CEDRE

Cycle des Évaluations Disciplinaires Réalisées sur Échantillons

Rapport technique

Espagnol 2016

Collège

Auteurs :

Stéphane BOUCE
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 de l'évaluation

1.1 Objectifs

Le cycle des évaluations disciplinaires réalisées sur échantillons (CEDRE) établit des bilans nationaux des acquis des élèves en fin d'école et de collège. Il couvre les compétences des élèves dans la plupart des domaines disciplinaires en référence aux programmes scolaires. La présentation des résultats permet de situer les performances des élèves sur des échelles de niveau allant de la maîtrise pratiquement complète de ces compétences à une maîtrise bien moins assurée, voire très faible, de celles-ci. Renouvelées 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 totalement 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 : elles informent sur les compétences et les connaissances des élèves à la fin d'un cursus ; elles éclairent sur l'attitude et la représentation des élèves à l'égard de la discipline ; elles interrogent les pratiques d'enseignement au regard des programmes ; elles contribuent à enrichir la réflexion générale sur l'efficacité et la performance de notre système éducatif.

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

CEDRE a 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).

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

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

1.2 Les compétences et connaissances visées

1.2.1 Constition des épreuves

En espagnol, l'évaluation a été proposée dans quatre activités langagières : la compréhension de l'oral, la compréhension de l'écrit, l'expression écrite et, pour la première fois en 2016, l'expression orale en continu. Cette évaluation a été élaborée à partir des objectifs fixés par les programmes officiels entrés en vigueur à la rentrée 2006 (BO n°6 du 25 aout 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 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 mettant ces informations en relation, en inférant à partir de l'explicite, en identifiant l'implicite, en synthétisant. Pour évaluer cette activité langagière, des supports de nature variée ont été proposés : des extraits d'interviews, d'émissions radiophoniques, de publicités, de conversations téléphoniques. Pour la pre-

mière fois, des supports vidéo ont également été proposés à un sous-échantillon d'élèves. Afin de pouvoir apprécier l'impact de l'image sur la compréhension, un autre sous-échantillon a été évalué à partir des seules bandes-son de ces vidéos.

En compréhension de l'écrit, on a mesuré les aptitudes des élèves à reconnaître dans un support écrit 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, en synthétisant. Les élèves ont été évalués à partir de textes littéraires (poèmes, extraits de romans et de contes), de textes informatifs issus de la presse ou de sites Internet, de textes argumentatifs (extraits de blogs, de forums).

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, des textes articulés et nuancés. A partir de situations contextualisées (rédaction d'un courriel, d'un message, participation à un forum de discussion) ou 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 des élèves à décrire et à justifier un choix à partir d'un scénario proposé et de supports iconographiques variés, déclencheurs de parole et à fort ancrage culturel. s Pour des raisons de faisabilité, seul un sous-échantillon de trois élèves sélectionnés par la DEPP au sein de chacune des classes concernées a été évalué dans cette activité langagière. Lors de la passation, les productions de ces élèves ont été enregistrées. Elles devaient être ensuite envoyées à la DEPP afin d'être évaluées par un groupe de correcteurs formés au préalable.

L'élaboration des grilles de compétences (tableaux 2, 3, 4 et 5) en vue de la construction des items a pris appui sur les documents de référence : les programmes officiels qui étaient en vigueur (BO n°6 du 25 aout 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 un lexique de la vie quotidienne Repérer des éléments culturels simples Repérer des indices 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**	
Repérer l'information pertinente	Reconnaître des expressions figées Reconnaître un lexique de la vie quotidienne
Identifier l'information pertinente	Connaître des repères culturels Identifier le(s) thème(s) Repérer l'information explicite Identifier des repères spatiaux / temporels Identifier l'élèment 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, par exemple 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 items fait intervenir des concepteurs, généralement des enseignants. La coordination est assurée par un chargé d'études. Une application dédiée leur permet de créer, modifier ou éditer leur unité ; en outre cette application permet au chargé d'études 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 par 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, au final validé 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 ”facile”, ”moyennement facile” ou ”difficile” est recherché (correspondant pour les langues, aux niveaux A1, A2 et tendant vers B1 du Cadre Européen). En compréhension comme en expression écrite, 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 formes de productions écrites. Elles supposent la mise en place d'un dispositif de corrections expertes à distance pour l'épreuve finale, nécessitant la formation technique des correcteurs et l'élaboration d'un cahier des charges strict de corrections pour 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 ont été saisies de manière automatisée et les questions ouvertes ont été corrigées par des experts via une interface Internet. Certaines questions, notamment celles constituant un ensemble de vrai/faux, ont été 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 ont été 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 Espagnol 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 Espagnol 2016 est constituée d'items de 2010 (représentant un peu moins de 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 ces 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 Espagnol 2016, chaque cahier comprend deux séquences de 51 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 pour que l'évaluation soit passée dans les meilleures conditions quel que soit l'établissement ; la collecte de l'information s'est faite par questionnaires " papier-crayon ".

L'anonymat des élèves et des personnels est respecté, chaque cahier étant repéré par un numéro. Une fois l'évaluation terminée, les cahiers et questionnaires sont renvoyés dans des conditionnements prévus à cet effet, pré-affranchis et pré-étiquetés. Les fichiers MP3 correspondant aux productions des élèves évalués en expression orale en continu ont été également transmis à la DEPP. 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 Espagnol Collége

	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 ESP LV2 des classes >= 10	6 232	528 209
On retire Socle 3ème, Cedre All. et Ang.	5 407	451 692
Base de tirage CEDRE Esp. 3e	**5 407**	**451 692**

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 Espagnol Collége

Strate	**Établissements**	**Élèves**
1. Public hors EP	3 786	340 268
2. EP	896	72 758
3. Privé	1 550	115 183
Total	**6 232**	**528 209**

Échantillon

Le tableau 8 présente la répartition de l'échantillon selon le secteur d'enseignement. Au total, 189 collèges ont été sélectionnées.

Tableau 8 – Répartition dans l'échantillon - CEDRE 2016 Espagnol Collége

Strate	**Établissements**	**Élèves**
1. Public hors EP	120	2 582
2. EP	29	566
3. Privé	40	880
Total	**189**	**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.

97.4 % des établissements de l'échantillon ont répondu à l'évaluation (tableau 9). 89.2 % des éffectifs attendus ont participé (tableau 10).

Tableau 9 – Non-réponse des établissements - CEDRE 2016 Espagnol Collége

Strate	Nb établissements attendus	Nb établissements répondants	% d'établissements répondants
1. Public hors EP	120	119	99.2 %
2. EP	29	29	100 %
3. Privé	40	36	90 %
Total	**189**	**184**	**97.4 %**

Tableau 10 – Non-réponse des élèves - CEDRE 2016 Espagnol Collége

Strate	Nb élèves attendus	Nb élèves répondants	% d'élèves répondants
1. Public hors EP	2 582	2 335	90.4 %
2. EP	566	498	88 %
3. Privé	880	760	86.4 %
Total	**4 028**	**3 593**	**89.2 %**

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 :

- 3.1 items pour la séquence 1
- 3.1 items pour la séquence 2

On considère que :

- 174 élèves n'ont pas vu la séquence 1, dont :
  - 159 n'ont répondu à aucun item de la séquence
  - 15 ont répondu à moins de 50 % de la séquence
- 221 élèves n'ont pas vu la séquence 2, dont :
  - 208 n'ont répondu à aucun item de la séquence
  - 13 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 :

- 2.5 items pour la séquence 1
- 2.3 items pour la séquence 2

On considère que :

- 187 élèves n'ont pas vu la séquence 1, dont :
  - 183 n'ont répondu à aucun item de la séquence
  - 4 ont répondu à moins de 50 % de la séquence

— 205 élèves n'ont pas vu la séquence 2, dont :
  — 200 n'ont répondu à aucun item de la séquence
  — 5 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 37 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 Espagnol Collége

Modalité	Variable	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population
Retard	1	88 491.98	94 329	16.75	17.86
	2	439 717.04	433 880	83.25	82.14
Sexe	1	262 681.36	262 063	49.73	49.61
	2	265 527.66	266 146	50.27	50.39
Strate	1	340 268.02	340 268	64.42	64.42
	2	72 758	72 758	13.77	13.77

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

Modalité	Variable	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population
Retard	1	86 992.42	94 329	16.47	17.86
	2	441 216.53	433 880	83.53	82.14
Sexe	1	261 762.74	262 063	49.56	49.61
	2	266 446.22	266 146	50.44	50.39
Strate	1	340 267.95	340 268	64.42	64.42
	2	72 758	72 758	13.77	13.77

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 Espagnol Collége

	Année	Score moyen	Erreur standard
Compréhension de l'écrit	2010	250	1.49
	2016	255.6	1.27
Expression écrite	2010	250	1.58
	2016	260.3	1.23
Compréhension de l'oral	2010	250	1.72
	2016	247.2	1.35

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 2.86 (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 est significative (5.17).
Compréhension de l'oral : l'évolution du score entre 2010 et 2016 n'est pas significative (1.26).

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 Espagnol Collége

	Année	Groupe <1	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5
Compréhension de l'écrit	2010	5.2	9.9	29	30.1	15.9	10
	2016	0.6	7.8	30.5	34.7	16.3	10.1
Expression écrite	2010	5.9	9.1	21.5	31.1	22.4	10
	2016	NA	4.1	24	34	28	10
Compréhension de l'oral	2010	5.6	9.4	30	30.1	14.8	10
	2016	0.9	11.1	36.2	33	12.7	6.1

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

	Année	Groupe <1	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5
Compréhension de l'écrit	2010	0.4	0.7	1	0.9	0.7	0.7
	2016	0.1	0.6	1.1	0.9	0.8	0.8
Expression écrite	2010	0.4	0.6	0.9	0.9	1	0.8
	2016	NA	0.4	1.2	0.9	1.2	0.8
Compréhension de l'oral	2010	0.5	0.6	1.1	0.9	0.7	0.9
	2016	0.2	0.8	1.2	1	0.8	0.7

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 Espagnol Collége

Année	Erreur Standard	Erreur SAS	*Design Effect*
2010	1.49	0.76	1.96
2016	1.27	0.67	1.91

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 1.91 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.

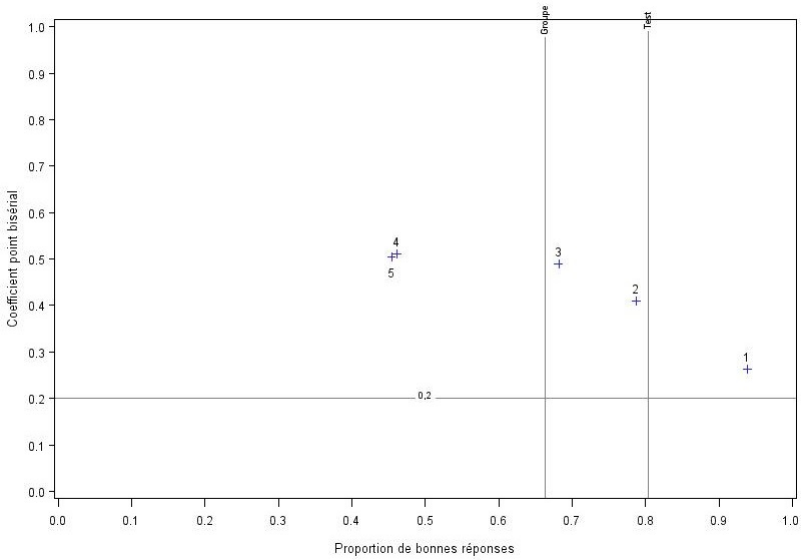
— Non-réponse terminale : l'élève s'est arrêté au cours de l'épreuve, potentiellement en raison d'un manque de temps. Des choix sont effectués pour déterminer le traitement de ces valeurs. Nous considérons que si un élève a passé moins de 50 % d'une séquence, il n'a pas vu la séquence, les valeurs manquantes 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 2 items dont les indices *rbis-point* sont trop faibles :

— 1 item commun à 2010 et 2016
— 1 item de 2016

Expression écrite

Aucun item n'avait un indice *rbis-point* trop faible.

Compréhension de l'oral

Le calcul des indices de discrimination conduit à éliminer 1 item dont l'indice *rbis-point* est trop faible :

— 1 item commun à 2010 et 2016

4 Modélisation

4.1 Méthodologie

4.1.1 Modèle de réponse à l'item

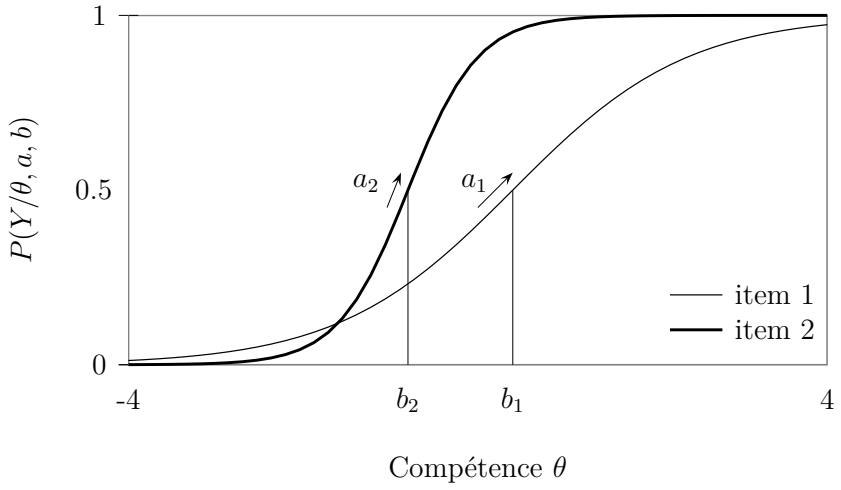
Le modèle de mesure utilisé est un modèle de réponse à l'item à deux paramètres avec une fonction de lien logistique (MRI 2PL) :

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

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 2010).

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 :

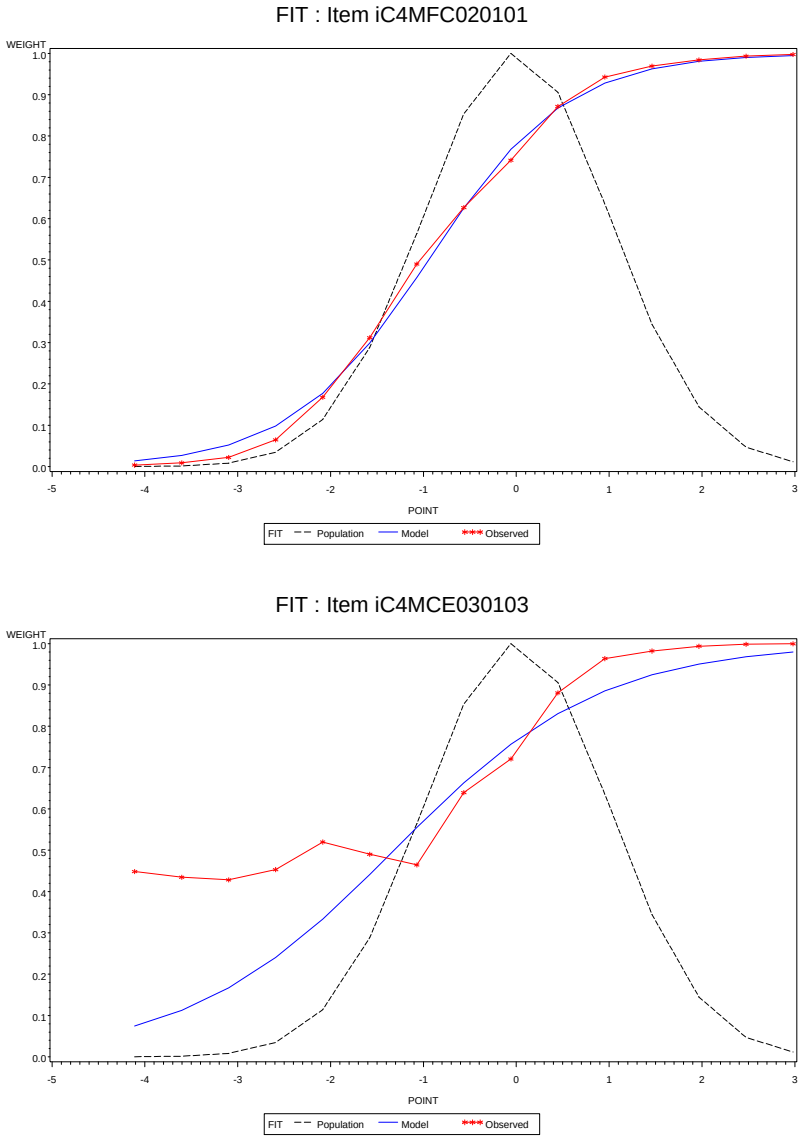
$$\frac{\partial lnL}{\partial \theta_i} + \frac{J}{2I} = 0 \tag{16}$$

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
1 item a été éliminé des calculs :

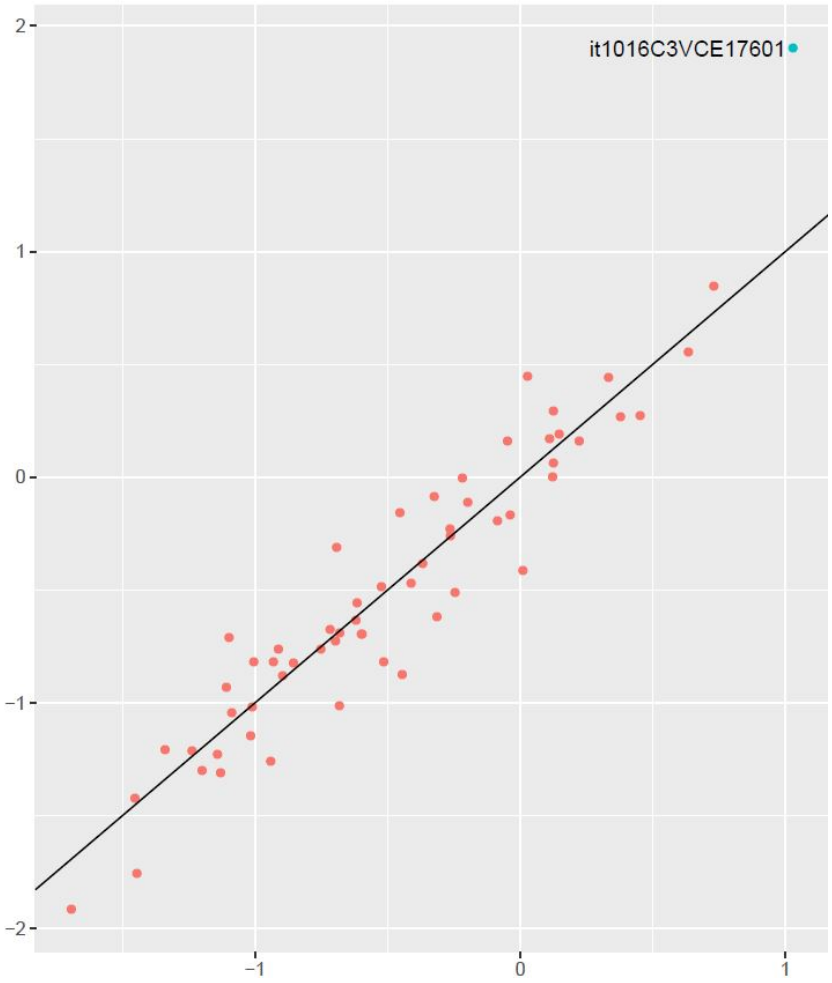
— 1 item pour 2010-2016

Expression écrite
Aucun fonctionnement différentiel n'a été détecté.

Compréhension de l'oral
3 items ont été éliminés des calculs :

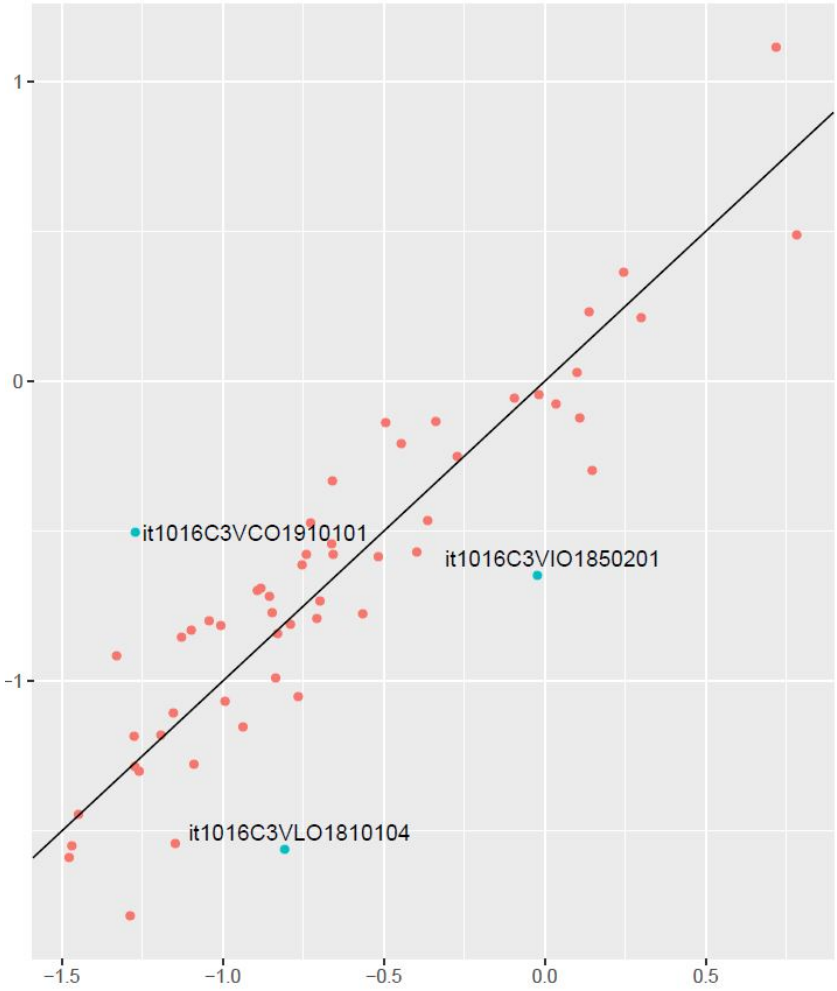
— 3 items pour 2010-2016

Figure 4 – Comparaison des paramètres de difficulté 2010-2016 : Compréhension de l'écrit - (CEDRE Espagnol 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.

Figure 5 – Comparaison des paramètres de difficulté 2010-2016 : Compréhension de l'oral - (CEDRE Espagnol 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 2 années, il y avait au départ :

— 15 items de 2010
— 74 items de 2016
— 60 items d'ancrage 2010-2016

Cela représente 149 items passés par les élèves en tout, dont 134 en 2016.

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

— 15 items de 2010
— 73 items de 2016
— 58 items d'ancrage 2010-2016

146 items sont donc conservés dans l'analyse, dont 131 utilisés dans l'évaluation 2016.

Expression écrite

En considérant l'ensemble des items sur les 2 années, il y avait au départ :

— 20 items de 2010
— 22 items de 2016
— 19 items d'ancrage 2010-2016

Cela représente 61 items passés par les élèves en tout, dont 41 en 2016.

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

— 20 items de 2010
— 22 items de 2016
— 19 items d'ancrage 2010-2016

61 items sont donc conservés dans l'analyse, dont 41 utilisés dans l'évaluation 2016.

Compréhension de l'oral

En considérant l'ensemble des items sur les 2 années, il y avait au départ :

— 16 items de 2010
— 80 items de 2016
— 57 items d'ancrage 2010-2016

Cela représente 153 items passés par les élèves en tout, dont 137 en 2016.

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

— 16 items de 2010
— 80 items de 2016

— 53 items d'ancrage 2010-2016

149 items sont donc conservés dans l'analyse, dont 133 utilisés dans l'évaluation 2016.

4.3 Calcul des scores

Comme indiqué précédemment, une analyse conjointe des données des 2 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 2010. Le tableau 17 présente les résultats obtenus.

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

	Année	Score moyen	Écart-type
Compréhension de l'écrit	2010	250	50
	2016	255.6	39.9
Expression écrite	2010	250	50
	2016	260.3	34.8
Compréhension de l'oral	2010	250	50
	2016	247.2	37.5

5 Construction de l'échelle

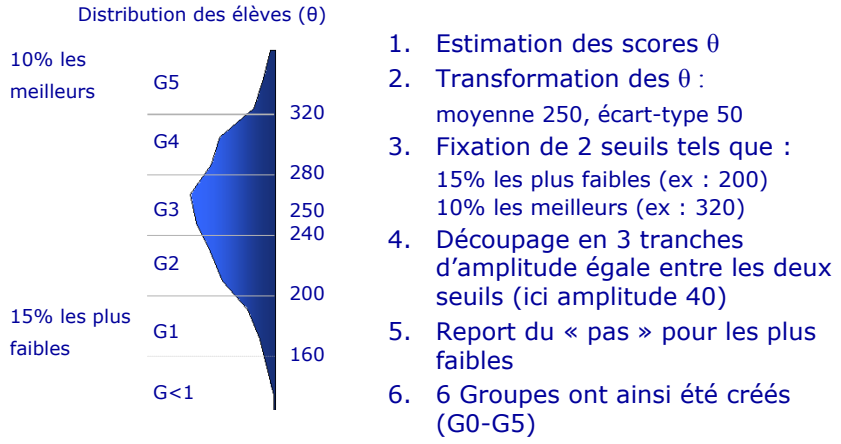
5.1 Méthode

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

Les scores en Espagnol 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 2010. 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 l'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 (0,9 % 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 (11,1 % 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 savent repérer une information explicite concernant la vie quotidienne (formules de salutation, civilités) quand elle est isolée en début ou fin de message. Ils reconnaissent dans le message sonore des éléments culturels très connus et courants. Ils savent repérer certains éléments simples de la présentation (*se llama*).

Groupe 2 (36,2 % des élèves)

Dans un message oral, les élèves savent repérer un lexique courant de la vie quotidienne concernant la description physique, l'alimentation, l'environnement proche, les sensations/sentiments, les loisirs, les saisons, la famille, les couleurs.

Ils repèrent des éléments culturels (la gastronomie). Ils savent repérer des indications chiffrées simples (indication simples de l'heure, des dizaines et des unités). Ils reconnaissent des expressions figées (civilités, expressions de la conversation téléphonique, de la langue de la classe) Ils repèrent des éléments plus complexes de la présentation (nationalité, âge...). Ils commencent à identifier certains éléments de la situation d'énonciation.

Groupe 3 (33,0 % des élèves)

Dans un message oral, les élèves savent repérer un lexique plus étendu et plus riche (description physique, sensations ; ils connaissent des synonymes) même quand le débit est assez rapide. Ils repèrent des expressions lexicalisées (consignes de travail). Ils savent repérer des informations chiffrées plus complexes. Ils ont des repères culturels plus étendus. Ils identifient la situation d'énonciation, le thème d'une conversation. Ils commencent à mettre en relation des informations explicites (lexique de la description physique, repères dans l'espace) et ils savent les associer à un document iconographique. Ils infèrent à partir d'éléments explicites. Ils commencent à identifier l'information implicite et à synthétiser.

Groupe 4 (12,7 % des élèves)

Dans un message oral, les élèves savent mettre en relation des informations explicites pour accéder à du sens même si le contexte est moins familier. Ils savent repérer des informations chiffrées complexes. Ils savent garder en mémoire les expressions et les mots porteurs de sens pour déduire, même quand le débit est rapide. Ils savent synthétiser des informations de nature diverse et inférer à partir de l'explicite.

Groupe 5 (6,1 % des élèves)

Dans un message oral, les élèves savent mettre en relation des informations explicites pour accéder à du sens même si le contexte est moins familier. Ils savent repérer des informations chiffrées complexes. Ils savent garder en mémoire les expressions et les mots porteurs de sens pour déduire, même quand le débit est rapide. Ils savent synthétiser des informations de nature diverse et inférer à partir de l'explicite.

5.2.2 Compréhension de l'écrit

Groupe $<$ 1 (0,6 % 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 (7,8 % des élèves)

Les élèves ont une connaissance limitée du lexique familier 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 (30,5 % des élèves)

Les élèves ont une certaine connaissance du lexique concernant leur environnement proche (la description physique, la maison, les animaux, l'alimentation). Ils savent identifier la situation d'énonciation à partir d'indices lexicaux transparents ou assez courants ou concernant leur environnement familier. Ils sont capables de mettre en relation des informations simples et de les associer à un document iconographique en s'appuyant sur un lexique transparent et/ou de la vie quotidienne. Ils savent retrouver l'ordre chronologique d'un texte en s'appuyant sur un lexique courant, des connecteurs temporels, l'indication de l'heure. Ils sont capables de retrouver une information explicite facilement identifiable.

Groupe 3 (34,7 % des élèves)

Les élèves connaissent des expressions figées et les champs d'un lexique courant. Ils ont des repères culturels sur lesquels ils peuvent prendre appui. Ils savent identifier la situation d'énonciation à partir du repérage d'un lexique plus étendu et/ou concernant un environnement moins familier. Ils savent s'appuyer sur des indices temporels (temps, connecteurs). Ils savent identifier le thème d'un document. Ils savent retrouver l'ordre logique et chronologique d'un texte en s'appuyant sur des connecteurs temporels variés, sur le temps des verbes. Ils commencent à synthétiser à partir d'une mise en relation d'indices lexicaux ; ils peuvent associer cette synthèse à un document iconographique. Ils commencent à identifier l'implicite.

Groupe 4 (16,3 % 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 savent identifier les repères spatiaux. Ils sont capables d'identifier l'élément qui justifie une affirmation en s'appuyant sur des expressions mémorisées et sur leur connaissance du lexique. Ils savent déduire le sens d'une expression ou d'une phrase en prenant appui sur tous les éléments explicites ou non qui la constitue. Ils sont capables de synthétiser à partir d'une mise en relation d'informations de nature diverse. Ils savent identifier l'information implicite.

Groupe 5 (10,1 % des élèves)

Les élèves savent retrouver l'ordre logique ou chronologique d'un texte en s'appuyant sur un repérage précis d'expressions, d'éléments du lexique, de connecteurs et grâce à leur maîtrise de la compétence pragmatique (la structure de la lettre, de la carte postale...). Ils savent déduire le sens d'une phrase même lorsque le message écrit est d'une certaine complexité. Ils sont capables de synthétiser.

5.3 Exemples d'items

5.3.1 Compréhension de l'oral

Parmi les nouvelles situations crées en 2016, était proposé un extrait de la série espagnole Celia (1993) du réalisateur José Luis Borau et inspirée des ouvrages d'Elena Fortún. Dans ce court extrait, une petite fille, *Celia*, prend congé de sa famille.

L'extrait de cette série est d'un niveau découverte A1 avec un questionnement de difficulté variée (item 1 à 4). Cet extrait a été proposé à une partie des élèves de l'échantillon sous forme de vidéo tandis qu'une autre partie de l'échantillon a été évalué à partir de la seule bande-son de cet extrait.

Figure 7 – Situation

Situación

Vas a ver un vídeo 2 veces. Lee las informaciones siguientes.
Mira el vídeo.

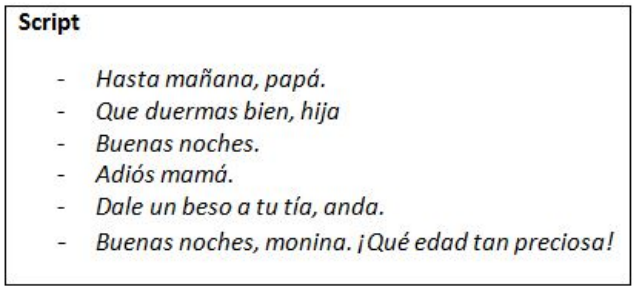
1. Marca la respuesta correcta.
La niña le dice a su padre:
1 □ hasta mañana.
2 □ buenas noches.
3 □ hasta la vista.
4 □ adiós.

2. Marca la respuesta correcta.
La niña le dice a su madre:
1 □ hasta mañana.
2 □ buenas noches.
3 □ hasta la vista.
4 □ adiós.

3. Marca la respuesta correcta.
La niña está con su padre, su madre y:
1 □ una amiga de su madre.
2 □ su abuela.
3 □ su tía.
4 □ una vecina.

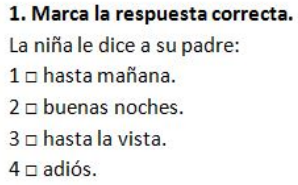
4. Marca la respuesta correcta.
La niña se va:
1 □ a la escuela.
2 □ a jugar con una amiga.
3 □ a estudiar.
4 □ a dormir.

Figure 8 – Script de la situation



5.3.1.a Item caractéristique du groupe 1

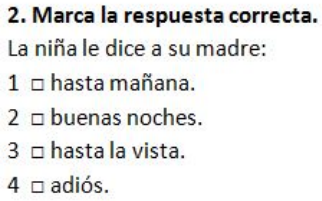
Figure 9 – Item réussi par le groupe 1



Dans l'item n°1 de la situation, les élèves ont à repérer une information explicite : la formule de salutation utilisée par la petite fille pour prendre congé de son père. Le taux de réussite atteint ici est de 97,7 % (il est de 93,1 % à partir de la seule bande son). Ce taux très satisfaisant peut s'expliquer par le fait que l'information à repérer - la formule de salutation - est isolée en début de conversation. Elle correspond aussi à la première possibilité proposée dans l'item. Repérer une information explicite concernant la vie quotidienne quand elle est isolée en début ou fin de message est une compétence que les élèves de ce groupe n°1 savent mettre en oeuvre.

5.3.1.b Item caractéristique du groupe 2

Figure 10 – Item réussi par le groupe 2

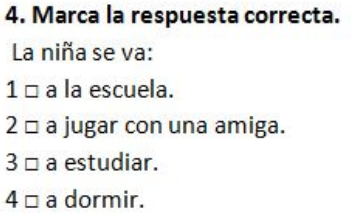


Dans cet item n°2, il s'agit de repérer là aussi une information explicite : la formule de salutation utilisée par la petite fille pour prendre congé de sa mère. Contrairement à l'item n°1, l'information explicite n'est pas ici isolée dans le message oral. D'autre part, parmi les distracteurs, la quatrième proposition est prononcée par la petite fille et la deuxième par la mère, ce qui peut augmenter le niveau de difficulté.

L'image permet davantage d'associer le personnage avec le fragment du discours qui contient l'information. Avec l'image, l'item est réussi par les élèves du groupe 2. Les élèves de ce groupe savent reconnaître des expressions figées (les civilités) dans le message sonore avec ici un taux de réussite de 80,4 %. (Sans l'image, il est réussi par les élèves du groupe 3 avec un taux de 69,8 %.)

5.3.1.c Item caractéristique du groupe 3

Figure 11 – Item réussi par le groupe 3



Dans cet item, il est demandé aux élèves de construire du sens en inférant à partir des informations explicites repérées. C'est une compétence que les élèves du groupe 3 savent mettre en oeuvre. En effet, plus précisément, c'est en prenant appui sur le repérage de certaines des formules de salutations utilisées, et parmi celles-ci, celles spécifiques du soir (*" hasta mañana, buenas noches, que duermas*

bien ") et c'est en les mettant en relation que les élèves de ce groupe savent construire le sens et inférer où se rend la petite : *a dormir*.

A partir du support audio, l'item est réussi par le groupe 3 avec un taux de 65,4 % (avec le support vidéo par le groupe 2 avec un taux de 87,6 %. Les éléments visuels extralinguistiques -objets, ambiance de la pièce, lumière, personnages présents ... - sont des aides, en sus des informations explicitement énoncées, pour déduire).

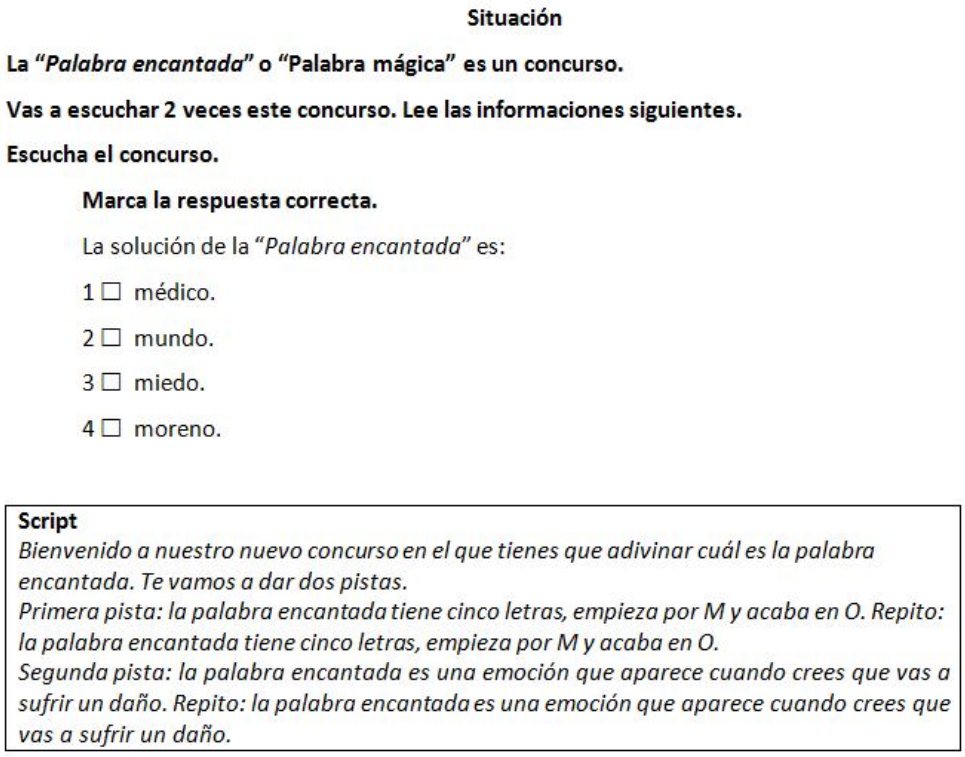
5.3.1.d Item caractéristique du groupe 4

Parmi les nouvelles situations crées en 2016, il a été proposé aux élèves, comme aux auditeurs de l'émission *La palabra encantada* (émission de *Radio 5*, radio nationale espagnole), un jeu radiophonique qui consiste à deviner quel mot est défini en suivant plusieurs pistes.

Dans cet item, les élèves ont à inférer et à synthétiser à partir du repérage d'informations explicites de nature variée. Au sein de la première piste, les élèves peuvent prendre appui sur le repérage d'une indication chiffrée (le nombre de lettres du mot à deviner) et de lettres de l'alphabet (la première et dernière lettre du mot). Au sein de la deuxième piste, ils peuvent s'appuyer sur des éléments transparents du lexique de la définition donnée (*" una emoción, sufrir ... "*). La connaissance d'un lexique moins courant *" daño "* (mal) est une aide supplémentaire pour déduire.

Les élèves du groupe n°4 savent repérer et garder en mémoire ces expressions et mots porteurs de sens, ils savent les mettre en relation et synthétiser. Cet item est réussi par le groupe 4 avec un taux de 46,6 %.

Figure 12 – Item réussi par le groupe 4



Situación

La "*Palabra encantada*" o "Palabra mágica" es un concurso.

Vas a escuchar 2 veces este concurso. Lee las informaciones siguientes.

Escucha el concurso.

Marca la respuesta correcta.

La solución de la "*Palabra encantada*" es:

1 ☐ médico.

2 ☐ mundo.

3 ☐ miedo.

4 ☐ moreno.

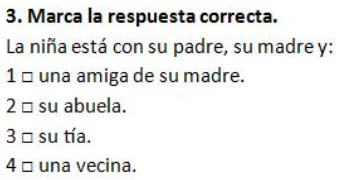
Script
Bienvenido a nuestro nuevo concurso en el que tienes que adivinar cuál es la palabra encantada. Te vamos a dar dos pistas.
Primera pista: la palabra encantada tiene cinco letras, empieza por M y acaba en O. Repito: la palabra encantada tiene cinco letras, empieza por M y acaba en O.
Segunda pista: la palabra encantada es una emoción que aparece cuando crees que vas a sufrir un daño. Repito: la palabra encantada es una emoción que aparece cuando crees que vas a sufrir un daño.

5.3.1.e Item caractéristique du groupe 5

Dans l'item n°3 de la situation proposée à partir de l'extrait de la série espagnole *Celia*, il s'agit de repérer une information explicite relative à la situation d'énonciation : identifier qui est la troisième personne présente dans la scène en sus des parents. Le lien de parenté de cette personne est explicitement énoncé : *" ¡Dale un beso a tu tía, anda ! "*

La complexité du repérage est due ici à la rapidité du débit et à la difficulté à segmenter la chaîne sonore. Avec la seule la bande son, c'est une compétence que les élèves du groupe 5 savent mettre en oeuvre. Le taux de réussite est de 36 %. (Avec le support vidéo, c'est une compétence qui est mise en oeuvre dès le groupe 3. Encore une fois, l'image permet de visualiser le personnage, sa gestuelle et de l'associer au fragment du discours qui inclut l'information.)

Figure 13 – Item réussi par le groupe 5



5.3.2 Compréhension de l'écrit

5.3.2.a Item caractéristique du groupe 2

Parmi les items créés en 2010, était reprise la situation ci-dessous où sont proposées 4 annonces de travail tirées de la presse écrite. Après avoir lu ces 4 annonces, les élèves doivent identifier un élément de la situation d'énonciation : le lieu où travailleront les personnes employées, et ceci pour chacune des annonces. Les situations des annonces B et D sont identifiées par les élèves du groupe 2.

Dans ces deux annonces, pour identifier le lieu, les élèves de ce groupe savent prendre appui sur des éléments d'un lexique transparent : (dans l'annonce D) *" autocares, conductores, permiso D "*, (dans l'annonce B) *" cafetería "*. Dans cette dernière, les élèves de ce groupe savent aussi prendre appui sur des éléments d'un lexique très courant *" bocadillo "* (sandwich) ou sur des repères culturels très connus *" tapas "*.

Figure 14 – Situation

Situación

ANUNCIOS

A. ¿Quieres trabajar como azafata o auxiliar de vuelo? En Iberworld, te ayudamos a conseguirlo. Llámanos al 975 256 322.
B. Camareras para cafetería, buena presencia, conocimientos de barra, preparación de tapas y bocadillos. Incorporación inmediata. Interesadas llamar al 600 509 477.
C. ¿Te gusta el contacto con la gente? Trabaja como ayudante especialista de enfermería, demanda continua de profesionales en el sector sanitario. Llamada gratuita: 200 548 256.
D. Empresa de autocares precisa conductores con permiso D. Interesados llamar al 785 665 125 en horario comercial.

Indica el lugar que corresponde a cada oferta de trabajo.

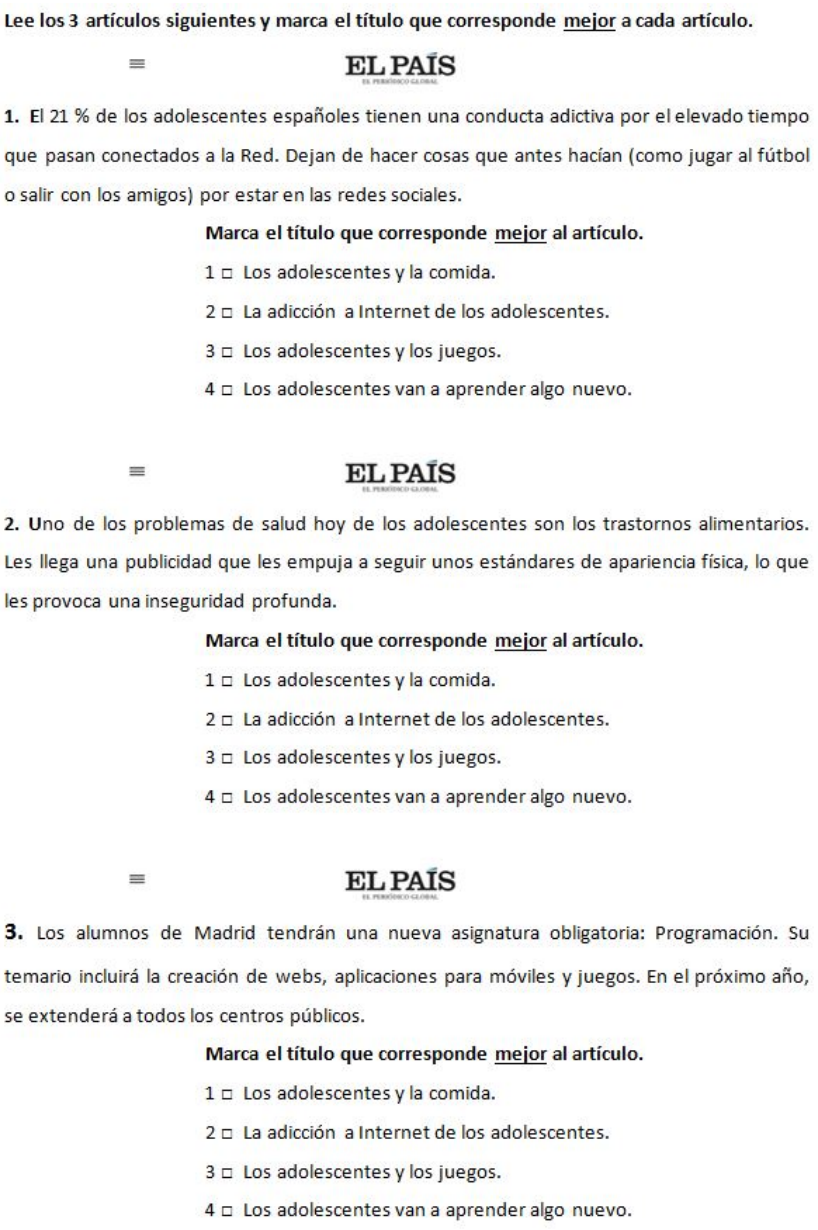
	Un hospital	Un colegio	Un bar	Un avión	Un autobús
A.	☐1	☐2	☐3	☐4	☐5
B.	☐1	☐2	☐3	☐4	☐5
C.	☐1	☐2	☐3	☐4	☐5
D.	☐1	☐2	☐3	☐4	☐5

5.3.2.b Item caractéristique du groupe 3

Dans la situation ci-dessous créée en 2016, sont proposées trois extraits d'articles du journal *El País*. Ces trois articles traitent de problématiques de la vie des adolescents : le premier, de l'addiction à internet et aux réseaux sociaux ; le deuxième, des standards physiques mis en avant par la publicité qui peuvent entraîner des troubles alimentaires et des problèmes de santé ; le troisième, d'un nouvel enseignement proposé dans les collèges de Madrid afin de savoir créer et programmer des sites web, des applications pour les téléphones mobiles, des jeux, etc.

Après avoir lu ces 3 textes, les élèves doivent synthétiser différentes informations et sélectionner lequel des quatre titres proposés correspond le mieux à l'article. Ce sont pour chacun des textes les mêmes titres qui sont proposés. L'item n°2 est réussi par les élèves du groupe 3, les deux autres - items n°1 et 3 - sont respectivement réussis par les groupes 4 et 5.

Figure 15 – Situation



Lee los 3 artículos siguientes y marca el título que corresponde mejor a cada artículo.

≡ EL PAÍS

1. El 21 % de los adolescentes españoles tienen una conducta adictiva por el elevado tiempo que pasan conectados a la Red. Dejan de hacer cosas que antes hacían (como jugar al fútbol o salir con los amigos) por estar en las redes sociales.

Marca el título que corresponde mejor al artículo.

1 □ Los adolescentes y la comida.

2 □ La adicción a Internet de los adolescentes.

3 □ Los adolescentes y los juegos.

4 □ Los adolescentes van a aprender algo nuevo.

≡ EL PAÍS

2. Uno de los problemas de salud hoy de los adolescentes son los trastornos alimentarios. Les llega una publicidad que les empuja a seguir unos estándares de apariencia física, lo que les provoca una inseguridad profunda.

Marca el título que corresponde mejor al artículo.

1 □ Los adolescentes y la comida.

2 □ La adicción a Internet de los adolescentes.

3 □ Los adolescentes y los juegos.

4 □ Los adolescentes van a aprender algo nuevo.

≡ EL PAÍS

3. Los alumnos de Madrid tendrán una nueva asignatura obligatoria: Programación. Su temario incluirá la creación de webs, aplicaciones para móviles y juegos. En el próximo año, se extenderá a todos los centros públicos.

Marca el título que corresponde mejor al artículo.

1 □ Los adolescentes y la comida.

2 □ La adicción a Internet de los adolescentes.

3 □ Los adolescentes y los juegos.

4 □ Los adolescentes van a aprender algo nuevo.

L'item n°2 qui évoque les troubles alimentaires et les problèmes de santé, est réussi par les élèves du groupe 3 avec un taux de 74,1 %. Ces élèves savent prendre appui sur une synthèse des éléments du lexique, un lexique ici trans-

parent ou familier (champs lexicaux concernant l'alimentation, le physique) *" problemas de salud de los adolescentes, alimentarios, aparencia física "*, pour sélectionner le meilleur titre.

5.3.2.c Item caractéristique du groupe 4

L'item n°1 est réussi par les élèves du groupe 4. Les élèves de ce groupe savent ne pas uniquement prendre appui sur le lexique familier *" jugar al futbol "* (jouer au football), *" salir con sus amigos "* (sortir avec leurs amis), ce qui correspond aux activités évoquées dans l'article avant l'importance prise par les réseaux sociaux dans la vie des adolescents.

Les élèves de ce groupe savent s'appuyer sur des champs lexicaux moins familiers (*" conducta adictiva "*), sur la connaissance d'expressions telles que *" dejan de hacer cosas que antes hacían "* (les adolescents cessent de faire des choses qu'ils faisaient avant). Cet item est réussi par les élèves de ce groupe avec un taux de 50,8 %.

5.3.2.d Item caractéristique du groupe 5

Le dernier article, l'item n°3, évoque un nouvel enseignement proposé à Madrid afin de savoir créer et programmer des sites web, des applications pour les téléphones mobiles, des jeux ... Pour parvenir à synthétiser et à identifier le titre pertinent, les élèves de ce groupe savent mettre en oeuvre plusieurs stratégies. Dans ce texte, il y a plusieurs champs lexicaux. Prendre appui sur le seul repérage d'un lexique transparent ou familier ne permet pas d'identifier le titre le plus pertinent.

Pour accéder au sens du texte et pour sélectionner le titre le plus pertinent, les élèves de ce groupe 5 savent mettre en oeuvre d'autres compétences : ils savent s'appuyer sur leur connaissance d'un lexique plus étendu et mettre en relation différents champs lexicaux ; ils peuvent prendre appui sur la valeur des temps utilisés (ici par exemple le futur). C'est par la mise en oeuvre de ces différentes stratégies que les élèves de ce groupe parviennent à sélectionner le titre le plus pertinent. Le taux de réussite est de 33,2 %.

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, Marchois, & Boucé, 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 espagnol - Compréhension de l'écrit)

Indice moyen de l'établissement	Année	Répartition (%)	Score moyen	Écart type
1er quart	2010	24.8	236	48
1er quart	2016	24.0	240	**35**
2e quart	2010	24.8	246	48
2e quart	2016	25.0	**255**	**40**
3e quart	2010	25.1	255	46
3e quart	2016	25.7	260	**40**
4e quart	2010	25.3	263	53
4e quart	2016	25.3	266	**39**

Note de lecture : en 2016, le score moyen des élèves appartenant au 2ème quart augmente de 9 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 espagnol - Expression écrite)

Indice moyen de l'établissement	Année	Répartition (%)	Score moyen	Écart type
1er quart	2010	24.8	233	53
1er quart	2016	24.0	**246**	**32**
2e quart	2010	24.0	248	47
2e quart	2016	25.0	**259**	**35**
3e quart	2010	25.1	255	47
3e quart	2016	25.7	**264**	**34**
4e quart	2010	25.3	264	48
4e quart	2016	25.3	271	**34**

Note de lecture : en 2016, le score moyen des élèves appartenant au quart des collèges les plus défavorisés (1er quart) augmente de 13 points par rapport à 2010. Les évolutions significatives sont indiquées en gras.

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

Indice moyen de l'établissement	Année	Répartition (%)	Score moyen	Écart type
1er quart	2010	25.5	236	45
1er quart	2016	24.0	233	**34**
2e quart	2010	24.3	246	47
2e quart	2016	24.9	247	**38**
3e quart	2010	25.0	253	46
3e quart	2016	25.8	252	**39**
4e quart	2010	25.2	265	57
4e quart	2016	25.3	**256**	35

Note de lecture : en 2016, le score moyen des élèves appartenant au quart des collèges les plus favorisés (4ème quart) diminue de 9 points par rapport à 2010. 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 participant 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 for-

mation 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, il leur est demandé d'évaluer la difficulté de l'épreuve et leur degré d'implication à faire le test.

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.

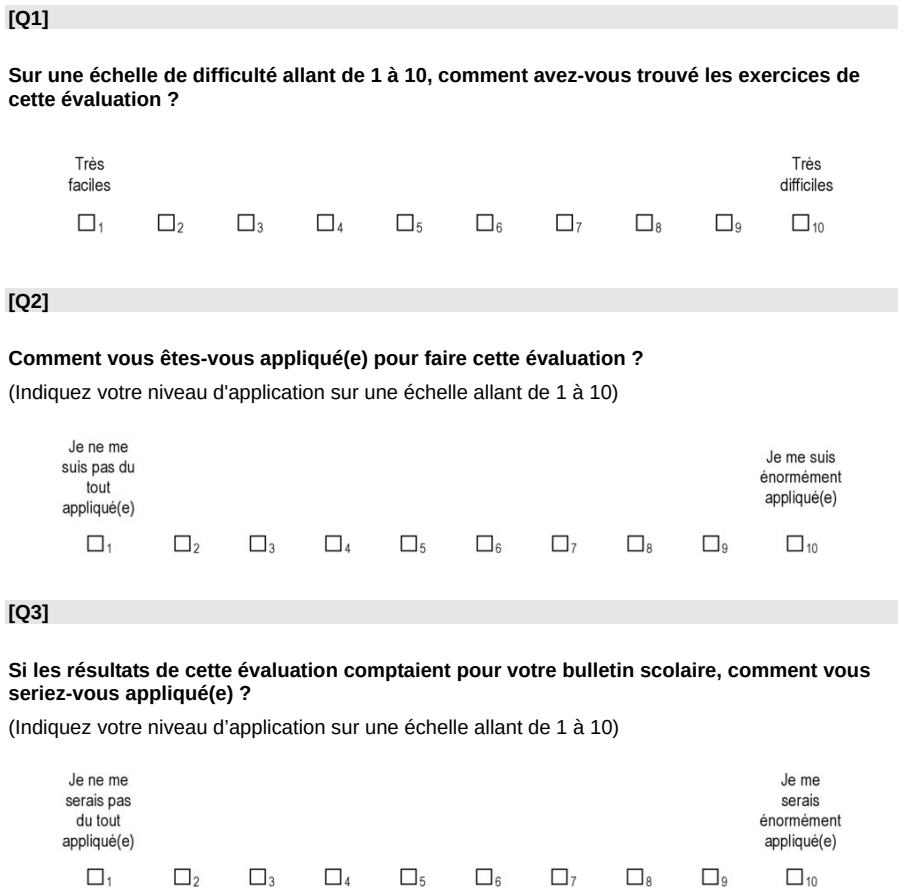
Un instrument pour mesurer la motivation a été adapté à partir du « thermomètre d'effort » proposé dans PISA (Keskpaik. & Rocher, 2015). Cet instrument (cf. figure 16) a été introduit dans plusieurs évaluations conduites au niveau national par la DEPP, y compris dans CEDRE maîtrise de la langue. Les 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 Espagnol 2016)

	Moyenne	Erreur standard
Difficulté perçue du test	5,5	0,06
Motivation au test	6,4	0,06
Motivation au test si les résultats comptaient pour le bulletin scolaire	8,7	0,04

Figure 16 – 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., Marchois, C., & Boucé, S. (2017). CEDRE 2004 - 2010 - 2016 - espagnol et allemand en fin de collège : des progrès à l'écrit, une stabilité à l'oral. *Note d'information*, *21*.

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