CEDRE

Cycle des Évaluations Disciplinaires Réalisées sur Échantillons

Rapport technique

Mathématiques 2019

Collège

Auteurs :

Louis PHILBERT
Vincent BERNIGOLE
Louis-Marie NINNIN
Reinaldo DOS SANTOS
Marion LE CAM
Franck SALLES
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

Decembre 2022

Table des matières

Introduction **3**

1 Cadre d'évaluation **4**
1.1 Objectifs . . . 4
1.2 Connaissances et compétences visées . . . 4
1.3 Construction du test . . . 10
1.4 Déroulement des épreuves . . . 14

2 Sondage **17**
2.1 Méthodes . . . 17
2.2 Echantillonnage . . . 23
2.3 État des lieux de la non-réponse . . . 25
2.4 Redressement . . . 26
2.5 Précision . . . 27

3 Analyse des items **30**
3.1 Méthodologie . . . 30
3.2 Codage des réponses aux items . . . 33
3.3 Résultats . . . 37

4 Modélisation **38**
4.1 Méthodologie . . . 38
4.2 Résultats . . . 47
4.3 Calcul des scores . . . 48

5 Construction de l'échelle **48**
5.1 Méthode . . . 48
5.2 Caractérisation des groupes de niveaux . . . 50
5.3 Exemples d'items . . . 52

6 Variables contextuelles et non cognitives **55**
6.1 Variables sociodémographiques et indice de position sociale . . . 55
6.2 Élaboration des questionnaires de contexte . . . 56
6.3 Motivation des élèves face à la situation d'évaluation . . . 56

7 Annexe **58**

Références **61**

Introduction

La Direction de l'Evaluation, de la Prospective et de la Performance (DEPP) met en place des dispositifs d'évaluation des acquis des élèves reposant sur des épreuves standardisées. Elle est également maître d'oeuvre 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'Evaluations Disciplinaires Réalisées sur Echantillons) 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

Les évaluations-bilans des acquis des élèves conduites par le bureau de l'évaluation des élèves de la Direction de l'Evaluation, de la Prospective et de la Performance (DEPP-B2) ont pour objectif de faire le point sur les connaissances et les compétences des élèves dans des disciplines (domaine cognitif) ou des attitudes (domaine conatif), en fin d'école primaire et en fin de collège, au regard des objectifs fixés par les programmes d'enseignement.

Ces évaluations n'ont pas valeur de délivrance de diplômes, ni d'examen de passage ou d'attestation de niveau ; elles donnent une photographie instantanée de ce que savent et savent faire les élèves à la fin d'un cursus scolaire. En ce sens, il s'agit bien d'un bilan. Destinées à être renouvelées périodiquement, ces évaluations-bilans permettent également de disposer d'un suivi de l'évolution des acquis des élèves dans le temps. Pour cette raison, les épreuves ne peuvent pas être rendues publiques car, devant être en grande partie reprises lors des 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, bien sûr, sur les compétences et les connaissances des élèves à la fin d'un cursus, mais elles éclairent également sur l'attitude et la représentation des élèves à l'égard de la discipline. Elles interrogent les pratiques d'enseignement au regard des programmes et elles contribuent à enrichir la réflexion générale sur l'efficacité et la performance de notre système éducatif.

Ces évaluations étant passées auprès d'échantillons statistiquement représentatifs de la population scolaire de France métropolitaine, aucun résultat par élève, établissement ni même par département ou académie ne peut être calculé. Le cycle des évaluations disciplinaires réalisées sur échantillon (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).

1.2 Connaissances et compétences visées

Évaluer les connaissances et compétences d'un élève mises en jeu au cours de la résolution de tâches dans un domaine mathématique donné, c'est évaluer l'activité mathématique sur ce domaine. Pour sélectionner les items de l'évaluation CEDRE, les deux approches psychométrique et didactique sont prises en compte

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

pour obtenir une évaluation valide des points de vue du contenu mathématique et de l'élève.

D'un point de vue didactique, plusieurs indicateurs sont mis en jeu. Les tâches du test doivent a priori être représentatives du domaine évalué. Dans le test CEDRE, une tâche correspond à un exercice à résoudre et le(s) item(s) à la(les) question(s) du test relative(s) à leur résolution. L'analyse de leur résolution doit permettre de repérer les processus de réponses des élèves et d'évaluer les connaissances, erreurs, compétences, en particulier démarches et raisonnements que le CEDRE cherche à évaluer en fin de cycle 4. Dans le cas des QCM, les processus de réponses au regard du savoir évalué sont liés aux distracteurs proposés comme réponses possibles. L'ensemble des tâches d'évaluation doit aussi globalement recouvrir le domaine des savoirs évalués et la variété de la complexité de tâches proposées dans les programmes (Grapin et Grugeon-Allys, 2018).

Le cadre d'évaluation de CEDRE 2019 en fin de collège est organisé autour de deux dimensions :

— La dimension des contenus (savoirs), qui précise les domaines mathématiques évalués, en lien avec les programmes du cycle 4 ,
— La dimension cognitive, qui précise les différents niveaux d'activité mathématique des élèves relatifs aux domaines mathématiques qui figurent dans les programmes en fin de cycle 4.

Les domaines mathématiques évalués dans CEDRE 2019 en fin de collège correspondent aux thèmes qui structurent le programme officiel de mathématiques de cycle 4 de 2015 : Nombres et calculs (thème A), Organisation et gestion de données, fonctions (thème B), Grandeurs et mesures (thème C), Espace et géométrie (thème D), Algorithme et programmation (thème E).

Les sous-thèmes retenus dans CEDRE 2019 en fin de collège sont résumés dans

Tableau 2 – Thèmes, sous-thèmes et répartition des pourcentages cibles des items d'évaluation dans les différents thèmes et sous-thèmes

A	Nombres et calculs	35%	A1	Utiliser les nombres pour comparer, calculer et résoudre des problèmes	20%
			A3	Utiliser le calcul littéral	15%
B	Organisation et Gestion de Données, Fonctions	30%	B1	Interpréter, représenter et traiter des données	5%
			B2	Comprendre et utiliser des notions élémentaires de probabilités	5%
			B3	Résoudre des problèmes de proportionnalité	15%
			B4	Comprendre et utiliser la notion de fonction	5%
C	Grandeurs et Mesures	10%	C1	Calculer avec des grandeurs mesurables ; exprimer les résultats dans les unités adaptés	10%
D	Espace, géométrie	20%	D1	Représenter l'espace	5%
			D2	Utiliser les notions de géométrie plane pour démontrer	15%
E	Algorithmique et programmation	5%	E1	Ecrire, mettre au point, exécuter un programme	5%

le tableau ci-dessous, ainsi que les pourcentages cibles des items d'évaluation consacrés à chacun des thèmes et des sous-thèmes en lien avec les attendus de fin de cycle 4. En psychométrie, on entend par item, chacune des questions d'un test.

Etant donné le temps limité prévu pour la passation de l'évaluation CEDRE, des choix ont dû être réalisés tout en rendant compte globalement de l'organisation des contenus mathématiques dans les programmes du cycle 4, en particulier des notions mathématiques associées à chaque thème. Ces choix sont représentés dans la répartition des pourcentages cible des items d'évaluation dans les différents thèmes et sous-thèmes. Le sous-thème Comprendre et utiliser les notions de divisibilité et de nombres premiers (A2) a été fusionné avec le sous-thème A1 Utiliser les nombres pour comparer, calculer et résoudre des problèmes. Certains sous-thèmes comme A1, A3, B3 et D2 sont surreprésentés, cela afin de maximiser l'information sur des sous-thèmes dont les acquis jouent un rôle considéré comme majeur dans la poursuite d'études générales ou professionnelles des élèves.

Le cadre de conception de l'évaluation CEDRE 2019 en fin de collège est conçu

à partir du programme scolaire, qui n'est pas un cadre d'évaluation. Cette partie introduit les éléments théoriques et méthodologiques nécessaires à la définition du cadre de conception.

Le programme indique, pour chaque thème du cycle 4, « les attendus de fin de cycle », et, pour chaque attendu, les aspects de l'activité mathématique visés. Il précise à cette fin les « connaissances et compétences associées » à mobiliser pour résoudre des tâches.

Par exemple, pour le thème A Nombres et calculs, un des trois attendus de fin de cycle est « utiliser le calcul littéral ». Plusieurs connaissances sont attendues, en particulier, les « notions de variable et d'inconnue ». Deux des compétences visées sont : « développer et factoriser les expressions algébriques dans des cas très simples » ; « utiliser le calcul littéral pour prouver un résultat général, pour valider ou réfuter une conjecture ».

Les connaissances correspondent aux notions et aux propriétés du savoir mathématique à enseigner. Les compétences associées correspondent à différents aspects de l'activité mathématique visée dans le domaine impliqué. Cependant, certaines sont trop larges pour définir des tâches d'évaluation au regard d'objectifs d'évaluation scolaires comme dans l'exemple ci-dessus. Il est donc nécessaire de les décrire plus précisément.

Dans la Théorie Anthropologique du Didactique (TAD), au regard d'une référence épistémologique reliée à un cycle scolaire, l'activité mathématique mise en jeu dans la résolution de tâches d'un domaine donné est modélisée par deux blocs : celui du savoir-faire décrit par des types de tâches et des techniques les résolvant, et celui du savoir décrit par des « technologies » reposant sur des propriétés et des raisonnements justifiant les techniques et une théorie les justifiant (Chevallard, 1999). Dans ce cadre, les connaissances issues des programmes relèvent du bloc du savoir et les compétences relèvent du bloc du savoir-faire. Dans des formulations du programme, certaines compétences portent sur différents types de tâches. Par exemple la compétence « développer et factoriser les expressions algébriques dans des cas très simples » regroupe deux types de tâches « développer les expressions algébriques » et « factoriser les expressions algébriques ». De même, « Utiliser le calcul littéral » regroupe les types de tâches : « prouver un résultat général », « valider ou réfuter une conjecture » qui doivent être prises en compte par le présent cadre.

Afin d'assurer la validité didactique du cadre d'évaluation, les types de tâches à évaluer ont été définis pour les sous-thèmes surreprésentés en complément des connaissances et compétences associés des programmes.

A titre d'exemple, les objectifs d'évaluation et les types de tâches évalués pour repérer les acquis des élèves visés en fin de cycle 4 sont présentés pour le sous-thème

A3 dans le (tableau 3). Les compétences attendues associées au programme de 2015 sont mises en perspective avec les types de tâches correspondants. Les connaissances évaluées sont mises en italique dans les paragraphes introductifs à chaque sous-thème.

Principales connaissances et compétences associées des programmes	**Types de tâches**
Utiliser le calcul littéral pour prouver un résultat général, pour valider ou réfuter une conjecture.	— Conjecturer — Prouver à l'aide d'un contre-exemple — Résoudre un problème de généralisation — Résoudre un problème de modélisation via une expression algébrique (production d'expression) — Prouver une propriété ou l'équivalence de PC — Démontrer une propriété d'une expression algébrique — Prouver l'équivalence ou pas entre deux expressions algébriques — Prouver l'équivalence ou pas entre deux équations
Mettre un problème en équation en vue de sa résolution.	— Mettre un problème en équation — Traduire par une équation — Associer
Absent, mais présent dans le document d'accompagnement des programmes « Utiliser le calcul littéral » (mars 2016) et apparaît dans les programmes consolidés 2018	— Traduire (programme de calcul, périmètre, aire, arbre...) par une expression algébrique — Associer
Absent, mais présent dans le document d'accompagnement des programmes « Utiliser le calcul littéral » (mars 2016)	— Déterminer la structure d'une expression algébrique (somme, produit) — Reconnaître la structure d'une équation
Absent des compétences mais développé dans les exemples de situations : « Tester sur des valeurs numériques une égalité littérale pour appréhender la notion d'équation »	— Tester si un nombre est solution d'une équation
Développer et factoriser des expressions algébriques dans des cas très simples.	— Développer — Factoriser
Absent, mais présent dans le document d'accompagnement des programmes « Utiliser le calcul littéral » (mars 2016)	— Réduire (forme développée réduite avec le moins de signe possible, et cas 3n+5n=8n) — Réécrire une expression algébrique (terme ou facteur, par exemple $16x^2$ en $(4x)^2$) — Substituer
Résoudre des équations ou des inéquations du premier degré.	— Résoudre une équation

TABLEAU 3 – Objectifs d'évaluation et types de tâches évalués pour le sous-thème « Utiliser le calcul littéral » (A3)

Parmi les types de tâches décrits, certains sont retenus alors qu'ils ne correspondent pas à des compétences du sous-thème. Mais ils apparaissent dans les documents d'accompagnement cités dans le tableau ou dans les exemples de situations, d'activités et de ressources pour l'enseignant. De plus des travaux en didactique des mathématiques montrent qu'ils sont souvent implicites alors qu'ils participent à l'activité mathématique à développer chez les élèves (Grugeon-Allys et al., 2012 ; Pilet, 2015).

Ces types de tâches ont permis de proposer des instructions pour la conception des situations 2019.

1.3 Construction du test

L'évaluation CEDRE 2019 sur support numérique se compose d'un ensemble de modules, constitués de sections, qui sont elles-mêmes composées d'items. La préparation des modules et de leurs constituants fait intervenir des concepteurs (enseignants, formateurs, chercheurs et inspecteurs). Le groupe de conception est coordonné par un chef de projet (généralement personnel du bureau DEPP-B2.1) sous la responsabilité de la cheffe du bureau DEPP-B2.1

1.3.1 Elaboration des questionnaires

Les items sont le fruit d'un travail collectif des concepteurs, encadré par le chargé d'étude (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, au final validé par le chargé d'étude et l'inspection. L'item fait alors l'objet d'un cobayage, c'est-à-dire d'une passation auprès d'une ou plusieurs classes pour estimer sa difficulté et recueillir les réactions des élèves.

Une tâche d'un type donné peut être résolue avec différentes techniques justifiées par des notions, propriétés et raisonnements eux-mêmes différents. Ces techniques peuvent être adaptées ou non, efficaces ou non, même si elles ne sont pas celles attendues dans les programmes, à un niveau scolaire donné. Dans CEDRE, le niveau d'activité mathématique (Grugeon-Allys, 2018) d'un élève relatif à un domaine mathématique est déterminé à partir des notions, propriétés et raisonnements qu'ils mobilisent dans la résolution des tâches des types de tâche de ce domaine, compte tenu des attendus de fin de cycle 4. C'est pourquoi, les techniques utilisées par les élèves pour résoudre une tâche informent sur leur activité mathématique au cours de la résolution de la tâche, et donc sur leurs acquis. C'est un moyen d'accéder, pour les élèves, à leur conceptualisation des notions d'un domaine mathématique ou d'un champ conceptuel (Vergnaud, 1986, 1990).

La complexité d'une tâche est une des variables didactiques qui permet de contraindre la nécessité d'utiliser une technique plutôt qu'une autre. La complexité d'une tâche au regard de l'activité mathématique visée lors de sa résolution est entendue au sens du niveau de mise en fonctionnement des connaissances et raisonnements à mobiliser pour sa résolution (Robert, 1998 ; Roditi & Salles, 2015). Cela concerne en particulier, le caractère nouveau ou ancien des contenus mathématiques en jeu, les étapes de résolution laissées à la charge de l'élève ou le niveau d'adaptation des connaissances mises en jeu. Plus la tâche est complexe, plus sa résolution nécessite un raisonnement « expert » pour un niveau donné.

D'un point de vue psychométrique, les items doivent vérifier deux indicateurs calculés a posteriori, leur difficulté et leur indice de discrimination (Vrignaud 2006). En effet, une fois le test CEDRE passé, la difficulté d'un item estime la proportion d'élèves ayant donné une réponse correcte tandis que l'indice de discrimination, estime « la quantité et la qualité d'informations apportées par l'item pour déterminer la compétence du sujet » (Rocher, 2015, p 43-44). La complexité d'une tâche informe sur le niveau de mise en fonctionnement des connaissances visées a priori, alors que la difficulté informe a posteriori sur la proportion d'élèves qui réussissent cette tâche et est utilisée dans la construction de l'échelle de score. En effet, les élèves sont répartis dans une échelle de score en six groupes à partir de leur réussite aux items. À chaque groupe d'élèves sont associés des items caractéristiques. Lorsqu'un item caractéristique est associé à un groupe d'élèves, cela signifie que les élèves de ce groupe ont une probabilité d'au moins 50 % d'avoir répondu correctement à cet item (Arzoumanian & Dalibard, 2015).

Pour un même type de tâche, le cadre doit donc permettre de concevoir des tâches de complexités différentes en attribuant des valeurs différentes aux variables didactiques les caractérisant afin d'accéder aux niveaux d'activité mathématique développés par les élèves sur un domaine donné. Ce choix permet de sélectionner a priori les items caractéristiques des groupes de l'échelle de scores pour chaque type de tâches.

Pour chaque sous-thème mathématique, trois niveaux d'activité mathématique ont été définis, ce qui permet d'associer à un item un niveau d'activité mathématique visé et de rendre compte d'un niveau de conceptualisation d'élèves relativement au champ conceptuel évalué. Le niveau A est celui attendu par l'institution, le niveau C relève d'un niveau de conceptualisation minimal, le niveau B est intermédiaire.

Un équilibre de proportion entre les items considérés selon leur complexité a priori est recherché. Les items des cinq thèmes sont pour certains identiques à ceux proposés en 2014 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...), calcul mental et items interactifs.

Les questions dites ouvertes ainsi que certains items interactifs 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 de consignes de correction strictes pour éviter toute subjectivité ou la validation de réponses trop imprécises ou succinctes. Les réponses sous format QCM sont prises en charge automatiquement par la plateforme de passation et les questions ouvertes et certains items interactifs ont été corrigés par des experts via une interface Internet. Les items interactifs donnant lieu à des types de données non structurés ou semi-structurés, des analyses complémentaires ont été nécessaires afin de traiter ces nouveaux items.

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.

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 (cf. plus loin l'encadré « GEODE »). L'édition, le stockage et la visualisation des items mais aussi la gestion des évaluations sont pris en charge par la plateforme d'évaluation (TAO).

1.3.2 Constitution des modules

L'évaluation Cedre mathématiques 2019 était composée de deux séquences de 50 minutes ainsi qu'une séquence de questionnaire de contexte de 30 minutes (tableau 4).

Une séquence contient une trentaine de questions de mathématiques. La première séquence commençait par des questions de calcul mental pendant lesquelles la calculatrice n'était pas autorisée. Les questions étaient de formats divers, majoritairement des questions à choix multiples et des réponses construites courtes. L'élève était ensuite orienté vers une nouvelle section, couvrant l'ensemble des domaines, parmi les 13 construites. La distribution des deux dernières sections de la séquence 1 se réfère à la méthodologie des cahiers tournants qui permet d'évaluer un nombre important d'items sans allonger le temps de passation. 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.

Dans cette première séquence, quatre questions proposaient à l'élève un environnement interactif lui permettant de mettre en oeuvre ses connaissances mathématiques, dans une tâche à prise d'initiative, notamment par l'utilisation d'outils numériques tels qu'un tableur ou un outil de géométrie dynamique.

La deuxième séquence était adaptative. Selon les réponses apportées à une section commune (starter), les élèves étaient orientés vers un niveau haut ou un niveau bas. Le module adaptatif portait seulement sur les domaines « nombres et calcul », « calcul littéral » et « géométrie ».

Tableau 4 – Design de l'évaluation CEDRE mathématiques 2019

Séquences	**Modules**	Sections		
Séquence 1	**19BMC01**	CM1	MA01	MA01
	19BMC02	CM2	MA02	MA02
	19BMC03	CM3	MA03	MA03
	19BMC04	CM4	MA04	MA04
	19BMC05	CM5	MA05	MA05
	19BMC06	CM6	MA06	MA06
	19BMC07	CM1	MA07	MA07
	19BMC08	CM2	MA08	MA08
	19BMC09	CM3	MA09	MA09
	19BMC10	CM4	MA10	MA10
	19BMC11	CM5	MA11	MA11
	19BMC12	CM6	MA12	MA12
	19BMC13	CM1	MA13	MA13
Séquence 2	**19BMCAD**	Starter	Niveau bas	Niveau haut
Séquence 3	**19BMCQQ**	Questionnaire de contexte		

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 Déroulement des épreuves

La passation de l'évaluation finale a eu lieu en mai 2019. Comme en 2014, 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 d'établissements.

Dans chaque établissement, une personne a été désignée comme étant l'administrateur du test, son rôle étant de veiller au strict respect de la procédure à suivre

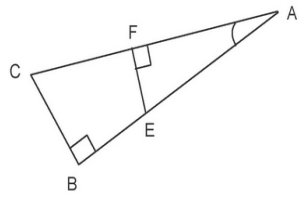
pour que l'évaluation soit passée dans les mêmes conditions quel que soit l'établissement. La collecte de l'information s'est faite via la plateforme informatique TAO en ligne (ordinateur-clavier-souris).

Chaque séquence était séparée par une pause de 5 minutes. Les deux premières séquences interrogeaient les élèves sur leurs connaissances et compétences en mathématiques alors que la troisième séquence était une partie de « contexte » permettant d'éclairer les réponses des élèves et de nuancer certaines différences de niveaux qui peuvent apparaître (notamment entre types d'établissements fréquentés).

Les professeurs de mathématiques de la classe ou des classes concernées ont également dû renseigner un questionnaire de contexte en ligne au moment de la passation des épreuves par les élèves. L'anonymat des élèves et des personnels a été respecté.

Une fois l'évaluation terminée, les données anonymisées ont été enregistrées. Aucun travail de correction n'a été demandé aux établissements.

FIGURE 1 – Exemple 1 : série de Vrai-Faux



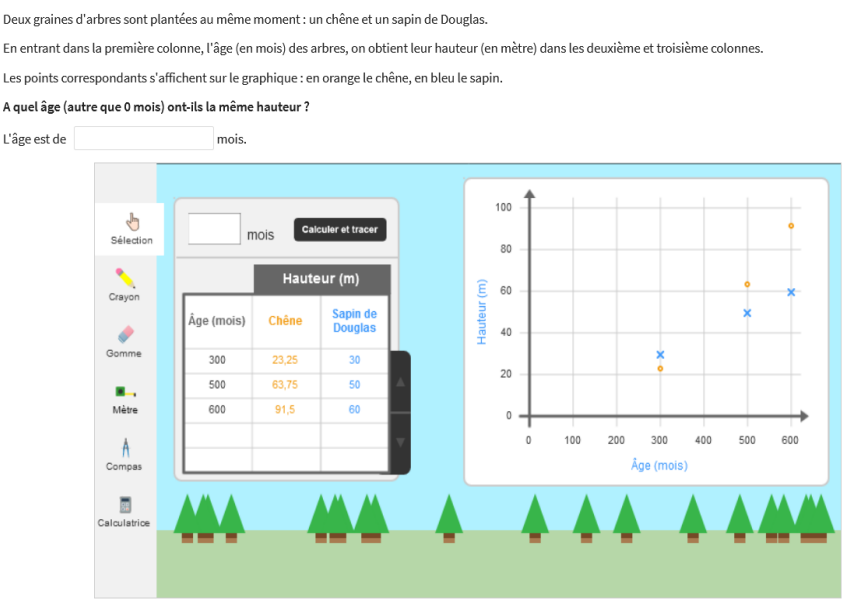
Le triangle ABC est rectangle en B, le triangle AEF est rectangle en F.

Cocher soit VRAI soit FAUX pour chacune des égalités suivantes.

Vous devez effectuer exactement 4 choix.

	Vrai	Faux
$tan\ \widehat{A} = \frac{BC}{AC}$	☐	☐
$tan\ \widehat{A} = \frac{AE}{EF}$	☐	☐
$tan\ \widehat{A} = \frac{BC}{AB}$	☐	☐
$tan\ \widehat{A} = \frac{EF}{AF}$	☐	☐

FIGURE 2 – Exemple 2 : Item interactif



2 Sondage

2.1 Méthodes

2.1.1 Tirage équilibré de classes de 3e

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

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

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

Cette propriété s'écrit :

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

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

Tirage d'établissements *versus* Tirage de classes

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

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

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

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

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

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

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

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

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

2.1.2 Redressement de la non réponse : calage sur marges

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

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

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

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

Tirage équilibré après élimination de la base des échantillons précédemment tirés

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

Formulation du problème et notations

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

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

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

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

Équilibrage

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

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

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

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

Ainsi, on aura :

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

Or, en espérance on a

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

La condition d'équilibrage initiale est donc remplie.

Condition fondamentale

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

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

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

2.1.3 Calcul de précision : méthode

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

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

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

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

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

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

2.2 Echantillonnage

Champ

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

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

Stratification

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

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

Modalités de sélection

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

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

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

Echantillon 2019

L'échantillon vise 8 000 élèves répartis proportionnellement selon les trois strates.

Base de sondage

Le tableau 5 présente les exclusions dans la population ciblée.

Tableau 5 – Exclusions pour la base de sondage - Mathématiques Collège

	Établissements	**Elèves**
Etab. accueillant des élèves de 3e	8 479	837 080
On retire les COM	8 439	832 669
On retire les étab hors contrat	8 206	829 542
On retire les EREA	8 139	828 331
On retire les UPE2A	8 127	827 183
On retire les ULIS	8 118	825 564
On ne garde que les collèges	6 929	797 445
On ne garde que les 3ème générales	6 925	771 958
Base CEDRE 3e	6 925	771 958

Le tableau 6 présente la répartition de la population ciblée selon le secteur d'enseignement.

Tableau 6 – Répartition dans la base de sondage - Mathématiques Collège

Strate	**Établissements**	**Élèves**
1.PU	4 193	483 586
2.REP	1 087	123 786
3.Privé	1 645	164 583
Total	**6 925**	**771 958**

Échantillon

Le tableau 7 présente la répartition de l'échantillon selon le secteur d'enseignement. Au total, 310 écoles ont été sélectionnées.

Tableau 7 – Répartition dans l'échantillon - Mathématiques Collège

Strate	**Établissements**	**Élèves**
1.PU	96	2 511
2.REP	68	2 525
3. Privé	35	1 084
Total	**199**	**6 120**

2.3 État des lieux de la non-réponse

2.3.1 Non-réponse totale

Parmi la non-réponse totale, nous distinguons la non-réponse des établissements de la non-réponse des élèves des établissements participants. Les chiffres suivants ont été observés pour 2019.

90 % des établissements de l'échantillon ont répondu à l'évaluation (tableau 8). 84.8 % des effectifs attendus ont participé (tableau 9).

Tableau 8 – Non-réponse des établissements - Mathématiques Collège

Strate	Nb établissements attendus	Nb établissements répondants	% d'établissements répondants
1.PU	114	107	93.9 %
2.REP	122	110	90.2 %
3. Privé	74	62	83.8 %
Total	**310**	**279**	**90 %**

Tableau 9 – Non-réponse des élèves - Mathématiques Collège

Strate	Nb élèves attendus	Nb élèves répondants	% d'élèves répondants
1.PU	3 001	2 722	90.7 %
2.REP	2 962	2 423	81.8 %
3. Privé	2 032	1 633	80.4 %
Total	**7 995**	**6 778**	**84.8 %**

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 module.
— La non-réponse terminale : un élève s'est arrêté avant la fin du module 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 les dernières réponses d'un élève sont vides alors il n'a pas eu le temps de les voir. La non réponse terminale a été étudiée par séquence et par module. Toutes les observations en non-réponses terminales sont enlevées des données

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

— 1.2 items pour la séquence 1
— 2.2 items pour la séquence 2

2.4 Redressement

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

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

TABLEAU 10 – Comparaison entre les marges de l'échantillon et les marges dans la population - Mathématiques Collège

Variable	Modalité	Marge pré-calage	Marge post-calage	Marge population	Pourcentage pré-calage	Pourcentage post-calage	Pourcentage population
Retard	1	82 372	97 408	97 408	10,7	12,6	12,6
	2	689 586	674 550	674 550	89,3	87,4	87,4
Sexe	1	382 636	386 576	386 576	49,6	50,1	50,1
	2	389 322	385 382	385 382	50,4	49,9	49,9
Strate	1	483 589	483 589	483 589	62,6	62,6	62,6
	2	123 786	123 786	123 786	16	16	16
	3	164 583	164 583	164 583	21,4	21,4	21,4

2.5 Précision

L'erreur standard (*se*) peut être calculée sur le score moyen de chaque année (tableau 11).

TABLEAU 11 – Scores moyens et erreurs standard associées - Mathématiques Collège

Année	Score moyen	Erreur standard
2008	250	1.6
2014	243	1.65
2019	236.8	1.47

Pour savoir par exemple si l'évolution entre 2014 et 2019 est significative , il faut calculer la valeur suivante :

$$\frac{\left|\hat{\bar{Y}}_{2019} - \hat{\bar{Y}}_{2014}\right|}{\sqrt{se^2_{\hat{\bar{Y}}_{2019}} + se^2_{\hat{\bar{Y}}_{2014}}}} \tag{3}$$

Entre 2014 et 2019, on obtient une valeur de 2.8 (supérieure à 1.96). Cela signifie que l'évolution du score moyen est statistiquement significative.

Les erreurs standards sont également calculées pour les répartitions dans les différents groupes de niveaux (tableaux 12 et 13).

Tableau 12 – Répartitions en % dans les groupes de niveaux - Mathématiques Collège

Année	Groupe <1	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5
2008	2.2	12.8	26.8	29.6	18.6	10
2014	3.6	15.9	27.8	28.3	15.3	9.1
2019	6.4	18.2	28.2	25	14.3	7.9

Tableau 13 – Erreurs standards des répartitions en % dans les groupes de niveaux - Mathématiques Collège

Année	Groupe <1	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5
2008	0.4	0.8	0.9	1	1	0.7
2014	0.4	0.9	0.9	0.8	0.7	0.7
2019	0.6	0.7	0.7	0.7	0.6	0.5

Design effect

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

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

Tableau 14 – Effet du plan de sondage - Mathématiques Collège

Année	Erreur Standard	Erreur SAS	*Design Effect*
2008	1.85	0.56	10.87
2014	0.92	0.51	3.23
2019	1.4	0.75	3.49

Dans le cas d'un sondage en grappes, la précision est dégradée en comparaison d'un sondage aléatoire simple. Cela signifie qu'en 2019, un sondage aléatoire simple avec un effectif 3.49 fois moins important aurait conduit au même niveau de précision.

3 Analyse des items

3.1 Méthodologie

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

3.1.1 Approche classique

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

Réussite et score

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

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

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

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

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

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

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

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

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

Fidélité

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

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

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

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

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

Indices de discrimination

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

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

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

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

En effet, on peut montrer que

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

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

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

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

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

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

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

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

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

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

3.2 Codage des réponses aux items

3.2.1 Valeurs manquantes

Trois types de valeurs manquantes sont distinguées :

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

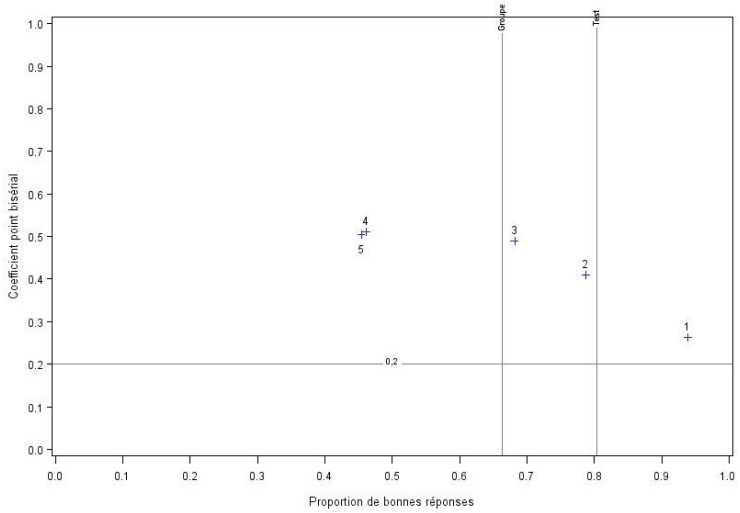
— Non-réponse terminale : l'élève s'est arrêté au cours de l'épreuve, potentiellement en raison d'un manque de temps. Des choix sont effectués pour déterminer le traitement de ces valeurs. Nous considérons que si les dernières réponses d'un élèves sont vides alors il n'a pas eu le temps de les voir. Toutes les observations en non-réponses terminales sont enlevées des données. Les valeurs manquantes sont alors traitées de manière structurelle. Sinon, elles sont traitées comme des échecs.

3.2.2 Regroupement des items

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

Le regroupement de ces items consiste à faire la somme des indicatrices de réussite et à déterminer un seuil de maîtrise. Une visualisation graphique est utilisée pour fixer les scores « seuils » (cf. figure 3). Ce graphique représente le taux de réussite pour chaque seuil possible en fonction de la discrimination obtenu pour le seuil. Il permet de choisir la combinaison la mieux adaptée. Le score seuil doit préserver la discrimination de l'item regroupé et la difficulté peut être modulée en fonction des objectifs.

FIGURE 3 – Représentation graphique utilisée pour le regroupement d'items



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

3.2.3 Traitement des données et correction des questions ouvertes

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

AGATE : un outil de correction à distance des questions ouvertes

Objectifs

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

Principes fonctionnels

Le chef de projet paramètre la session de correction. Il définit les groupes de correcteurs et supervise chaque groupe. Il intègre et vérifie les items mis en correction et ajuste les paramètres de double correction. Son rôle consiste également à répondre aux questions des correcteurs par le biais d'une messagerie intégrée au logiciel et à communiquer sa réponse également aux autres correcteurs. Le superviseur gère son groupe de correcteurs. Il anime la session de formation, qui consiste d'une part à communiquer aux télécorrecteurs une grille de correction très précises et d'autre part à corriger collectivement à blanc un nombre défini d'imagettes pour s'assurer de la compréhension et de la bonne mise en oeuvre des consignes. Puis, pendant la télécorrection, il arbitre les litiges lors des doubles-corrections. Le correcteur corrige les items en portant un codage de réussite/erreur sur chaque item. En cas de doute, il peut se référer à son superviseur de groupe. Une messagerie interne complète le dispositif et permet un échange de point de vue entre les différents acteurs.

3.3 Résultats

3.3.1 Pouvoir discriminant des items

9 items ont été éliminés pour cause de *rbis-point* trop faible :

— 9 items de 2019

4 Modélisation

4.1 Méthodologie

4.1.1 Modèle de réponse à l'item

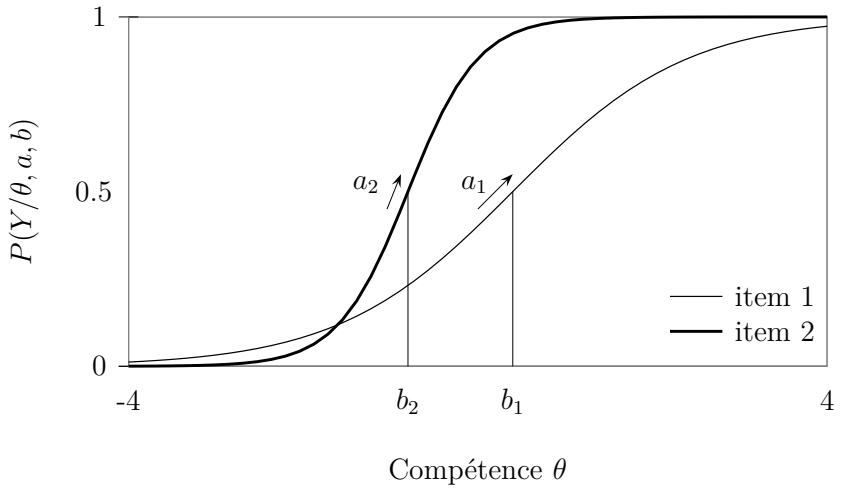
Le modèle de mesure utilisé est un modèle de réponse à l'item à deux paramètres avec une fonction de lien logistique (MRI 2PL) :

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

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

FIGURE 4 – 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 4).

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 [5], la fonction de vraisemblance s'écrit :

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

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

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

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

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

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

avec g la densité de θ.

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

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

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

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

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

Estimation des niveaux de compétence

Une fois les paramètres des items estimés, ils sont considérés comme fixes et il est possible d'estimer les θ_i, par exemple *via* la maximisation de la vraisemblance donnée par l'équation (12).

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

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

avec

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

et

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

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

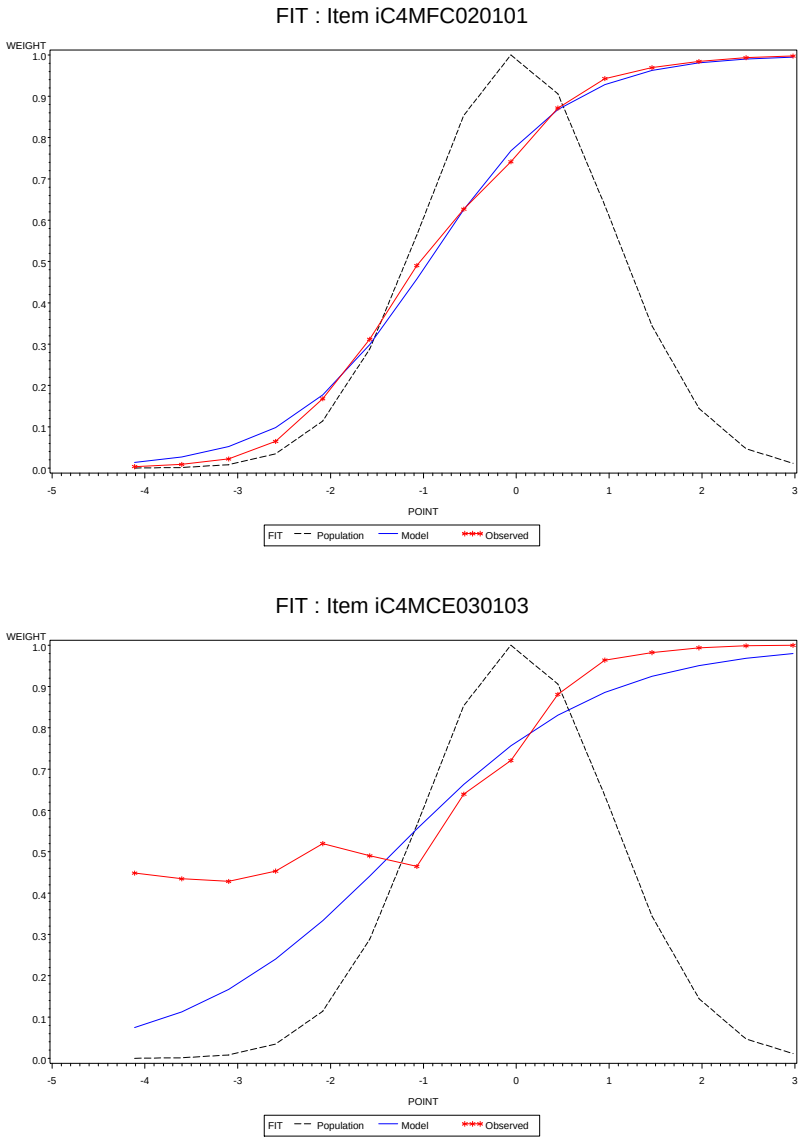
$$\frac{\partial lnL}{\partial \theta_i} + \frac{J}{2I} = 0 \tag{16}$$

4.1.3 Indice d'ajustement (FIT)

L'ajustement des items au modèle est étudié. Graphiquement, cela revient à comparer les courbes caractéristiques estimées avec les résultats observés (cf. figure 5). 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 5 – Exemples d'ajustements (FIT)



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

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

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

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

4.1.4 Fonctionnement Différentiel d'Item (FDI)

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

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

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

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

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

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

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

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

4.1.5 L'information du test

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

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

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

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

4.1.6 Transition papier-numérique : étude de comparabilité

Contexte

L'enquête CEDRE est une série temporelle, c'est-à-dire qu'elle a pour objectif premier de pouvoir comparer les performances des élèves de cycle en cycle. Cette caractéristique implique que les différentes générations de l'enquête soient comparables et que le construit testé à chaque cycle soit donc identique.

La DEPP s'est engagée dans la transition d'enquêtes réalisées sur papier vers des enquêtes au format numérique. Cette transition offre de nombreux avantages, aussi bien sur le plan technique qu'en termes de potentialités d'études. Toutefois, la modification du mode d'administration des items aux élèves ne va pas sans poser certaines questions d'ordre méthodologique, qui peuvent mettre en péril la comparabilité des résultats entre les cycles.

Objectifs

Pour assurer cette comparabilité, la Théorie de Réponse à l'Item fournit un ensemble d'outils mé-thodologiques robustes. L'enquête CEDRE s'appuie notamment sur l'utilisation d'items dits d'ancrage, c'est-à-dire repris à l'identique d'un cycle sur l'autre. Ce sont ces items qui permettent de mettre sur la même échelle de performance les résultats des élèves des différents cycles.

Toutefois, la théorie psychométrique impose un certain nombre de contraintes pour que son usage soit pertinent. Une de ces contraintes, essentielle, est l'invariance locale des items. Autrement dit, chaque item doit mesurer le même trait latent, et avec la même précision, pour l'ensemble des sujets, quel que soit son cycle.

Comparer les élèves évalués en 2019 avec ceux des cohortes précédentes ne pouvait donc se faire que sous l'hypothèse que les items restaient parfaitement identiques (notamment en termes de difficulté) lors de leur changement de mode (passage du papier au numérique).

Il était donc nécessaire de construire une cohorte intermédiaire, soumise à une enquête au format mixte, à la fois papier et numérique, servant de "pont" entre les cycles au format papier et les cycles au format numérique.

Méthodologie

L'étude de comparabilité effectuée en 2018 était composée d'items de 2014, repris à l'identique, permettant une comparaison diachronique et leur transposition au format numérique. Ces items étaient répartis en deux cahiers papier et deux modules numériques.

L'échantillon a été construit selon la même méthodologie que pour toutes les enquêtes Cedre, à savoir un tirage équilibré de classes de 3ème. Ce tirage est stratifié selon la nature de l'établissement (public, privé, éducation prioritaire), et équilibré selon le sexe et le retard (étant considérés "en retard" les élèves ayant redoublé au moins une fois).

Cette enquête de comparabilité, ou "bridge study", était essentiellement définie par deux choses : d'une part, le design qui a présidé à sa construction, et d'autre part, les hypothèses statistiques qui sous-tendaient ce design.

Lorsqu'on parle de design expérimental, il s'agit à la fois de déterminer le choix des items qui constitueront l'enquête, mais aussi le plan de rotation, c'est-à-dire quels items seront vus par quels élèves.

En ce qui concerne les items, l'ensemble des items d'ancrage ont été repris sous leurs deux formats, papier et numérique. Comme un élève ne peut pas rencontrer deux fois le même item, sans quoi l'effet d'apprentissage serait incontestable, ceux-ci ont été répartis en deux groupes d'items A et B.

Les élèves ont également été répartis dans deux groupes 1 et 2. Les élèves du groupe 1 se sont vu soumettre les items du groupe A au format papier et les items du groupe B au format numérique, tandis que les élèves du groupe 2 se sont vu soumettre les items du groupe A au format numérique et les items du groupe B au format papier. Ainsi, les difficultés des items dans leurs deux modes (papier et numérique) ont pu être calculées distinctement. L'écart de difficulté entre les versions papier et numérique des items (aussi appelé "effet mode") a ensuite été reporté sur la passation CEDRE 2019, afin de la rendre comparable avec les cohortes précédentes.

Précisons que le choix de répartition des élèves au sein de chaque groupe s'est fait au niveau de la classe de manière aléatoire. Cette consigne stricte est régie par la théorie psychométrique. En effet, en construisant deux designs distincts comme nous l'avons fait, rien ne permet a priori de dire que les deux échelles de performance seront équivalentes. Pour cela, il faut que les deux échantillons d'élèves soient représentatifs de la même population.

Effets fixes et effets aléatoires

En théorie, le tirage aléatoire de deux sous-échantillons au sein d'une même population sont également représentatifs de la population. Malheureusement, la méthode de tirage (tirage équilibré par strate) ne garantit pas le côté totalement aléatoire. On peut distinguer les biais subis par le plan de sondage entre effets communs et effets distincts aux deux groupes. Les effets fixes correspondent à la variabilité interclasse, c'est-à-dire aux biais de sondage qui pèsent de manière identique sur les deux sous-échantillons. Les effets distincts corres-pondent à la variabilité intra-classe, c'est-à-dire au biais créé par la scission de chaque classe en deux sous-groupes distincts.

Variabilité interclasse

Les biais de sondage liés à la variabilité interclasse sont identiques à ceux constatés lors des cycles précédents. Ils peuvent être corrigés par une repondération adaptée (calage sur marges), et sont pris en compte lors des calculs de précision. De plus, ils n'impactent pas les deux sous-groupes de l'étude de comparabilité, puisque ceux-ci les portent de manière identique (après repondération). Ils s'apparentent à des effets fixes, et seront donc traités comme tel.

Principes fonctionnels

Cette méthode présente deux avantages. Le tirage aléatoire simple minimise le biais de sélection pour chacun des sous-groupes, qui ne dépend plus que de la taille de chaque sous-groupe. De plus, il limite les effets aléatoires aux individus, rendant ainsi fixes les effets portés par les variables de niveau supérieur (classe, établissement, ...).

4.2 Résultats

4.2.1 Identification des fonctionnements différentiels d'items (FDI)

2 items ont été éliminés des analyses pour cause de fonctionnements différentiels.

— 2 items d'ancrage 2014-2019

4.2.2 Bilan de l'analyse des items

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

— 79 items de 2008
— 73 items de 2014

— 263 items de 2019
— 46 items d'ancrage 2008-2014
— 41 items d'ancrage 2014-2019
— 22 items d'ancrage 2008-2014-2019

Cela représente 524 items passés par les élèves en tout, dont 326 en 2019.

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

— 79 items de 2008
— 73 items de 2014
— 263 items de 2019
— 43 items d'ancrage 2008-2014
— 26 items d'ancrage 2014-2019
— 19 items d'ancrage 2008-2014-2019

503 items sont donc conservés dans l'analyse, dont 308 utilisés dans l'évaluation 2019.

4.3 Calcul des scores

Comme indiqué précédemment, une analyse conjointe des données des 3 années a permis d'estimer les paramètres des items, puis les niveaux de compétences θ des élèves. Afin de lever l'indétermination du modèle, la moyenne des θ a été fixé à 250 et leur écart-type à 50, pour l'échantillon de 2008. Le tableau 15 présente les résultats obtenus.

Tableau 15 – Niveaux de compétences (moyennes des scores et écarts-types) Mathématiques Collège

Année	Score moyen	Écart-type
2008	250	50
2014	243	50.3
2019	236.8	54

5 Construction de l'échelle

5.1 Méthode

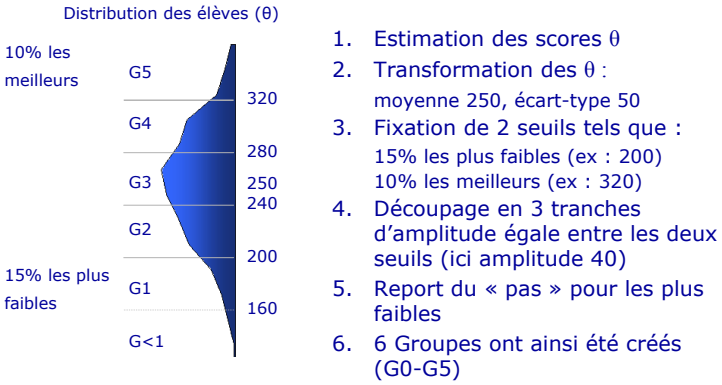
Les modèles de réponse à l'item permettent de positionner sur une même échelle les paramètres de difficulté des items et les niveaux de compétences des élèves.

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

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

FIGURE 6 – Principes de construction de l'échelle



5.2 Caractérisation des groupes de niveaux

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

Groupe < 1 (6,4 % 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 début de cycle 3).

Groupe 1 (18,2 % 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 (28,2 % des élèves)

Les élèves du groupe 2 possèdent de réelles compétences pour **réaliser des calculs sur les nombres entiers et décimaux relatifs**. Ils peuvent résoudre un problème impliquant des nombres décimaux dans un environnement numérique interactif et représenter un nombre décimal dans différents registres. La maîtrise des programmes de calcul est également très satisfaisante. Ils parviennent à en proposer les expressions littérales associées. 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. En géométrie, ils peuvent mettre en relation un programme de construction avec une figure et reconnaître une symétrie axiale.

Groupe 3 (28,3 % des élèves)

Les élèves du groupe 3 peuvent conduire des **raisonnements à une étape déductive**. Le calcul sur les puissances de 10 est acquis. Leurs aptitudes à **réaliser des calculs algébriques** et à mettre une situation en équation sont étendues. Ils sont capables de développer une expression algébrique simple, de la factoriser en utilisant la distributivité de la multiplication par rapport à l'addition ainsi que de reconnaître une somme ou un produit. En outre, ils utilisent la proportionnalité comme un outil permettant de résoudre les problèmes et savent utiliser les grandeurs composées. En géométrie, ils savent **mettre en œuvre certains théorèmes** du programme dans des cas simples. Le calcul d'aire par dénombrement d'unités et la conversion de durées entre les systèmes sexagésimal et décimal sont acquis. Enfin, en algorithmique, ils savent interpréter le langage de programmation par blocs dans des cas simples incluant variable, boucle et condition.

Groupe 4 (14,3 % 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 et les puissances. Dans le domaine des fonctions, ils comprennent le formalisme f(a)=b et la notion d'image et d'antécédent. Ils établissent des liens entre différentes représentations d'une fonction, notamment dans un environnement numérique interactif. L'interprétation de programme par blocs plus complexes, contenant notamment une fonction, est acquis.

Groupe 5 (7,9 % 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. En géométrie, les seules transformations acquises sont les symétries et la translation. Ils sont capables de résoudre un problème à l'aide des nombres en écriture fractionnaire ou les puissances. Enfin, les notions sur les fonctions sont mieux comprises et exploitées par ces élèves, par exemple pour re-

présenter dans une forme algébrique une grandeur géométrique dépendant d'une variable.

5.3 Exemples d'items

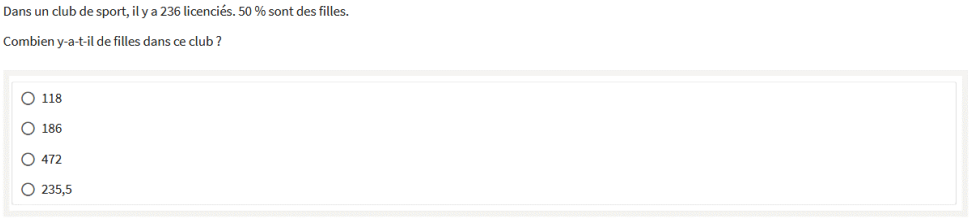
5.3.1 Item caractéristique du groupe < 1

FIGURE 7 – Exemple groupe <1

Calculer un événement le plus probable dans un contexte familier en extrayant l'information.

5.3.2 Item caractéristique du groupe 1

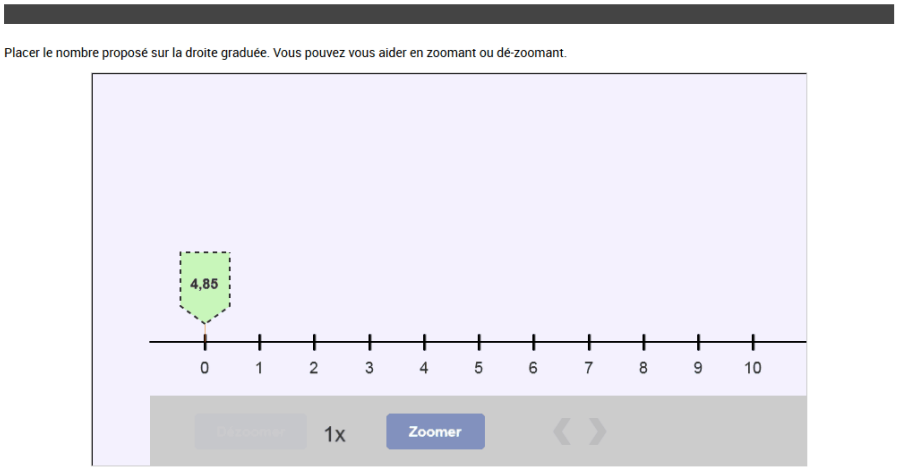
FIGURE 8 – Exemple groupe 1



Calcul automatisé d'un pourcentage.

5.3.3 Item caractéristique du groupe 2

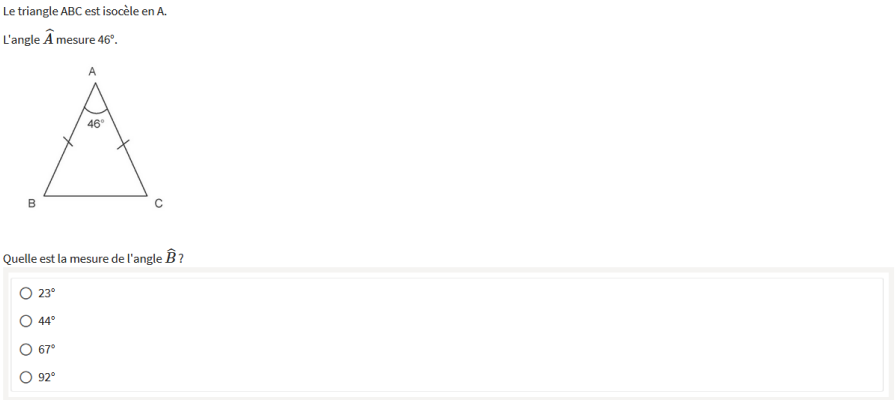
FIGURE 9 – Exemple groupe 2



Placer un nombre décimal sur une droite graduée avec la possibilité de zoomer sur la droite.

5.3.4 Item caractéristique du groupe 3

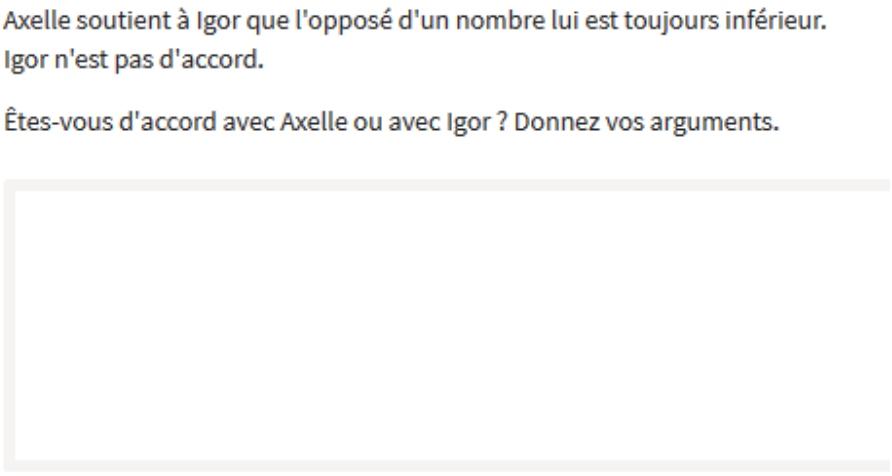
FIGURE 10 – Exemple groupe 3



Utiliser un théorème de géométrie (ici la somme des angles dans un triangle)

5.3.5 Item caractéristique du groupe 4

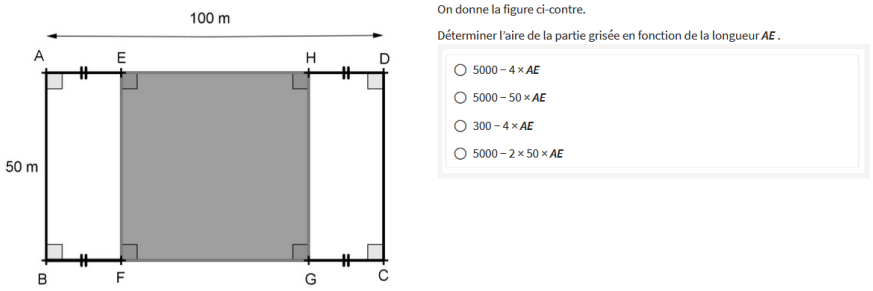
FIGURE 11 – Exemple groupe 4



Produire un raisonnement en citant éventuellement un contre-exemple pour invalider un énoncé.

5.3.6 Item caractéristique du groupe 5

FIGURE 12 – Exemple groupe 5



Exprimer une grandeur géométrique en utilisant une expression algébrique dépendant d'une variable.

6 Variables contextuelles et non cognitives

6.1 Variables sociodémographiques et indice de position sociale

Un certain nombre de variables sociodémographiques permettent d'enrichir l'analyse des résultats. Le score moyen des élèves est ainsi analysé en fonction du genre, du retard scolaire et quand les effectifs le permettent en fonction du secteur d'enseignement. Le lecteur est invité à consulter la Note d'Information pour plus de détails (Salles, Dos Santos, & Ninnin, 2019).

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

Il n'a été possible d'établir des comparaisons qu'en termes de niveau social des écoles, et non au niveau individuel. En effet, en 2019, la PCS des parents est disponible pour chaque élève, mais elle ne l'était pas dans les cycles antérieurs. Pour chaque établissement des échantillons de 2008, 2014 et 2019, 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 6.1).

Indice moyen de l'etab.	Année	Score moyen	Ecart type
Groupe 1 (25% les plus défavorisés)	2008	228	47
	2014	**218**	45
	2019	220	**50**
Groupe 2	2008	246	47
	2014	242	48
	2019	**231**	**52**
Groupe 3	2008	257	47
	2014	**244**	46
	2019	242	**54**
Groupe 4 (25% les plus favorisés)	2008	268	51
	2014	268	50
	2019	**254**	**54**

Note de lecture : en 2019, le score moyen des élèves appartenant au quart des collèges les plus favorisées (Groupe 4) diminue de 14 points par rapport à 2008. Les évolutions significatives sont indiquées en gras.

6.2 Élaboration des questionnaires de contexte

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

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

Le questionnaire élève interroge des dimensions dites conatives intéressantes à mettre en lien avec le score obtenu à l'épreuve - temps de travail personnel estimé par l'élève, type de travail personnel le plus demandé, pratiques culturelles en lien avec les disciplines évaluées... De plus, les élèves sont demandés d'évaluer la difficulté de l'épreuve et leur degré d'implication à faire le test.

6.3 Motivation des élèves face à la situation d'évaluation

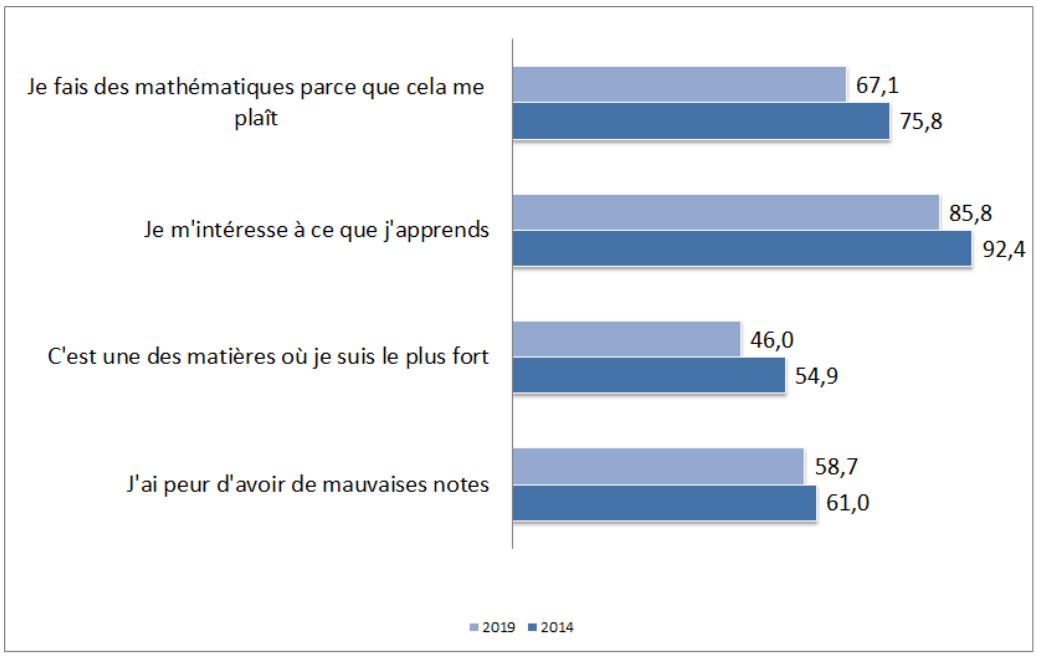
En 2019, la grande majorité des élèves se déclarent anxieux vis-à-vis des notes. Ainsi, 74,2% d'entre eux sont d'accord avec l'affirmation suivante : « Je m'inquiète à l'idée d'avoir de mauvaises notes en mathématiques » . En 2014, ils étaient 72% (13). La nervosité associée à la résolution de problèmes concerne

une proportion moins importante d'élèves (38,1%), mais est plus affirmée qu'en 2014 (33,8%).

En 2019, les formes d'apprentissage en collaboration sont encore très appréciées des élèves. Ainsi, 77,4 % affirment aimer travailler en groupe avec d'autres élèves (80,4% en 2014) et 58,9% déclarent qu'ils apprennent mieux en mathématiques en travaillant avec d'autres élèves de leur classe (62,4% en 2014).

Enfin, les élèves gardent une image positive de la discipline. Ils sont 71,2% à être d'accord avec l'affirmation selon laquelle les mathématiques sont une matière importante parce qu'elles sont nécessaires pour de futures études (68% en 2014).

FIGURE 13 – Rapport des élèves aux mathématiques



Note de lecture : 58,9 % des élèves répondants en 2019 déclarent être « D'accord » ou « Tout à fait d'accord » avec l'affirmation « J'apprends mieux en mathématiques quand je travaille avec d'autres élèves de ma classe », contre 62,4 % en 2014.

7 Annexe

Certification AFNOR pour les évaluations CEDRE

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

Les finalités de la certification

Les finalités sont les suivantes :

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

Les enjeux pour la DEPP

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

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

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

La démarche qualité

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

Les engagements de service

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

Les engagements de service de la DEPP

Des objectifs clairs et partagés

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

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

Des évaluations fondées sur l'expertise pédagogique

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

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

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

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

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

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

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

Une mesure fiable et des comparaisons temporelles pertinentes

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

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

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

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

Des analyses enrichies par des données de contexte

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

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

Transparence des méthodes et partage des résultats

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

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

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

Références

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

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

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.

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

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

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

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

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

Salles, F., Dos Santos, R., & Ninnin, L.-M. (2019). CEDRE 2008-2014-2019 - mathématiques en fin d'école : des résultats en baisse. *Note d'information*, *16*.

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 nº3*, 427-450.

Liste des tableaux

1 Cycle des évaluations disciplinaires réalisées sur échantillons (CEDRE) depuis 2003 5
2 Thèmes, sous-thèmes et répartition des pourcentages cibles des items d'évaluation dans les différents thèmes et sous-thèmes . . . 6
3 Objectifs d'évaluation et types de tâches évalués pour le sous-thème « Utiliser le calcul littéral » (A3) 9
4 Design de l'évaluation CEDRE mathématiques 2019 13
5 Exclusions pour la base de sondage - Mathématiques Collège . . 24
6 Répartition dans la base de sondage - Mathématiques Collège . . 24
7 Répartition dans l'échantillon - Mathématiques Collège 25
8 Non-réponse des établissements - Mathématiques Collège 25
9 Non-réponse des élèves - Mathématiques Collège 25
10 Comparaison entre les marges de l'échantillon et les marges dans la population - Mathématiques Collège 27
11 Scores moyens et erreurs standard associées - Mathématiques Collège 27
12 Répartitions en % dans les groupes de niveaux - Mathématiques Collège 28
13 Erreurs standards des répartitions en % dans les groupes de niveaux - Mathématiques Collège 28
14 Effet du plan de sondage - Mathématiques Collège 29
15 Niveaux de compétences (moyennes des scores et écarts-types) Mathématiques Collège 48

Table des figures

1 Exemple 1 : série de Vrai-Faux 15
2 Exemple 2 : Item interactif 16
3 Représentation graphique utilisée pour le regroupement d'items . 35
4 Modèle de réponse à l'item - 2 paramètres 38
5 Exemples d'ajustements (FIT) 42
6 Principes de construction de l'échelle 50
7 Exemple groupe <1 52
8 Exemple groupe 1 52
9 Exemple groupe 2 53
10 Exemple groupe 3 53
11 Exemple groupe 4 54
12 Exemple groupe 5 54
13 Rapport des élèves aux mathématiques 57