CEDRE

Cycle des Évaluations Disciplinaires Réalisées sur Échantillons

Rapport technique

Mathématiques 2014

Collège

Auteurs :
Philippe ARZOUMANIAN
Etienne DALIBARD
Saskia KESKPAIK
Marion LE CAM
Thierry ROCHER

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

Décembre 2015

Table des matières

Introduction

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

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

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

1 Cadre d'évaluation

1.1 Objectifs

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

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

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

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

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

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

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

1.2 Connaissances et compétences visées

1.2.1 Principes généraux

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

Un balayage exhaustif des programmes étant impossible, cette évaluation est conçue à partir de leurs finalités majeures dans un triple objectif :
- connaître les aptitudes des élèves à résoudre des problèmes à caractère mathématique ;
- tester leur connaissance des définitions et des propriétés des principaux concepts ;
- évaluer leur aptitude à raisonner (mener un raisonnement déductif non formalisé à l'écrit, conduire une démonstration, trouver un contre-exemple ou contrôler un résultat).

Les principaux domaines évalués sont au nombre de quatre, identiques à 2008 :

Géométrie : dans le plan, dans l'espace, construction de figure, instruments (règle, équerre, compas, rapporteur), symétries, repérage...

Nombres et calculs : arithmétique, algèbre, calcul mental, calcul posé, calcul instrumenté, calcul exact, calcul approché, entiers, décimaux, fractions, radicaux, comparaison de nombres...

Organisation et gestion de données - Fonctions : proportionnalité, indicateurs statistiques, représentation de données, tableur, grandeur quotient, fonctions affines et linéaires...

Grandeurs et mesures : durée, longueur, aire, volume, unités, conversions, formules usuelles...

1.2.2 Socle et résolution de problèmes

La loi d'orientation du 8 juillet 2013, en son article 13, pose le principe du socle commun :
« La scolarité obligatoire doit garantir à chaque élève les moyens nécessaires à l'acquisition d'un socle commun de connaissances, de compétences et de culture, auquel contribue l'ensemble des enseignements dispensés au cours de la scolarité. Le socle doit permettre la poursuite d'études, la construction d'un avenir personnel et professionnel et préparer à l'exercice de la citoyenneté. Les éléments de ce socle commun et les modalités de son acquisition progressive sont fixés par décret, après avis du Conseil supérieur des programmes. »

Dans ce cadre, certaines des questions proposées ont mobilisé les compétences de résolution de problèmes suivantes [1] :

Chercher
Analyser un problème.
Extraire, organiser et traiter l'information utile.
Observer, s'engager dans une démarche, expérimenter en utilisant éventuellement des outils logiciels, chercher des exemples ou des contre-exemples, simplifier ou particulariser une situation, reformuler un problème, émettre une conjecture.
Valider, corriger une démarche, ou en adopter une nouvelle.

Modéliser
Traduire en langage mathématique une situation réelle (à l'aide d'équations, de suites, de fonctions, de configurations géométriques, de graphes, de lois de probabilité, d'outils statistiques ...).
Utiliser, comprendre, élaborer une simulation numérique ou géométrique prenant appui sur la modélisation et utilisant un logiciel.
Valider ou invalider un modèle.

Représenter
Choisir un cadre (numérique, algébrique, géométrique...) adapté pour traiter un problème ou pour représenter un objet mathématique.
Passer d'un mode de représentation à un autre.
Changer de registre.

Calculer
Effectuer un calcul automatisable à la main ou à l'aide d'un instrument (calculatrice, logiciel).

1. Ces compétences n'ont cependant pas été prises en compte dans la construction de l'échelle mais ont fait l'objet d'une analyse complémentaire.

Mettre en œuvre des algorithmes simples.
Exercer l'intelligence du calcul : organiser les différentes étapes d'un calcul complexe, choisir des transformations, effectuer des simplifications.
Contrôler les calculs (au moyen d'ordres de grandeur, de considérations de signe ou d'encadrement).

Raisonner
Utiliser les notions de la logique élémentaire (conditions nécessaires ou suffisantes, équivalences, connecteurs) pour bâtir un raisonnement.
Différencier le statut des énoncés mis en jeu : définition, propriété, théorème démontré, théorème admis...
Utiliser différents types de raisonnement (par analyse et synthèse, par équivalence, par disjonction de cas, par l'absurde, par contraposée, par récurrence...).
Effectuer des inférences (inductives, déductives) pour obtenir de nouveaux résultats, conduire une démonstration, confirmer ou infirmer une conjecture, prendre une décision.

Communiquer
Opérer la conversion entre le langage naturel et le langage symbolique formel.
Développer une argumentation mathématique correcte à l'écrit ou à l'oral.
Critiquer une démarche ou un résultat.
S'exprimer avec clarté et précision à l'oral et à l'écrit.

1.3 Construction du test

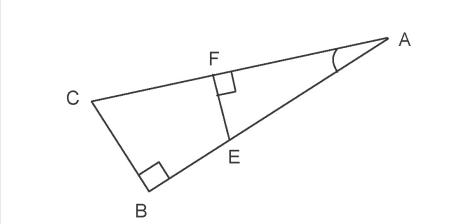
Le bureau de l'évaluation des élèves de la DEPP élabore des évaluations par disciplines et niveaux scolaires. La préparation des unités et de leurs constituants fait intervenir des concepteurs, généralement des enseignants. La coordination est assurée par un chef de projet, membre de l'équipe du bureau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

1.3.1 Elaboration des items

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

Exemples d'items

Exemple 1 : série de vrai-faux



Question

Le triangle ABC est rectangle en B, le triangle AEF est rectangle en F.
Cocher soit VRAI soit FAUX pour chacune des égalités suivantes.

$\tan \hat{A} = \frac{BC}{AC}$	1 ☐ VRAI	2 ☐ FAUX
$\tan \hat{A} = \frac{AE}{EF}$	1 ☐ VRAI	2 ☐ FAUX
$\tan \hat{A} = \frac{BC}{AB}$	1 ☐ VRAI	2 ☐ FAUX
$\tan \hat{A} = \frac{EF}{AF}$	1 ☐ VRAI	2 ☐ FAUX

C3MGE3510101
C3MGE3510102
C3MGE3510103
C3MGE3510104

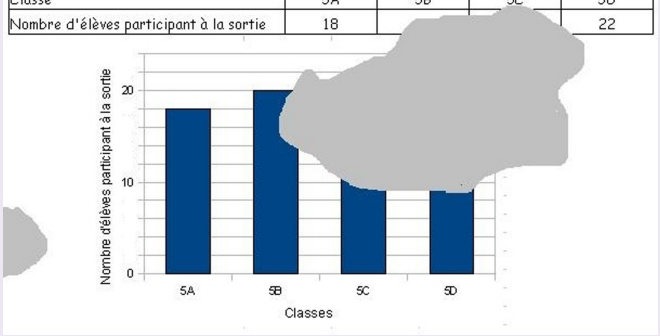
Un équilibre de proportion entre les items considérés comme étant de difficulté « facile », « moyenne » ou « difficile » est recherché. Les items des quatre domaines sont pour certains identiques à ceux proposés en 2008 afin d'assurer une comparabilité de qualité.

Trois formats de questions sont utilisés : questions à choix multiples (QCM), question ouverte appelant une réponse écrite (démonstration, calcul, construction géométrique...) et calcul mental dicté à partir d'un CD audio. Deux exemples d'items sont donnés en encadrés dans cette section.

Exemple 2 : question ouverte

Un collège propose une sortie au cinéma pour les élèves des quatre classes de cinquième.
Le prix d'une place est 4 €. Le collège va payer 288 € pour tous les élèves.
Le professeur a récapitulé, sous forme d'un tableau et d'un diagramme, le nombre d'élèves de chaque classe qui participent à la sortie. Malheureusement la fiche récapitulative a été tachée et certaines données ne sont plus lisibles.

Classe	5A	5B	5C	5D
Nombre d'élèves participant à la sortie	18			22



Le professeur croit se souvenir qu'il y a autant d'élèves de 5C qui participent à la sortie que d'élèves de 5A. Qu'en pensez-vous ?

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 distinguant les différentes compétences nécessaires pour répondre.

Les réponses sous format QCM ont été saisies de manière automatisée et les questions ouvertes ont été corrigées par des experts via une interface Internet (cf.

l'application « Agate » dans la partie 3 du présent rapport). Certaines questions, notamment celles constituant un ensemble de vrai/faux, ont été regroupées afin qu'une question 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.

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

1.3.2 Constitution des cahiers

L'évaluation CEDRE 2014 est constituée de 13 cahiers tournants intégrant un ensemble de 13 blocs d'évaluations contenant des items de 2008 repris à l'identique pour assurer une comparaison diachronique et de nouveaux items qui ont fait l'objet d'une expérimentation en 2013.

Au préalable à la résolution des items présents dans les cahiers, les élèves doivent répondre à une série de questions de calcul mental dicté à partir d'un CD-audio. Pour garantir la qualité de la comparaison avec 2008, notamment en termes de passation des épreuves, près d'un tiers des blocs d'évaluation de 2008 et l'ensemble du calcul mental dicté à partir d'un CD audio ont été conservés à l'identique.

Quatre blocs complémentaires ont été constitués en sélectionnant les items les plus pertinents dans les blocs non retenus intégralement. Les items de 2008 correspondent encore aux programmes de mathématiques pour les élèves de troisième pendant l'année scolaire 2012-2013. En revanche, les nouveaux items prennent en compte les nouveaux programmes progressivement mis en œuvre de la sixième à la troisième depuis la rentrée 2009, dans l'objectif de pouvoir mesurer leurs effets lors de la prochaine étude. Ces nouveaux items, expérimentés en 2013, concernent de nouvelles questions (par exemple « notion de probabilité ») ou approches pédagogiques (par exemple résolution d'un problème ouvert). Dans l'évaluation de 2014, l'analyse s'appuie sur 236 items dont 134 d'ancrage (identiques à 2008) soit 57 %, avant regroupements des items de type « vrai/faux » et analyses psychométriques (cf. parties 3 et 4 du rapport).

La méthodologie des cahiers tournants permet d'évaluer un nombre important d'items sans allonger le temps de passation. Les items sont ainsi repartis dans des blocs d'une durée de 20 minutes et les blocs sont ensuite distribués dans les cahiers tout en respectant certaines contraintes telles que chaque bloc devant

Tableau 2 – Répartition des blocs dans les cahiers pour l'évaluation CEDRE mathématiques collège 2014

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

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

Au final, pour l'évaluation CEDRE 2014, chaque cahier comprend deux séquences cognitives de 45 minutes chacune précédées d'une épreuve de 10 minutes de calcul mental dicté à partir d'un CD-audio. Elles sont complétées par une troisième séquence de 30 minutes (questionnaire de contexte), identique dans tous les cahiers, dans laquelle l'élève doit renseigner plusieurs éléments concernant l'environnement familial dans lequel il évolue, ses projets scolaires et professionnels, sa perception de la matière et de son environnement scolaire.

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

Objectifs

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

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

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

Principes fonctionnels

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

1.4 Passation des évaluations

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

nistrateur du test, son rôle étant de veiller au strict respect de la procédure à suivre pour que l'évaluation soit passée dans les mêmes conditions quel que soit l'établissement. La collecte de l'information s'est faite par questionnaires « papier-crayon ».

Chaque partie était séparée par une pause de 5 minutes. La première partie de l'évaluation correspondait à une épreuve de calcul mental dicté à partir d'un CD-audio. Les deux séquences suivantes interrogeaient les élèves sur leurs connaissances et compétences en mathématiques (tableau 2). Enfin, une dernière séquence consistait à répondre à un questionnaire de contexte. Les professeurs de mathématiques de la classe ou des classes concernées ont également dû renseigner un questionnaire de contexte en ligne deux moins avant le début de la passation des épreuves par les élèves.

L'anonymat des élèves et des personnels a été respecté, chaque cahier étant repéré par un numéro. Une fois l'évaluation terminée, les cahiers et questionnaires étaient renvoyés dans des conditionnements prévus à cet effet, pré-affranchis et pré-étiquetés. Aucun travail de correction n'a été demandé aux 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 [2]. Les évaluations CEDRE suivent la première option tandis que l'évaluation PISA suit la seconde. Des simulations ont permis de montrer que les niveaux de précision des deux options sont très proches, dès lors que le tirage est équilibré (cf. encadré « Tirage d'établissement *versus* tirage de classes »). Le choix de sondages par grappe est motivé par la facilité de gestion. En effet, le fait de sélectionner tous les élèves d'une classe au collège permet d'éviter de mettre en place des procédures de tirage au sort d'élèves une fois les établissements tirés.

On note U la population visée par une évaluation donnée, Y la variable d'intérêt (typiquement le score à l'évaluation, ou bien une indicatrice de difficulté), X une variable auxiliaire, c'est-à-dire connue pour l'ensemble des élèves de la population U. Un échantillon S d'élèves est sélectionné dans la population U. Chaque élève i a la probabilité 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 :

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

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

Tirage d'établissements *versus* Tirage de classes

Pour faciliter la logistique dans les collèges, nous réalisons un 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 :

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

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

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

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

2.2.1 Echantillon 2008

Le tirage est stratifié selon les trois strates :
– 55 collèges publics hors ZEP/REP
– 55 collèges publics en ZEP ou en REP
– 55 collèges privés
Dans chaque strate, un tirage proportionnel à la taille de la population des élèves de 3e du collège est réalisé. Puis dans chaque collège, un tirage aléatoire simple de 30 élèves de 3e, quelle que soit la classe, est réalisé.

2.2.2 Echantillon 2014

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. Le deuxième degré de sondage consiste à interroger tous les élèves de la classe sélectionnée (tirage par grappe).

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

Stratification

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

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

On vise environ 8 000 élèves.

Champ et exclusions

Pour l'année 2014, nous documentons le champ de l'évaluation qui est l'ensemble des élèves de 3e générale de collèges de France métropolitaine (tableau 3).

Préalablement au tirage, les établissements des échantillons d'autres opérations d'évaluations de la DEPP (D'COL et/ou Collège Connecté), ainsi que les établissements de l'échantillon de l'expérimentation de PISA, sont retirés de la base de sondage .

Tableau 3 – Exclusions pour la base de sondage (2014)

	Etab.	Elèves
Etablissements accueillant des élèves de 3e	8 389	815 181
On retire les EREA	8 320	813 897
On retire les étab hors contrat	8 156	811 590
On retire les TOM	8 119	807 392
On ne garde que les collèges	6 920	777 786
3e générales France métropolitaine	6 917	753 556
Base CEDRE Maths 3e	**6 685**	**718 432**

Base de sondage

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

Tableau 4 – Répartition dans la base de sondage (2014)

strate	collèges	classes	élèves
1. Public hors EP	4 072	18 238	465 056
2. EP	979	4 413	98 556
3. Privé	1 634	5 957	154 820
Total	**6 685**	**28 608**	**718 432**

Échantillon

Le tableau 5 présente la répartition de l'échantillon dans les différentes strates. Au total, 330 classes ont été sélectionnées dans 323 établissements, rassemblant 8 026 élèves.

Tableau 5 – Répartition dans l'échantillon 2014 (CEDRE mathématiques collège)

strate	collèges	classes	élèves
1. Public hors EP	118	118	3 010
2. EP	128	135	3 019
3. Privé	77	77	1 997
Total	**323**	**330**	**8 026**

2.3 Etat des lieux de la non-réponse

2.3.1 Non-réponse totale

Parmi la non-réponse totale, nous distinguons selon la non-réponse de classes entières ou la non-réponse d'élèves dans les classes participantes. Les chiffres suivants ont été observés pour 2014. Tout d'abord, 96,7 % des classes de l'échantillon ont répondu à l'évaluation (tableau 6).

Tableau 6 – Non réponse des classes (CEDRE mathématiques collège 2014)

strate	N classes attendues	N classes répondantes	% de classes répondantes
1- public hors EP	118	115	97,5 %
2- EP	135	128	94,8 %
3- privé	77	76	98,7 %
Total	**330**	**319**	**96,7 %**

Au final, 86,6 % des effectifs attendus ont participé (tableau 7).

Tableau 7 – Non réponse globale (classes + élèves, CEDRE mathématiques collège 2014)

strate	N élèves attendus	N élèves répondants	% élèves répondants
1- public hors EP	3 048	2 683	88,0 %
2- EP	3 077	2 495	81,1 %
3- privé	2 029	1 884	92,9 %
Total	**8 154**	**7 062**	**86,6 %**

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").

2008

Les élèves ont passé tout le cahier en une seule fois. La non réponse terminale a été étudiée par cahier. Parmi les élèves ayant de la non réponse terminale, il y en a en moyenne 15,3.

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

Au final, on supprime 30 élèves pour cause de non-réponse terminale dont :
- 20 élèves qui n'ont répondu à aucun item cognitif du cahier
- 10 élèves qui ont répondu à moins de 50 % des items cognitifs du cahier.

2014

Les cahiers élèves sont composés de trois séquences. La non réponse terminale a été étudiée par séquence et par cahier. Parmi les élèves ayant de la non réponse terminale, pour la 1ère séquence, il y en a en moyenne 3,8, pour la 2ème séquence il y en a en moyenne 6,2, et pour la 3ème séquence il y en a en moyenne 7,0.

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

Au final, pour 2014, on considère que :
- élèves n'ont pas vu la séquence 1 dont :
  - 236 n'ont répondu à aucun item de la séquence
  - 5 ont répondu à moins de 50 % de la séquence
- élèves n'ont pas vu la séquence 2 dont :
  - 162 n'ont répondu à aucun item de la séquence
  - 28 ont répondu à moins de 50 % de la séquence
- 205 élèves n'ont pas vu la séquence 3 dont :
  - 173 n'ont répondu à aucun item de la séquence

– 32 ont répondu à moins de 50 % de la séquence

Les élèves dont les trois séquences sont codées en « s » sont considérés comme de la non réponse totale. C'est le cas pour 53 élèves.

2.4 Redressement

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

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

Le tableau 8 montre que le calage concerne principalement les élèves en retard, plus souvent absents à l'évaluation et donc moins nombreux dans l'échantillon que dans la population (16,7 % contre 19,4 %).

Tableau 8 – Comparaison entre les marges de l'échantillon avant calage et les marges dans la population

	Modalité ou variable	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population
Retard	1	119 924.72	139 318	16.69	19.39
	2	598 507.26	579 114	83.31	80.61
Sexe	1	352 407.66	358 677	49.05	49.92
	2	366 024.32	359 755	50.95	50.08
Strate	1	465 055.98	465 056	64.73	64.73
	2	98 556.00	98 556	13.72	13.72
	3	154 820.00	154 820	21.55	21.55

2.5 Précision

L'erreur standard (*se*) peut être calculée sur le score moyen de chaque année (tableau 9).

Tableau 9 – Scores moyens et erreurs standard associées (mathématiques collège 2014)

Année	Score moyen	Erreur standard
2008	250	1.60
2014	243	1.29

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

$$\frac{\left|\hat{\bar{Y}}_{2014} - \hat{\bar{Y}}_{2008}\right|}{\sqrt{se^2_{\hat{\bar{Y}}_{2014}} + se^2_{\hat{\bar{Y}}_{2008}}}} \tag{3}$$

Avec une valeur de 3,42 (supérieure à 1,96), cela signifie que la baisse du score moyen observée entre 2008 et 2014 est statistiquement significative.

Les erreurs standards sont également calculées pour les répartitions dans les différents groupes de niveaux (tableaux 10 et 11).

Tableau 10 – Répartition en % dans les groupes de niveaux (CEDRE mathématiques collège)

Année	Groupe < 1	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5
2008	2.2	12.7	26.8	29.6	18.6	10.0
2014	3.6	15.9	27.8	28.3	15.3	9.1

Tableau 11 – Erreurs standards des répartitions en % dans les groupes de niveaux (CEDRE mathématiques collège)

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

Design effect

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

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

Dans le cas d'un sondage en grappes, la précision est dégradée en comparaison d'un sondage aléatoire simple. L'effet du plan de sondage est donc supérieur à 1 (tableau 12).

Tableau 12 – Effet du plan de sondage (CEDRE mathématiques collège)

Année	Erreur Standard	Erreur SAS	*Design Effect*
2008	1.85	0.56	10.87
2014	0.92	0.51	3.23

Cela signifie qu'en 2014, un sondage aléatoire simple avec un effectif 3 fois moins important aurait conduit au même niveau de précision. En 2008, ce rapport est encore plus élevé car le tirage était en fait à trois degrés (collèges puis classes puis tous les élèves) et non en grappes.

3 Analyse des items

3.1 Méthodologie

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 [3].

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 [4].

Indices de discrimination

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

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

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

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

En effet, on peut montrer que

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

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) [5]. Il s'agit d'identifier les différentes dimensions sous-jacentes aux réussites observées et surtout d'évaluer le poids de la dimension principale, dans la mesure où c'est une optique unidimensionnelle qui sera envisagée lors de la modélisation.

Dans le cas où les items sont dichotomiques, la matrice des corrélations entre items est en fait la matrice des coefficients 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.

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

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

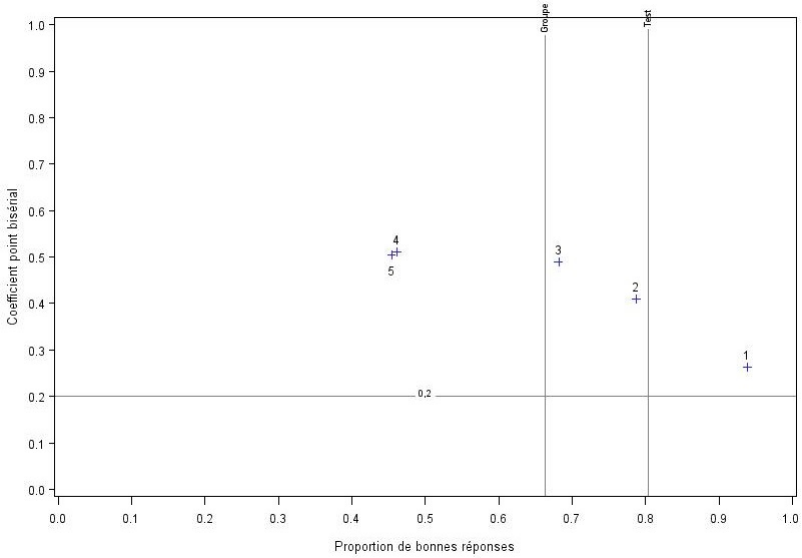
Les séries d'items comportant seulement deux réponses, comme les Vrai/Faux, font l'objet d'un traitement spécifique (cf. l'exemple 1 donné au paragraphe 1.3.1). Les items de ce type sont regroupés pour former un seul item à réponse binaire (réussite ou échec). En effet, la plus forte potentialité de réponse au hasard et l'inter-dépendance des items fragilisent leur utilisation individuelle.

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

3.2.3 Traitement des données et correction des questions ouvertes

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

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



Note de lecture : L'item présenté ici est une série de cinq questions de type « Vrai/Faux ». Chaque croix représente l'item correspondant au seuil de réussite retenu. Par exemple, si la réussite à l'ensemble est attribuée dès lors qu'une seule question est réussie, l'item obtenu a un taux de réussite d'environ 95 % et un coefficient 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 cetitem).

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

Aucun item n'est apparu faiblement discriminant (i.e. tous les *r-bis point* tous supérieurs à 0,2).

3.3.2 Dimensionnalité

Le tableau 13 présente les résultats de l'analyse factorielle des items effectuée sur l'année 2014.

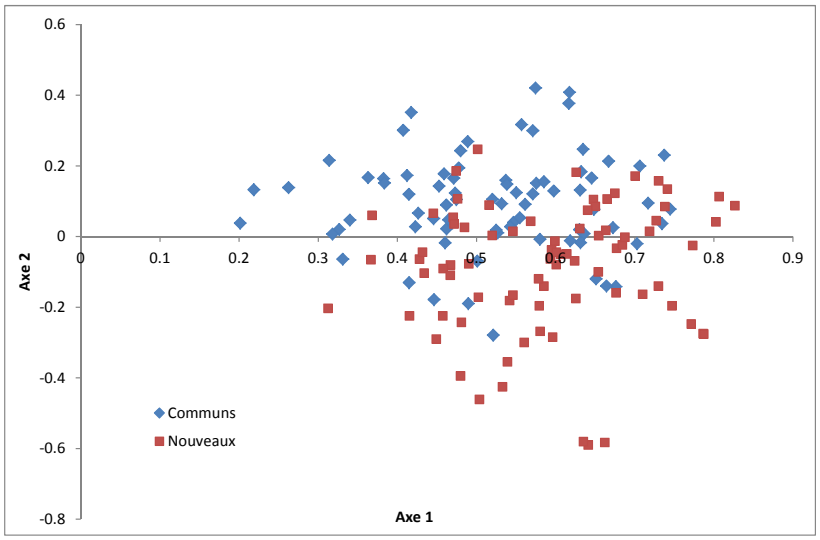
La structure des items est fortement unidimensionnelle : le « poids » de la première dimension est très important (valeur propre de 50,8 contre 5,2 pour la deuxième dimension). En outre, il apparaît une légère différentiation sur la

Tableau 13 – Analyse en composantes principales (CEDRE mathématiques collège 2014)

	Valeur Propre	Différence	Proportion	Proportion cumulée
1	50.8	45.6	0.33	0.33
2	5.2	0.8	0.03	0.36
3	4.4	1.0	0.03	0.39

deuxième dimension, entre les items nouveaux de 2014 et les items repris de 2008, comme le montre la figure 2, et qui est sans doute liée au ré-équilibrage opéré en matière de difficulté des items (cf. plus loin le paragraphe sur les courbes d'information).

Figure 2 – Premier plan factoriel des items de 2014 (CEDRE mathématiques 3e)



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

4 Modélisation

4.1 Méthodologie

4.1.1 Modèle de réponse à l'item

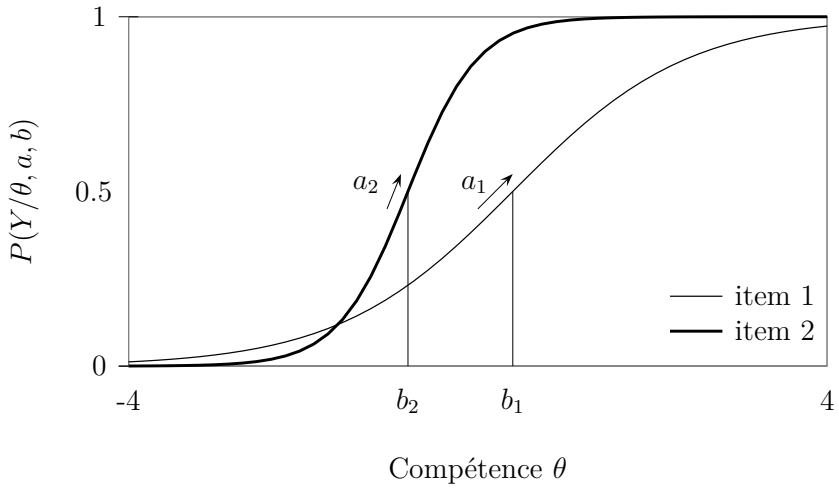
Le modèle de mesure utilisé est un modèle de réponse à l'item à deux paramètres avec une fonction de lien logistique (MRI 2PL) :

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

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

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



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

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

Autre avantage : le niveau de compétence des élèves et la difficulté des items sont placés sur la même échelle, par le simple fait de la soustraction $(\theta_i - b_j)$. Cette propriété permet d'interpréter le niveau de difficulté des items par rapprochement avec le continuum de compétence. Ainsi, les élèves situés à un niveau de compétence égal à b_j auront 50 % de chances de réussir l'item, ce que traduit visuellement la représentation des courbes caractéristiques des items (CCI) selon ce modèle (figure 3).

4.1.2 Procédures d'estimation

L'estimation est conduite en deux temps : l'estimation des paramètres des items puis l'estimation des 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 2008).

Sous l'hypothèse d'indépendance locale des items [6], la fonction de vraisemblance

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

s'écrit :

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

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 :

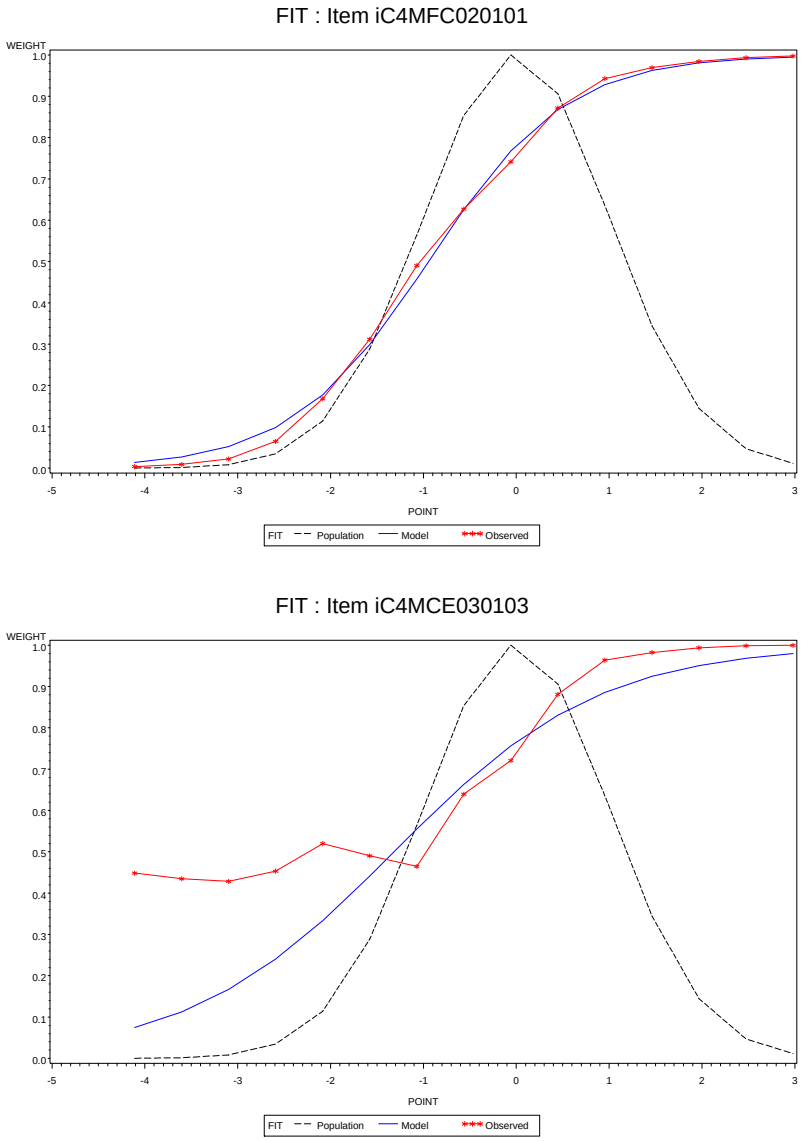
$$\frac{\partial lnL}{\partial \theta_i} + \frac{J}{2I} = 0 \tag{16}$$

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 4). Certaines procédures proposent de comparer directement les probabilités théorique avec les proportions de réussite de groupes d'élèves. Plus généralement, nous pouvons écrire les résidus de la manière suivante :

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

Figure 4 – Exemples d'ajustements (FIT)



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

Les carrés des résidus suivent typiquement une loi du 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 (cf. l'illustration plus loin dans la section 4.4). L'erreur de mesure étant inversement proportionnelle à l'information, cette courbe d'information permet de visualiser la précision avec laquelle le niveau de compétence θ des élèves est estimé.

4.2 Résultats

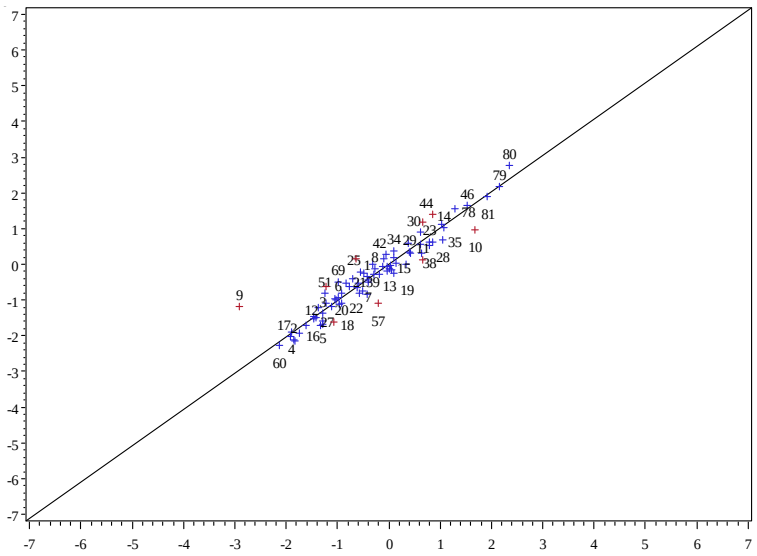
4.2.1 Identification des fonctionnements différentiels d'items (FDI)

L'analyse des FDI a permis de détecter 9 items : 5 items en faveur de 2008, 4 items en faveur de 2014 (figure 5). Ils ont été éliminés des calculs.

4.2.2 Identification des items présentant un mauvais ajustement (FIT)

L'analyse des ajustements (FIT) a permis de détecter 5 items problématiques, dont 1 item commun aux deux années, 1 items de 2008 et 3 items de 2014.

Figure 5 – Paramètres de difficulté 2008-2014 (CEDRE mathématiques 3e)



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

4.2.3 Bilan de l'analyse des items

Au départ, il y avait :
- 81 items communs
- 80 items de 2008
- 99 items de 2014

Après suppression des items présentant un fonctionnement différentiel ou un mauvais ajustement, il reste :
- 71 items communs
- 79 items de 2008
- 96 items de 2014

4.3 Calcul des scores

Comme indiqué précédemment, une analyse conjointe des données (2008et 2014) 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 2007. Le tableau 14 présente les résultats obtenus.

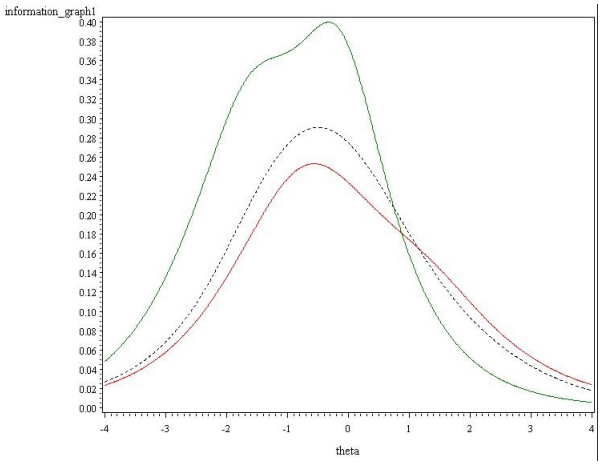
Tableau 14 – Scores CEDRE mathématiques 3e (moyennes et écarts-type)

annee	N	Moyenne	Ecart-Type
2008	4 351	250.0	50.0
2014	7 062	240.3	50.3

4.4 Courbes d'information

La figure 6 représente les courbes d'information pour les items des évaluations 2008 et 2014. La courbe rouge représente les items 2008 non repris en 2014, la courbe verte les items nouveaux de 2014 et la courbe en pointillé les items communs aux deux évaluations. Il ressort que les items nouveaux de 2014 sont globalement plus discriminants et plus facile que les items de 2008. En effet, l'évaluation 2008 comportait trop peu d'items simples adaptés aux élèves les plus faibles. En 2014, un ré-équilibrage a été opéré.

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



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

5 Construction de l'échelle

5.1 Méthode

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

Les scores en mathématiques estimés selon le modèle de réponse à l'item présenté dans la partie précédente ont été standardisés de manière à obtenir une moyenne de 250 et un écart-type de 50 pour l'année 2008. Puis, comme le montre la figure 7, la distribution des scores est « découpée » en six groupes de la manière suivante : nous déterminons le score-seuil en-deça duquel se situent 15 % des élèves (groupes < 1 et 1), nous déterminons le score-seuil au-delà duquel se situent 10 % des élèves (groupe 5). Entre ces deux niveaux, l'échelle a été scindée en trois parties d'amplitudes de scores égales correspondant à trois groupes intermédiaires. Ces choix sont arbitraires et ont pour objectif de décrire plus précisément le continuum de compétence.

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

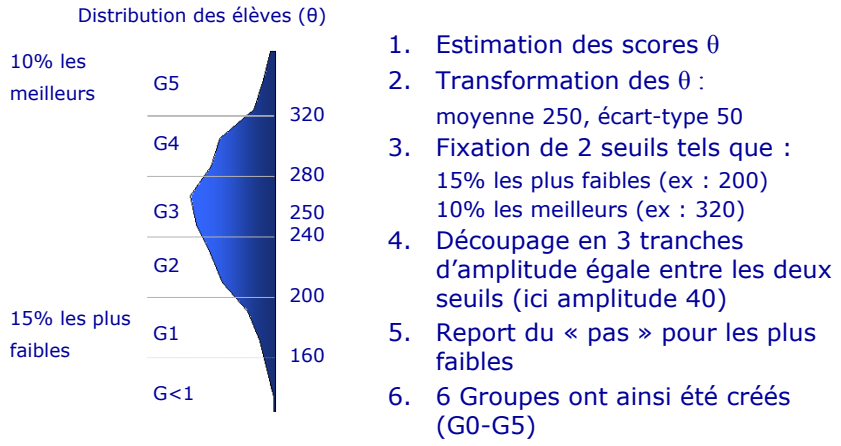
5.2 Caractérisation des groupes de niveaux

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

Groupe < 1 (3,6 % des élèves)

Les élèves du groupe < 1 sont capables de traiter des situations simples mobilisant des grandeurs ou données familières, d'extraire de l'information explicite exhaustive (sans inférence ni interprétation) et de réaliser des calculs avec les quatre opérations sur les entiers (attendus en fin de CM2, début de collège).

Figure 7 – Principes de construction de l'échelle



Groupe 1 (15,9 % des élèves)

Les élèves du groupe 1 manifestent des connaissances et donnent du sens à des situations simples de pourcentage, de représentation dans l'espace, d'unité de durée, et ils sont capables d'un premier pas vers l'interprétation ou la mise en relation.

Groupe 2 (27,8 % des élèves)

Les élèves du groupe 2 possèdent de réelles compétences pour réaliser des calculs sur les nombres entiers et décimaux relatifs. La maîtrise des programmes de calcul est également très satisfaisante. Ils parviennent en effet à remonter un programme de calcul et proposer les expressions littérales associées. Néanmoins, l'utilisation du calcul littéral reste une difficulté pour eux. La proportionnalité est bien utilisée dans des cas simples de la vie courante et reconnue à partir d'un tableau (recherche de l'information). Les conversions d'unités de longueur et de masse simples sont, elles aussi, maîtrisées. La notion de vitesse est globalement comprise, tout comme les calculs de durée (en heures et en minutes).

Groupe 3 (28,3 % des élèves)

Les élèves du groupe 3 peuvent conduire des raisonnements à une étape déductive. Leurs aptitudes à réaliser des calculs algébriques sont étendues. Ils sont capables de développer une expression algébrique simple ou de la factoriser en utilisant la distributivité de la multiplication par rapport à l'addition. En outre, ils utilisent la proportionnalité comme un outil permettant de résoudre les problèmes. En géométrie, ils savent mettre en œuvre certains théorèmes

du programme dans des cas simples. Enfin, le calcul d'aire par dénombrement d'unités et la conversion de durées entre les systèmes sexagésimal et décimal sont acquis.

Groupe 4 (15,3 % des élèves)

Les élèves du groupe 4 sont capables d'analyses à deux étapes déductives. C'est à partir de ce groupe qu'ils produisent des raisonnements formalisés dans une démonstration écrite et citent un contre-exemple pour invalider un énoncé trop général. Confrontés à une figure de géométrie complexe, ils identifient une sous-figure pertinente qui se base sur les conditions suffisantes du théorème usité. De plus, la proportionnalité et les nombres sont des éléments mieux maîtrisés par ces élèves. En effet, ils calculent une quatrième proportionnelle et réalisent des opérations sur les nombres en écriture fractionnaire. Dans le domaine des fonctions, ils comprennent le formalisme f(a)=b. A ce stade, ils ne confondent plus périmètre et aire.

Groupe 5 (9,1 % des élèves)

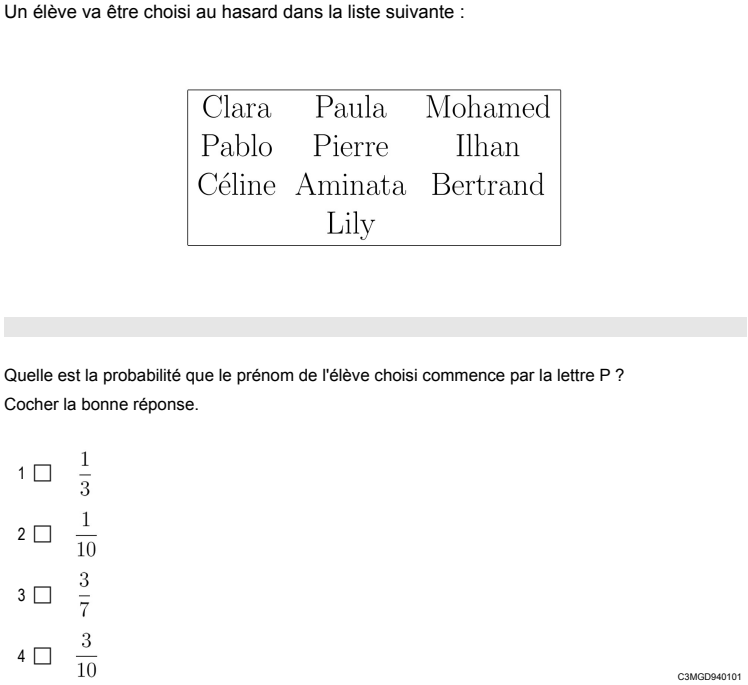
Les élèves du groupe 5 prennent des initiatives et argumentent leurs choix. Dans les différents champs mathématiques, ils mènent des raisonnements structurés. Ils mobilisent correctement un large éventail de définitions et de propriétés enseignées au collège. Ils sont capables de résoudre un problème à l'aide des nombres en écriture fractionnaire et d'effectuer des opérations sur des radicaux. Enfin, les notions sur les fonctions sont mieux comprises et exploitées par ces élèves. Ils établissent, par exemple, des liens entre définition algébrique, représentation graphique et tableau de valeurs, tous associés à une même fonction.

5.3 Exemples d'items

5.3.1 Item caractéristique du groupe <1

Le rapport de la commission de réflexion sur l'enseignement des mathématiques (Kahane, 2002) évoque dans les termes suivants l'enseignement au collège et au lycée : « l'objectif d'une initiation aux probabilités et à la statistique est d'enrichir le langage, de repérer des questions de nature statistique, de définir des concepts qui fonderont un mode de pensée pertinent, rassurant, remarquablement efficace. Selon comment sont conçus les exercices sur ce thème, il est important de souligner que même les groupes en moindre réussite parviennent à réaliser la tâche demandée. ». Cet item (figure 8) est réussi dès le groupe < 1 (probabilité de 0,58 pour les élèves de ce groupe), l'univers étant représenté exhaustivement (il suffit de dénombrer pour l'obtenir).

Figure 8 – Exemple groupe < 1



Un élève va être choisi au hasard dans la liste suivante :

Clara	Paula	Mohamed
Pablo	Pierre	Ilhan
Céline	Aminata	Bertrand
	Lily	

Quelle est la probabilité que le prénom de l'élève choisi commence par la lettre P ?
Cocher la bonne réponse.

1 ☐ $\frac{1}{3}$

2 ☐ $\frac{1}{10}$

3 ☐ $\frac{3}{7}$

4 ☐ $\frac{3}{10}$

C3MGD940101

5.3.2 Item caractéristique du groupe 1

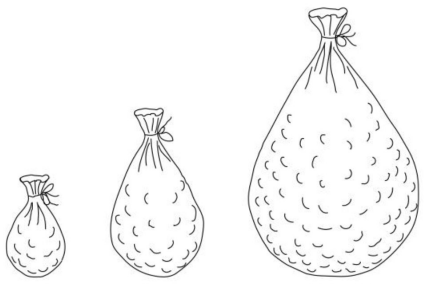
Cet item (figure 9) est réussi dès le groupe 1 (probabilité de 0,52) mais est à la frontière entre le groupe <1 et le groupe 1 (probabilité de 0,43), les univers étant multiples et non représentés. Par rapport à l'item précédent, le saut conceptuel est important. Il faut repérer les différents univers.

Figure 9 – Exemple groupe 1

On dispose de trois sacs de tailles différentes :

- le plus petit sac contient 10 billes,
- le sac de taille moyenne contient 100 billes,
- le grand sac contient 1000 billes.

Dans chaque sac, il n'y a qu'une seule bille rouge.
Sans regarder et au hasard on prend une bille de chacun des sacs.



Question

Quel sac doit-on choisir pour avoir le plus de chances de tirer une bille rouge ?
Cocher la bonne réponse.

1 ☐ Le sac contenant 10 billes.
2 ☐ Le sac contenant 100 billes.
3 ☐ Le sac contenant 1000 billes.
4 ☐ Il n'y a aucune différence.

5.3.3 Item caractéristique du groupe 2

Cet item (figure 10) est réussi dès le groupe 2 (probabilité de 0,62) mais est à la frontière entre le groupe 1 et le groupe 2 (0,44). Il faut construire l'univers (calculer son cardinal). La question « tiré un crayon au hasard » ne rend pas le contexte familier.

Figure 10 – Exemple groupe 2

Marc possède 6 crayons rouges, 4 crayons verts et 5 crayons bleus.
Il prend au hasard un crayon sans regarder.

Quelle est la probabilité que le crayon tiré au hasard soit vert ?
Cocher la bonne réponse.

1 ☐ $\frac{1}{3}$

2 ☐ $\frac{1}{4}$

3 ☐ $\frac{4}{11}$

4 ☐ $\frac{4}{15}$

5.3.4 Item caractéristique du groupe 3

Un des objectifs de l'enseignement mathématique au collège est que le calcul littéral prenne place dans les moyens d'expression et de résolution de problèmes disponibles pour les élèves, au côté du calcul numérique. Cependant, il ne faut pas que cela se fasse au détriment des procédures personnelles des élèves qui peuvent se révéler très efficaces d'autant plus que ce sont parfois les seules accessibles par les élèves en moindre réussite. Il est donc important de proposer des exercices qui permettent de valoriser ces procédures personnelles dans un premier temps sans oublier ensuite d'en montrer les limites.

Cet item « ouvert » (figure 11) a permis de faire connaître et valoriser différentes démarches de résolution. Il pointe l'aptitude à mobiliser des connaissances et des techniques et à raisonner dans le cadre de la résolution de problèmes. Le taux de non réponse est assez faible (14 %) car même les élèves en moindre réussite se sont engagés en mettant en œuvre une procédure personnelle.

Figure 11 – Exemple groupe 3

Brigitte va dans une librairie.
Elle y achète autant de livres que de magazines.
Les magazines coûtent 2 € chacun et les livres coûtent 6 € chacun.
Elle dépense en tout 40 €.

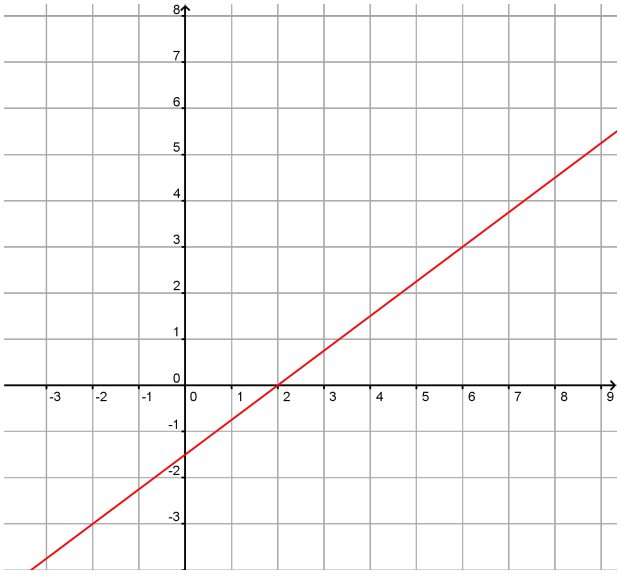
Combien de livres a-t-elle achetés ?
Ecrire les calculs et la réponse dans le cadre ci-dessous.

5.3.5 Item caractéristique du groupe 4

La notion de fonction est intéressante à étudier pour de multiples raisons. Elle permet de mettre en évidence et de décrire la dépendance entre des quantités. Elle est utile pour déterminer une quantité à partir d'une autre et même de comparer les variations de plusieurs quantités. Enfin, elle permet aussi de modéliser (afin d'interpoler, d'extrapoler,...). A partir de la rentrée 2009, un des objectifs du programme de troisième a été de faire émerger progressivement, sur des exemples, la notion de fonction en tant que processus faisant correspondre, à un nombre, un autre nombre. Dans l'exemple présenté ici (figure 12), il s'agit de lire graphiquement l'image d'un nombre par une fonction. Cet item est réussi par le groupe 4 (probabilité de 0,65).

Figure 12 – Exemple groupe 4

Une fonction affine f est représentée graphiquement ci-dessous.
Cocher la bonne réponse.



1 ☐ L'image de 6 par la fonction f est 3.
2 ☐ L'image de 3 par la fonction f est 6.
3 ☐ L'image de 2 par la fonction f est 2.
4 ☐ L'image de 0 par la fonction f est 0.

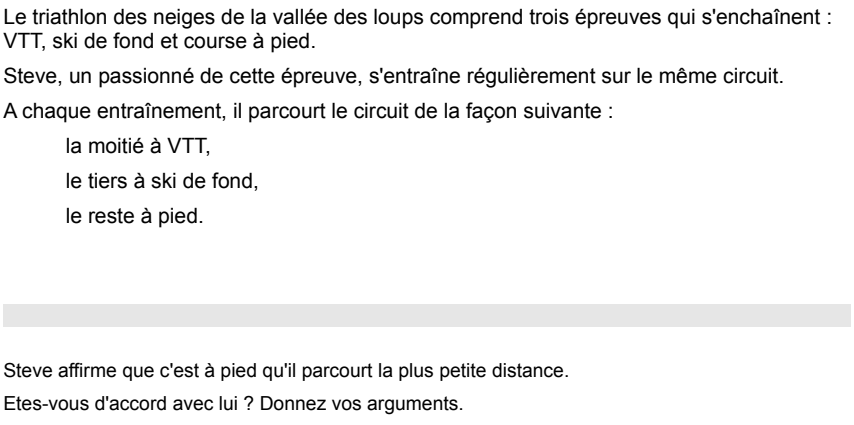
5.3.6 Item caractéristique du groupe 5

Il est clair que la question du calcul sur des nombres exprimés sous forme fractionnaire ne peut pas être traitée uniquement sur le plan de la mise en place de techniques de calcul, mais que celles-ci doivent être justifiées en appui sur les significations données aux écritures fractionnaires et mises en œuvre pour résoudre des problèmes.

C'est le cas ici pour l'exemple présenté (figure 13). Il est réussi à partir du groupe 5 (probabilité de 0,77).

Parmi l'ensemble des procédures qui ont permis aux élèves de produire un raisonnement correct, il est possible d'en mettre cinq en évidence : le calcul de la fraction restante suivi de la comparaison avec les deux connues, l'utilisation d'un exemple de distance 30 m ou 100 m, l'utilisation d'un segment partagé en 1/2 puis 1/3, l'utilisation d'une inconnue suivie de la résolution d'une équation ou la comparaison de 1/3 et de 1/4 puis de 1/2+1/3 et de 1/2+1/4.

Figure 13 – Exemple groupe 5



6 Contexte et dimensions conatives

6.1 Variables sociodémographiques

Un certain nombre de variables sociodémographiques permettent d'enrichir l'analyse des résultats. Le score moyen des élèves est ainsi analysé en fonction du genre, du retard scolaire et quand les effectifs le permettent en fonction du secteur d'enseignement. Le lecteur est invidé à consulter la Note d'Information pour plus de détails (Arzoumanian & Dalibard, 2015).

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

Il n'a été possible d'établir des comparaisons qu'en termes de niveau social des collèges, et non au niveau individuel. En effet, en 2014, la PCS des parents est disponible pour chaque élève, mais elle ne l'était pas en 2008. Pour chaque établissements des échantillons de 2008 et 2014, la moyenne de l'indice de position socio-scolaire a été calculée et la population a ensuite été découpée en quatre groupes selon les quartiles (tableau 15).

Tableau 15 – Score moyen selon l'indice de position sociale moyen de l'établissement (CEDRE mathématiques collège 2008-2014)

Indice moyen de l'étab.	Année	Score moyen	Écart type
Groupe 1 (25 % les plus défavorisés)	2008	227	47
	2014	**219**	45
Groupe 2	2008	251	47
	2014	**241**	48
Groupe 3	2008	254	48
	2014	**242**	45
Groupe 4 (25 % les plus favorisés)	2008	267	49
	2014	269	49

6.2 Élaboration des questionnaires de contexte

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

Pour pouvoir observer des évolutions dans le temps ainsi que de comparer les résultats à d'autres études, certaines questions ont été reprises des cycles précedents du CEDRE et d'autres évaluations nationales et/ou internationales, telles que PISA (Programme international pour le suivi des acquis des élèves). Des questions ont été testées lors du pré-test à l'année $n-1$.

Le questionnaire enseignant interroge les enseignants sur leur formation initiale et continue, leur ancienneté, les niveaux enseignés, la charge et la fonction dans l'établissement. Ce questionnaire inclut aussi des questions sur la qualité relationnelle avec les collègues. Des blocs de questions sont également consacrés aux pratiques pédagogiques. Enfin, les enseignants sont inscités à donner leur opinion sur l'image de la discipline qu'ils enseignent.

Le questionnaire élève contient des questions sur différents aspects du milieu familial et du parcours scolaire des élèves - leur statut d'immigration, la langue parlée à la maison, le redoublement, l'orientation scolaire choisie pour l'année suivante, les diplômes envisagés etc. Le questionnaire interroge également les élèves sur le temps consacré aux activités en lien avec les mathématiques ainsi que sur les pratiques de classe.

Le questionnaire interroge également certaines dimensions dites conatives interessantes à mettre en lien avec le score obtenu à l'épreuve - la motivation, l'anxiété et la perception de soi en mathématiques, la perception du climat d'apprentissage et du climat en classe (soutien par enseignant), l'efficacité de l'apprentissage (stratégies de contrôle et de mémorisation), l'intérêt et le plaisir pour les mathématiques, la préférence pour les formes d'apprentissage compétitif et collaboratif, et les attitudes de l'entourage vis-à-vis des mathématiques.

Enfin, les élèves sont demandés d'évaluer la difficulté de l'épreuve et leur degré d'implication à faire le test.

6.3 Construction des scores factoriels et des indicateurs

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

Pour pouvoir analyser l'évolution des différentes dimensions conatives dans le temps, les coefficients sont calculés sur les données du dernier cycle CEDRE mathématiques (2008), puis sont appliqués sur les données de 2014. Afin de vérifier si la structure factorielle des données reste stable entre les deux cycles, les analyses exploratoires sont effectuées au préalable.

Le tableau 16 présente en guise d'illustration les items d'une de ces dimensions, en l'occurrence l'intérêt pour les mathématiques.

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

Tableau 16 – Exemple de variable conative - intérêt pour les mathématiques (CEDRE collège)

Item du questionnaire	**1er Axe ACP**
Je fais des mathématiques parce que cela me plaît	0,87
J'attends mes cours de mathématiques avec impatience	0,84
Je m'intéresse aux choses que j'apprends en mathématiques	0,81
J'aime lire des textes qui traitent de mathématiques	0,79

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

6.4 Motivation des élèves face à la situation d'évaluation

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

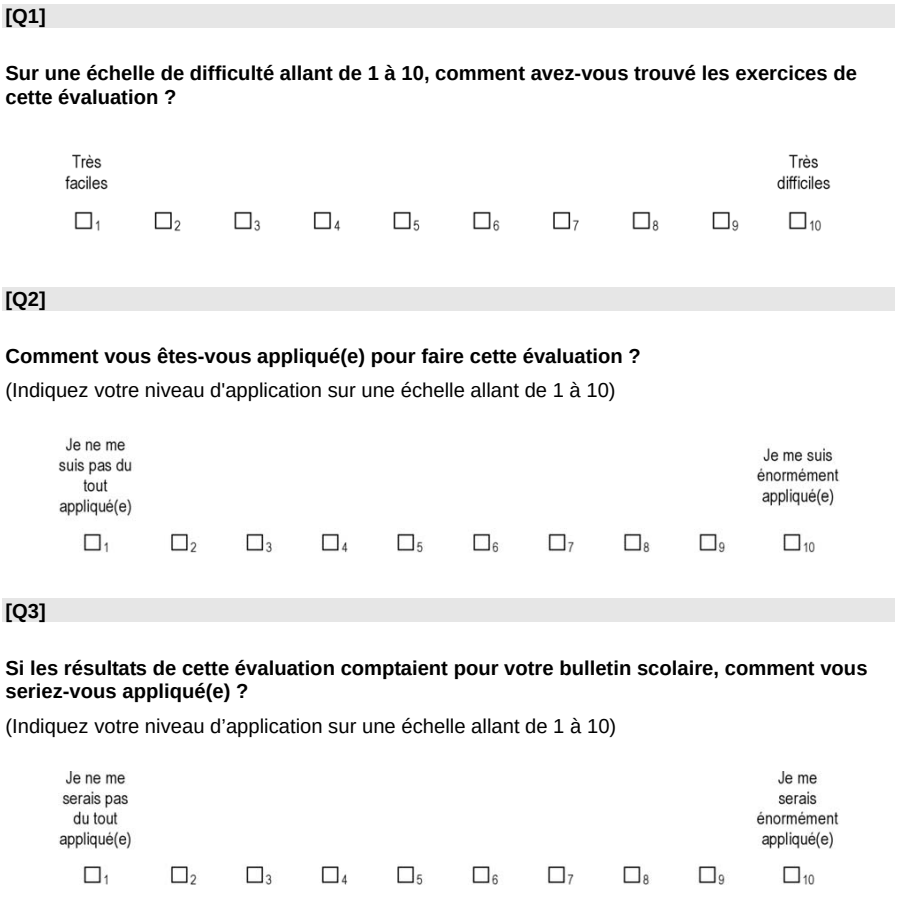
Un instrument pour mesurer la motivation a été adapté à partir du « thermomètre d'effort » proposé dans PISA (Keskpaik. & Rocher, 2015). Cet instrument (cf. figure 14) a été introduit dans plusieurs évaluations conduites au niveau national par la DEPP, y compris dans CEDRE mathématiques. Les données recueillies permettent de distinguer la motivation de l'élève de la difficulté perçue du test, et ainsi de mieux appréhender le lien entre la motivation des élèves français et leur performance. L'analyse de ces données renseigne en outre sur le rôle de certaines caractéristiques, des élèves ou des évaluations elles-mêmes, dans le degré de motivation à répondre aux questions de l'évaluation.

Le tableau 17 présente les grands résultats de cet instrument.

Tableau 17 – Résultats de l'instrument de mesure de la motivation au test (CEDRE mathématiques 2014)

	Moyenne	Erreur standard
Difficulté perçue du test	4,5	0,04
Motivation au test	6,3	0,04
Motivation au test si les résultats comptaient pour le bulletin scolaire	8,9	0,03

Figure 14 – 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.

Arzoumanian, P., & Dalibard, E. (2015). CEDRE 2014 - mathématiques en fin de collège : une augmentation importante du pourcentage d'élèves de faible niveau. *Note d'information*, *19*.

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

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

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

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

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

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

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

Rocher, T. (à paraître). Construction d'un indice de position sociale des élèves. *Éducation et Formations*, *90*.

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