'm too young. Have you?'

'Yes.'

'Really? How was it?'

'Scary at first but I really enjoyed it.'

'Have you ever driven a motorbike?'

'No. Have you?'

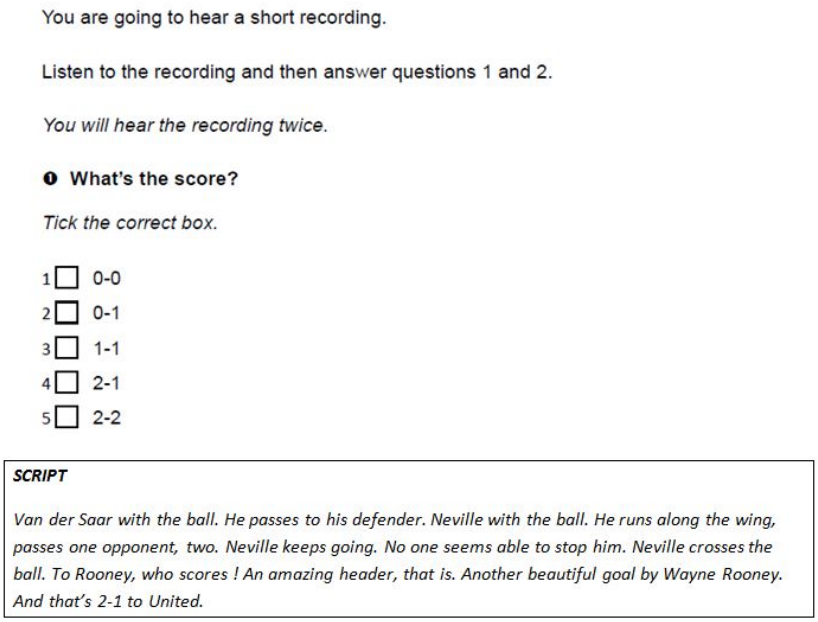
'I have. With my brother. It was terrifying!'

5.3.1.c Item caractéristique du groupe 2

Dans cette situation, les élèves entendent un commentaire sportif - un match de football - et doivent identifier le score final. L'objectif d'évaluation visé dans cet item (repérer l'information explicite : le score du match) est atteint par une grande majorité des élèves (84,8 %) et ce, dès le groupe 2. Le repérage de l'information essentielle était : le score du match et les élèves devaient être capables de repérer les chiffres (le nombre de buts marqués) dans le flux du continuum sonore et de les isoler comme éléments pertinents. Ce résultat souligne la capa-

cité, même chez les élèves les moins habiles, à effectuer des repérages, dès lors que l'information est explicite. Le contexte spécifique du document (le monde du sport et plus particulièrement l'univers du football) n'a semble-t-il pas constitué un écueil. La réussite massive à cet item peut s'expliquer par l'intérêt suscité par l'authenticité de la situation d'écoute, proche du réel : écouter la radio pour connaître le score d'un match.

Figure 11 – Exemple d'item réussi par le groupe 2


You are going to hear a short recording.

Listen to the recording and then answer questions 1 and 2.

You will hear the recording twice.

What's the score?

Tick the correct box.

1 ☐ 0-0
2 ☐ 0-1
3 ☐ 1-1
4 ☐ 2-1
5 ☐ 2-2

SCRIPT

Van der Saar with the ball. He passes to his defender. Neville with the ball. He runs along the wing, passes one opponent, two. Neville keeps going. No one seems able to stop him. Neville crosses the ball. To Rooney, who scores ! An amazing header, that is. Another beautiful goal by Wayne Rooney. And that's 2-1 to United.

5.3.1.d Item caractéristique du groupe 3

Voici un exemple d'item réussi dès le groupe 3, c'est-à-dire le groupe médian de l'échelle de performances. Dans cette situation, dont le support est une vidéo présentant un camp d'été pour jeunes en Angleterre, les élèves doivent comprendre à quelle tranche d'âge le programme Buckykid est destiné. Seul l'item 2 relevait du groupe 3 et il est réussi à 66 % par les élèves de ce groupe de performances. La compétence mise en oeuvre ici pour parvenir à la bonne réponse est : " inférer à partir d'éléments explicites ", une compétence plus difficile à maîtriser que le seul repérage d'un âge par exemple. Les items 1 et 3, plus faciles, impliquant d'effectuer un prélèvement simple, relèvent de groupes moins habiles : les groupes 1 et 2.

Figure 12 – Exemple d'item réussi par le groupe 3

You are preparing your next holidays.

Watch the video on a summer camp program in England.

First read the questions.

Watch the video.

>> *You will see it twice.*

Tick RIGHT or WRONG.

	RIGHT	WRONG
The school is based on two sites.	☐ 1	☐ 2
Buckykid program is for 7 to 15 year-olds.	☐ 1	☐ 2
There is a magicians' theme day.	☐ 1	☐ 2
You go on excursions two days a week.	☐ 1	☐ 2
You study English 12 hours a week.	☐ 1	☐ 2

Figure 13 – Script associé à l'item

SCRIPT

Summer programs are based in two campus environments, one is King Edward's School and the other is Plumpton College. We have a range of summer programs in these two centres beginning with Buckykid which is in King Edward's campus which is for seven to ten year-olds. This program combines theme days such as magic and magicians day and detectives day and princesses and monsters day with the English lessons so that the two work together. They have a full theme day.

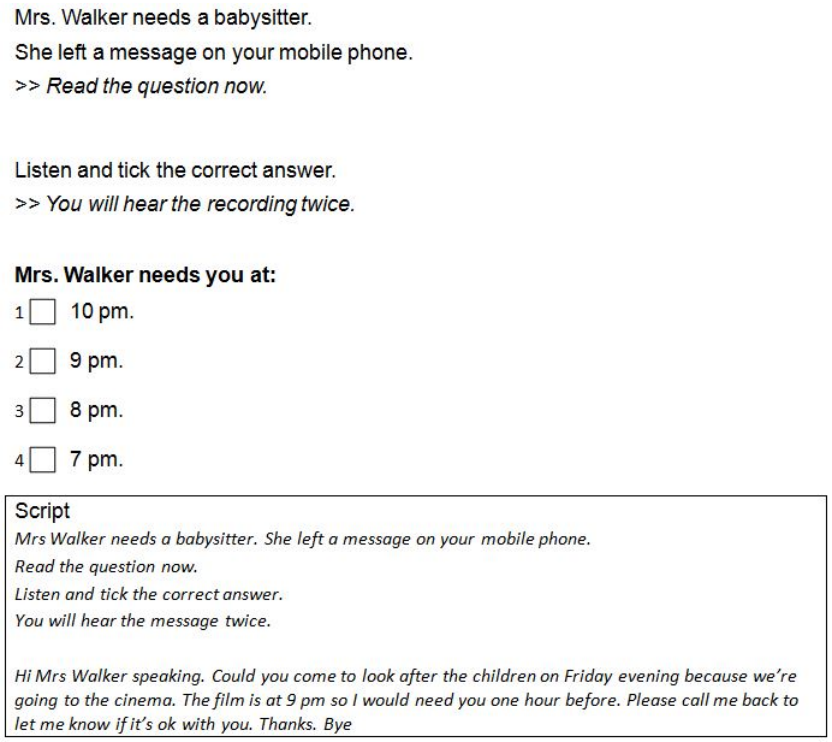
We then have Bucksmore summer programs in both King Edward and Plumpton which is English lessons and activities and excursions. They have two full-day excursions a week and two half-day excursions and they have twenty day, twenty hours of English lessons each week.

5.3.1.e Item caractéristique du groupe 4

Cette situation présentée pour le groupe 4 évalue, comme dans l'exemple précédent, la maîtrise de la compétence " inférer à partir d'éléments explicites ". Après avoir écouté un court message téléphonique dans lequel une mère de famille sollicite les services d'une baby-sitter, les élèves doivent trouver l'heure exacte à laquelle la jeune fille est attendue. Or, pour parvenir à la réponse, ils doivent s'appuyer sur ce qui est dit (*" the film is at 9pm "* et *" I would need you one hour before "*) et déduire l'information demandée, tout en faisant du sens au sein de la chaîne sonore. Cette situation est réussie à 49,9 % dans son

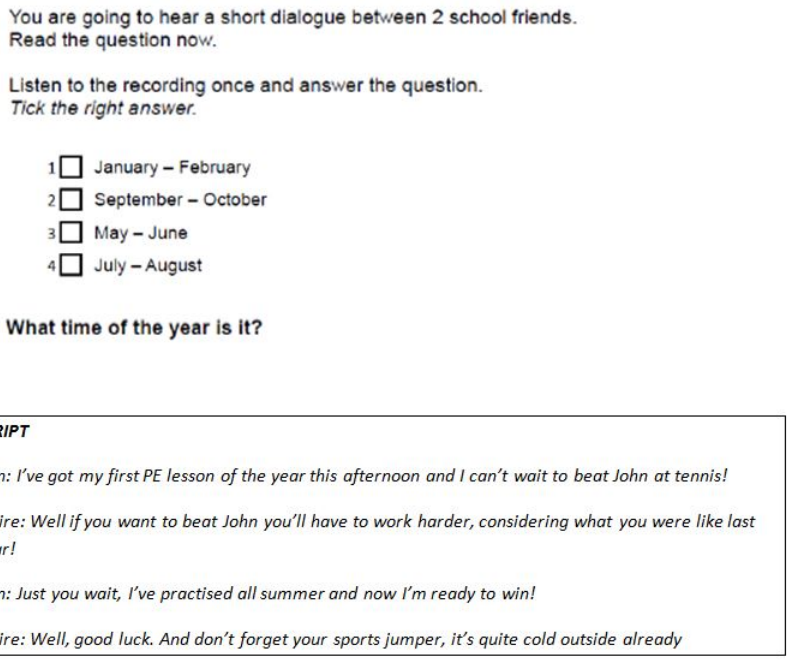
ensemble, donc par à peine un élève sur deux. Le support présentait pourtant peu de difficultés d'ordre lexical. La réponse 2, choisie par 46,2 % des élèves, souligne la complexité de la construction du sens, qui ne pouvait s'opérer avec un seul repérage de surface (*" the film is at 9pm "*). Il s'agissait en effet d'être capable d'inférer à partir des éléments explicites, présents dans le document : le film étant à 21 heures, la mère de famille avait besoin de la jeune fille une heure avant, soit à 20 heures.

Figure 14 – Exemple d'item réussi par le groupe 4



5.3.1.f Item caractéristique du groupe 5

Figure 15 – Exemple d'item réussi par le groupe 5



Dans cette situation, les élèves entendent un bref échange (ce qui contribue à la difficulté de cet item) entre deux camarades de classe. Ils doivent situer la scène dans l'année, selon les indices qu'ils auront recueillis : janvier-février, septembre-octobre, mai-juin ou juillet-aout. On évalue ici l'aptitude à déduire, à inférer à partir de ce qui est dit. Cette opération relève de la compétence " traiter l'information ". La difficulté réside dans un premier temps dans le relevé des indices présents dans le message, puis dans leur traitement : établir des liens entre les indices et la question posée. Les élèves doivent donc s'appuyer sur ce qui est dit (*" my first PE lesson of the year, last year, all summer, cold "*), l'interpréter pour aboutir à la conclusion logique que la scène se déroule en début d'année scolaire (septembre ou octobre). Les élèves doivent faire sens avec ce qui est dit pour comprendre ce qui n'est pas dit. Cette compétence de haut niveau commence à être maîtrisée par les élèves du haut de l'échelle (item réussi par le groupe 5 à 73 %), mais demeure encore hors de portée des élèves des groupes inférieurs. Le taux de réussite globale est faible : 32 %. Il est intéressant de relever que presqu'un tiers des élèves choisissent la réponse 4 : *"July-August"*. Ces élèves ont identifié correctement l'indication temporelle

"all summer", sans toutefois la mettre en relation avec *"now"*, qui marque une rupture. On peut dire que cette situation ne permettait pas aux élèves, capables de ne repérer qu'un seul indice, de trouver la réponse. Elle illustre la maîtrise des compétences dites " de haut niveau " dont font preuve les élèves à partir du groupe 5 seulement.

5.3.2 Compréhension de l'écrit

5.3.2.a Item caractéristique du groupe 2

Figure 16 – Exemple d'iem réussi par le groupe 2

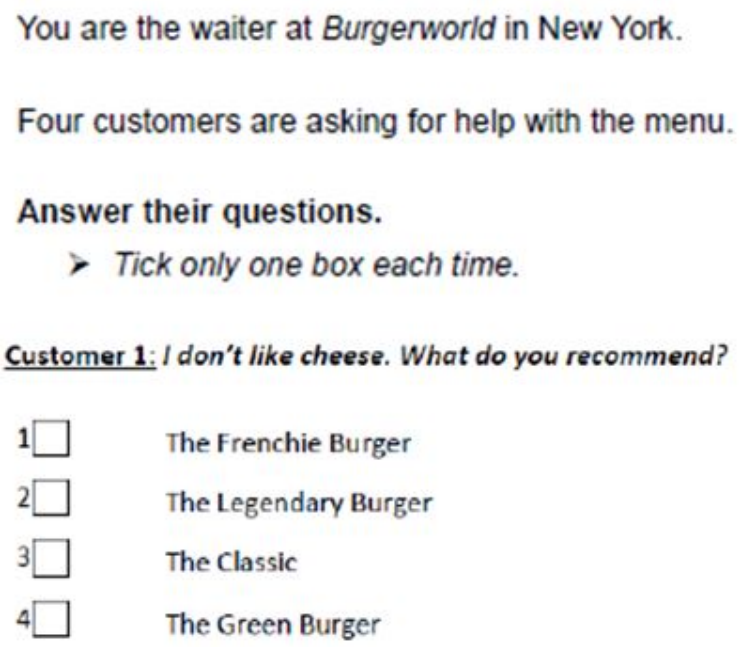
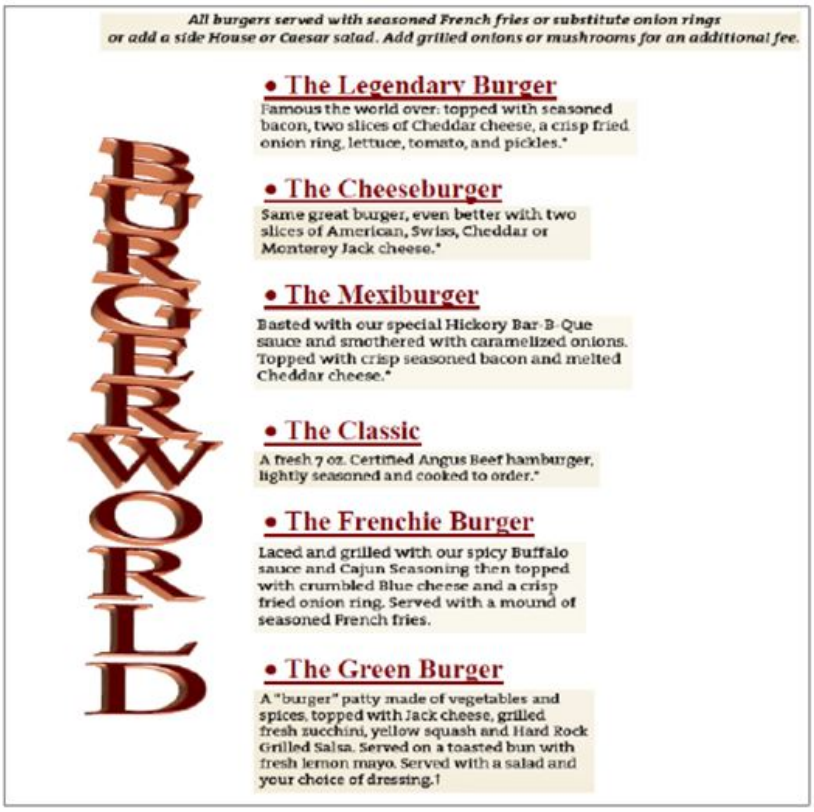


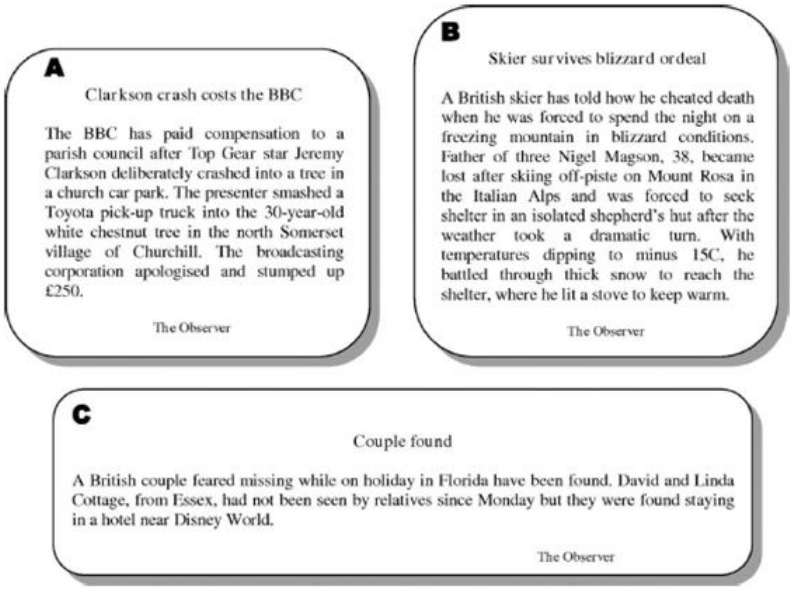
Figure 17 – Support de l'item



Dans cette situation, les élèves lisent le menu d'un restaurant et sont confrontés à des demandes spécifiques de clients, qu'ils doivent conseiller. Le premier client n'aime pas le fromage et demande quel burger le serveur peut lui recommander. La compétence évaluée ici est : identifier l'information explicite (explicite car la réponse se trouve dans la composition des burgers). Pour trouver la bonne réponse parmi les 4 proposées, les élèves doivent s'appuyer sur le repérage du mot cheese. Or seule la description du burger " classic " ne contenait pas ce mot (et cet ingrédient), à la différence des autres burgers. Avec un taux global de réussite de 78,3 % cet item était accessible aux élèves du groupe 2.

5.3.2.b Item caractéristique du groupe 3

Figure 18 – Support de l'item



Dans cette situation, présentant trois articles de presse, les élèves doivent trouver la phrase qui résume le mieux chaque article. La compétence évaluée ici est : construire le sens : synthétiser l'information et l'item 1, représentatif du groupe 3, est réussi à 72,9 %. Les élèves doivent repérer des éléments appartenant à différents champs lexicaux (le ski, la météo, le sport) puis les mettre en relation pour aboutir à une hypothèse conclusive pertinente : la phrase *" It was intensely cold, but at least he spent the night under cover "* était celle qui résumait l'article B.

Figure 19 – Consigne de l'item réussi par le groupe 3

Match the statements (1-4) to the articles (A, B, C).
Tick the letters in the corresponding boxes.

		Article A	Article B	Article C
1	It was intensely cold, but at least he spent the night under cover.	☐1	☐2	☐3
2	These people had no real reason to seek help.	☐1	☐2	☐3
3	This man had problems after he had left the unauthorized area.	☐1	☐2	☐3
4	What happened to this man was not an accident.	☐1	☐2	☐3

5.3.2.c Item caractéristique du groupe 4

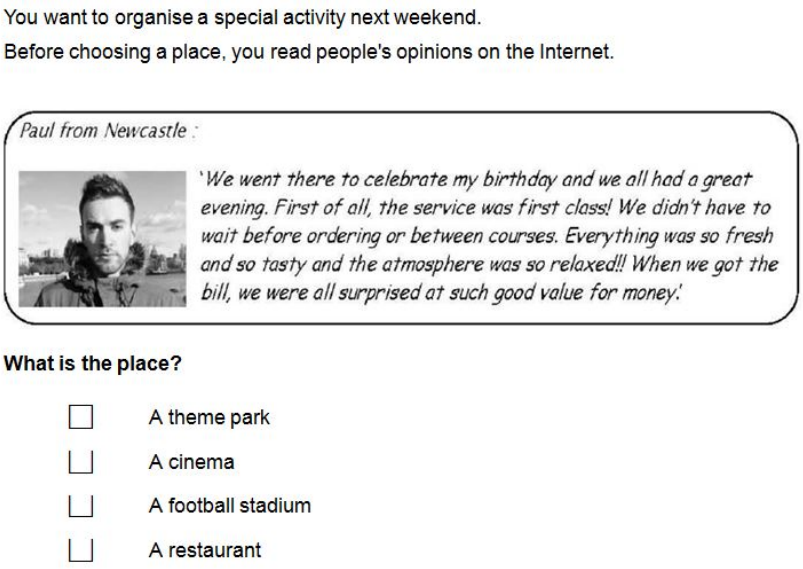
Figure 20 – Exemple d'item réussi par le groupe 4

Dans cette situation, présentant trois critiques de films, les élèves doivent déterminer quel film sera le plus adapté à une fillette de 5 ans, la soeur de l'adolescent dont on montre la photographie. Pour trouver la bonne réponse, les élèves de ce groupe doivent croiser les différents indices présents dans les trois documents et comprendre que le film adapté à un jeune public est *" The Wind "*. Ils doivent s'appuyer sur les indices (*"family film, inventive to kids and parents"*, notamment. Le taux global de réussite est de 58,8 % et il souligne la difficulté chez les élèves de construire du sens en effectuant une mise en réseau d'indices. 23,2 % d'entre eux ont fait porter leur choix sur le film The Rush, qualifié pourtant dans la critique de *" dark and mysterious "*, donc peu adapté à un jeune public et 16,4 % ont choisi de conseiller le film Going West, (*" a brilliant mixture of extreme violence "*), qui fait l'apologie d'une violence peu recommandable pour une fillette de 5 ans.

5.3.2.d Item caractéristique du groupe 5

L'item présenté pour le groupe 5 évalue la maîtrise de la compétence identifier l'information implicite. On propose aux élèves, qui doivent organiser un événement le weekend suivant, de choisir un lieu et pour ce faire, de lire des avis d'utilisateurs sur un site internet. On donne à lire l'avis de Paul, originaire de Newcastle, sur un lieu où il est possible de célébrer des anniversaires (*" We went there to celebrate my birthday "*) et on demande aux élèves de définir l'endroit dont il est question : s'agit-il d'un parc à thème, d'un cinéma, d'un stade de football ou d'un restaurant ?

Figure 21 – Exemple d'item réussi par le groupe 5



Le taux global de réussite est de 52,3 %, soit un peu plus d'un élève sur deux pour lesquels la description correspond bien à celle d'un restaurant. La répartition des taux de réussite selon les réponses possibles apporte un précieux éclairage sur les compétences des élèves en compréhension d'un support écrit en anglais. En effet, la réponse 3 (a football stadium) est choisie par 15,8 % des élèves. Ces derniers se sont appuyés sur *" courses ", " fresh ", " atmosphere "*, qu'ils ont mis en relation pour aboutir à la conclusion qu'il s'agissait d'un stade de football, où l'on peut courir et profiter du bon air frais. Dans leur cas, la compétence " mettre en relation " est maîtrisée mais elle ne s'appuie pas sur un socle de connaissances lexicales suffisant. En effet, le terme *" courses "* est un faux-ami, c'est-à-dire un mot perçu par les élèves francophones comme transparent mais qui ne l'est pas et a un sens différent de celui attendu. En l'occurrence, le mot *" courses "* ne signifie pas " courses à pied " mais signifie des plats, servis par exemple au restaurant, et il aurait dû guider les élèves vers le choix de la réponse *" restaurant "*.

6 Variables contextuelles et non cognitives

6.1 Variables sociodémographiques et indice de position sociale

Un certain nombre de variables sociodémographiques permettent d'enrichir l'analyse des résultats. Le score moyen des élèves est ainsi analysé en fonction du genre, du retard scolaire et quand les effectifs le permettent en fonction du secteur d'enseignement. Le lecteur est invité à consulter la Note d'Information pour plus de détails (Dalibard & Beuzon, 2017).

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).

Pour chaque établissement des échantillons de 2004, 2010 et 2016, la moyenne de l'indice de position socio-scolaire a été calculée et la population a ensuite été découpée en quatre groupes selon les quartiles (tableau 18).

Tableau 18 – Score moyen selon l'indice de position sociale moyen de l'établissement (CEDRE anglais - Compréhension de l'écrit)

Indice moyen de l'établissement	Année	Répartition (%)	Score moyen	Écart type
1er quart	2004	25.0	231	46
1er quart	2010	25.1	233	**54**
1er quart	2016	24.8	**258**	**69**
2e quart	2004	24.9	244	44
2e quart	2010	25.2	241	**55**
2e quart	2016	25.5	**273**	**71**
3e quart	2004	24.9	254	48
3e quart	2010	24.9	257	**58**
3e quart	2016	25.5	**275**	**71**
4e quart	2004	25.1	270	53
4e quart	2010	25.7	276	**66**
4e quart	2016	24.6	**308**	**76**

Note de lecture : en 2016, le score moyen des élèves appartenant au quart des collèges les plus défavorisés (1er quart) augmente de 25 points par rapport à 2010. Les évolutions significatives sont indiquées en gras.

Tableau 19 – Score moyen selon l'indice de position sociale moyen de l'établissement (CEDRE anglais - Expression écrite)

Indice moyen de l'établissement	Année	Répartition (%)	Score moyen	Écart type
1er quart	2004	25.0	230	52
1er quart	2010	25.0	223	47
1er quart	2016	24.8	220	47
2e quart	2004	24.9	246	47
2e quart	2010	24.9	**227**	48
2e quart	2016	25.5	230	46
3e quart	2004	25.1	255	46
3e quart	2010	24.4	**245**	46
3e quart	2016	25.2	**236**	47
4e quart	2004	25.1	269	47
4e quart	2010	25.7	**257**	46
4e quart	2016	24.6	256	**51**

Note de lecture : en 2010, le score moyen des élèves appartenant au quart des écoles les plus favorisées (4ème quart) diminue de 12 points par rapport à 2004. Les évolutions significatives sont indiquées en gras.

Tableau 20 – Score moyen selon l'indice de position sociale moyen de l'établissement (CEDRE anglais - Compréhension de l'oral)

Indice moyen de l'établissement	Année	Répartition (%)	Score moyen	Écart type
1er quart	2004	24.9	231	44
1er quart	2010	25.1	**222**	43
1er quart	2016	24.8	**236**	**51**
2e quart	2004	24.8	243	45
2e quart	2010	25.2	234	49
2e quart	2016	25.4	**249**	48
3e quart	2004	25.0	252	47
3e quart	2010	24.0	246	46
3e quart	2016	24.5	**256**	**52**
4e quart	2004	25.4	274	53
4e quart	2010	25.7	**259**	53
4e quart	2016	25.4	**282**	57

Note de lecture : en 2010, le score moyen des élèves appartenant au quart des écoles les plus favorisées (4ème quart) diminue de 15 points par rapport à 2004. 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 socioprofessionnelle des parents mais aussi le parcours de l'élève (raccourcissement de cycle ou maintien dans un cycle, orientation retenue etc.).

6.3 Motivation des élèves face à la situation d'évaluation

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

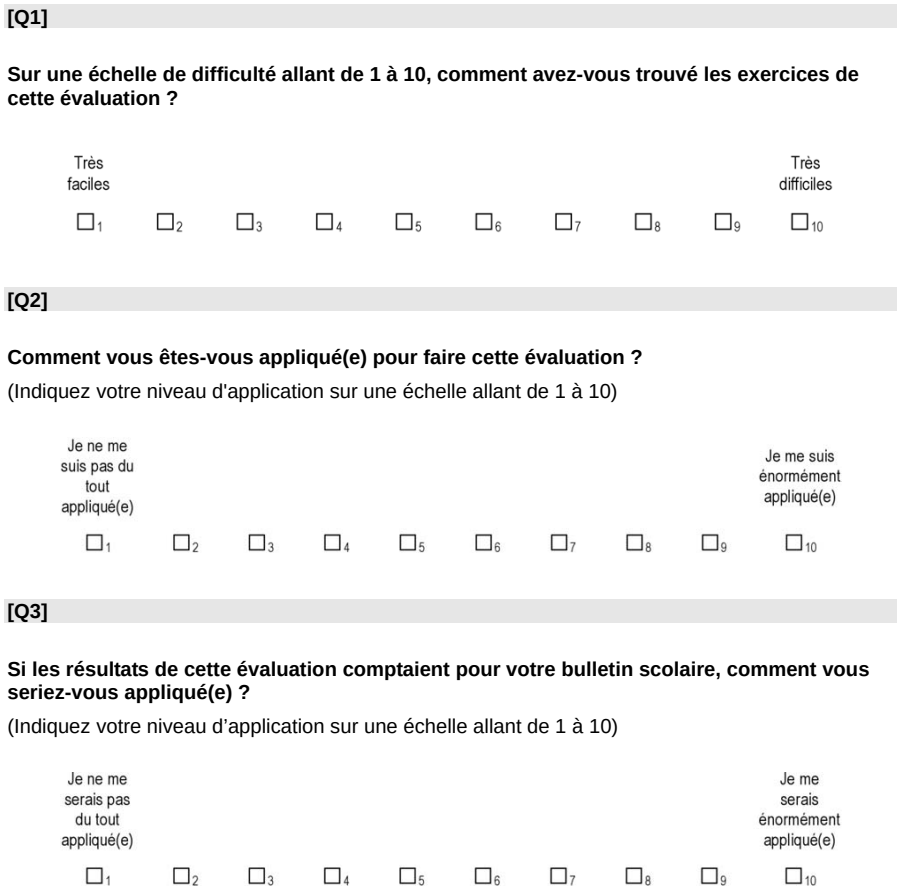
Un instrument pour mesurer la motivation a été adapté à partir du « thermomètre d'effort » proposé dans PISA (Keskpaik. & Rocher, 2015). Cet instrument (cf. figure 22) a été introduit dans plusieurs évaluations conduites au niveau national par la DEPP, y compris dans CEDRE maîtrise de la langue. Les 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 Anglais 2016)

	Moyenne	Erreur standard
Difficulté perçue du test	5,1	0,07
Motivation au test	6,3	0,06
Motivation au test si les résultats comptaient pour le bulletin scolaire	8,6	0,04

Figure 22 – Instrument de mesure de la motivation au test



7 Annexe

Certification AFNOR pour les évaluations CEDRE

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

Les finalités de la certification

Les finalités sont les suivantes :

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

Les enjeux pour la DEPP

Il y a deux enjeux forts pour la DEPP, l'un interne, l'autre externe :

- améliorer les processus de construction des instruments d'évaluation des acquis des élèves, fiabiliser ces processus par une démarche de contrôle-qualité ;
- valoriser l'enquête CEDRE comme un standard de qualité procédurale dans le domaine de l'évaluation.

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

La démarche qualité

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

Les engagements de service

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

Les engagements de service de la DEPP

Des objectifs clairs et partagés

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

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

Des évaluations fondées sur l'expertise pédagogique

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

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

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

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

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

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

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

Une mesure fiable et des comparaisons temporelles pertinentes

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

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

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

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

Des analyses enrichies par des données de contexte

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

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

Transparence des méthodes et partage des résultats

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

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

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

Références

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

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

Dalibard, E., & Beuzon, S. (2017). CEDRE 2004 - 2010 - 2016 - compétences en anglais en fin de collège : en 2016, les élèves sont plus performants en compréhension. *Note d'information*, *20*.

Garcia, E., Le Cam, M., & Rocher, T. (2015). Méthodes de sondage utilisées dans les programmes d'évaluation des élèves. *Éducation et Formations*, *85-86*, 101-117.

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.

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

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

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

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

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