CEDRE

Cycle des Évaluations Disciplinaires Réalisées sur Échantillons

Rapport technique

Sciences expérimentales 2018

Collège

Auteurs :

Anaïs BRET
Reinaldo DOS SANTOS
Saskia KESKPAIK
Marion LE CAM
Jessica NADEAU
Louis-Marie NINNIN
Thierry ROCHER
Léa ROUSSEL
Ronan VOURC'H

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

Septembre 2020

Table des matières

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

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

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

Ces évaluations apportent un éclairage qui intéresse tous les niveaux du système éducatif, des décideurs aux enseignants sur le terrain, en passant par les formateurs d'enseignants : elles informent sur les compétences et les connaissances des élèves à la fin d'un cursus, elles éclairent sur l'attitude et la représentation des élèves à l'égard de la discipline ; elles interrogent les pratiques d'enseignement au regard des programmes ; elles contribuent à enrichir la réflexion générale sur l'efficacité et la performance de notre système éducatif. Ces évaluations étant proposées à des échantillons statistiquement représentatifs de la population scolaire de France métropolitaine, aucun résultat par élève ne peut être calculé.

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

1.2 Connaissances et compétences visées

L'évaluation CEDRE Sciences collège a pour but de faire le point sur les connaissances et les compétences des élèves en physique-chimie et en sciences de la vie et de la Terre à la fin du collège (cycle 4).

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

Les connaissances et les compétences telles qu'elles sont définies dans les programmes officiels constituent une partie du cadre de cette évaluation.

1.2.1 Les programmes

En mai 2018, les élèves de troisième ont été évalués à la fois sur les programmes de 2008 et de 2016 en SVT et en physique-chimie.

	Programmes de 2008		**Programmes de 2016 (cycle 4)**	
	SVT	**Physique-chimie**	**SVT**	**Physique-Chimie**
Sixième	x			
Cinquième	x	x		
Quatrième			x	x
Troisième			x	x

1.2.2 le cadre d'évaluation

Les programmes ayant subi des modifications, il a été décidé de modifier le cadre d'évaluation de CEDRE Sciences. Ce cadre est articulé autour de cinq composantes :

— Le thème abordé
— Le type de connaissance
— La compétence principale mise en jeu
— La complexité de la tâche
— Le contexte

1.2.2.a Les thèmes

Chaque item est indexé suivant le thème qu'il aborde dans les programmes de 2016.

Les thèmes abordés en SVT

- La planète Terre, l'environnement et l'action humaine
- Le vivant et son évolution
- Le corps humain et la santé

Les thèmes abordés en Physique-Chimie

- Organisation et transformations de la matière
- Mouvement et interactions
- L'énergie et ses conversions
- Des signaux pour observer et communiquer

1.2.2.b Les types de connaissances

Chaque item est indexé suivant le type de connaissances qu'il évalue :

- Connaissances notionnelles : C'est connaître des concepts scientifiques fondamentaux et des théories explicatives et s'approprier des situations de la vie réelle en mobilisant ces concepts et ces théories à bon escient. Ces connaissances sont choisies dans les grandes thématiques au programme des cycles 3 et 4 de SVT et de Physique-chimie.
- Connaissances procédurales : C'est connaître les concepts et les méthodes, essentiels aux démarches scientifiques, utilisés pour collecter des données fiables, les traiter, valider les méthodes et les résultats.
- Connaissances épistémiques : C'est connaître les caractéristiques générales des savoirs scientifiques et les processus de construction des connaissances scientifiques ainsi que les rôles des sciences dans la société.

1.2.2.c Les compétences

Chaque item est indexé suivant la compétence principale mise en jeu. Les compétences évaluées sont issus du socle commun de connaissances, de compétences et de culture de 2016.

Maitriser les connaissances attendues

Pratiquer des démarches scientifiques

- Identifier ou formuler une question scientifique.
- Proposer une ou des hypothèses pour répondre à une question scientifique.
- Concevoir une expérience ou un protocole expérimental pour tester une ou des hypothèses.
- Mettre en oeuvre un protocole expérimental.
- Utiliser des instruments d'observation, de mesures ou des techniques de préparation.
- Interpréter des résultats et/ou conclure (répondre à une problématique, valider ou infirmer une hypothèse).
- Distinguer une relation de cause à effet d'une relation de corrélation.
- Identifier ou développer des modèles simples pour expliquer des faits d'observations.
- Communiquer sur ses démarches, ses résultats ou ses choix, en argumentant.
- Faire preuve d'esprit critique.

Pratiquer des langages

- Extraire des données, des informations, des résultats présentés sous différentes formes (tableau, graphique, diagramme, dessin, schéma, carte, image animée ...)
- Représenter des données, des informations, des résultats sous forme de tableau, graphique, diagramme, dessin, schéma, carte ...
- Passer d'un mode de représentation à un autre.
- Exploiter des simulations ou des modélisations numériques.
- Utiliser le vocabulaire spécifique et /ou les connecteurs logiques adaptés à la situation.
- Calculer.

Adopter un comportement éthique et responsable

- Connaître et/ou appliquer les règles de sécurité.
- Identifier les impacts des activités humaines sur l'environnement, en matière de santé et de préservation de ressources de la planète.
- Reconnaître et/ou distinguer les responsabilités individuelles et collectives.
- Distinguer ce qui relève d'une croyance ou d'une idée et ce qui constitue un savoir scientifique.

Se situer dans l'espace et le temps

- Appréhender différenst échelles de temps.
- Appréhender différentes échelles d'espace.

1.2.2.d La complexité de la tâche

Chaque item est indéxé suivant la complexité de la tâche demandée à l'élève, c'est-à-dire le nombre d'étapes cognitives que l'élève doit mettre en jeu pour répondre à la question posée.

1.2.2.e Le contexte

Chaque item est également indexé en fonction du contexte de la question :

- Personnel
- Scolaire
- Local-National
- Mondial

1.3 Construction du test

Le bureau de l'évaluation des élèves de la DEPP élabore des évaluations par disciplines et niveaux scolaires. La préparation des unités et de leurs constituants fait intervenir des concepteurs, généralement des enseignants. La coordination est assurée par un chef de projet, membre de l'équipe du bureau de l'évaluation des élèves. Une application dédiée leur permet de créer, modifier ou éditer leur unité ; en outre cette application permet au chargé d'étude de gérer l'ensemble de l'évaluation (cf. plus loin l'encadré « GEODE »). Le passage au numérique des évaluations CEDRE a été opéré pour la première fois sur la discipline Sciences collège. Une plateforme (TAO) a donc été mise en place pour permettre la création des items, la passation de l'évaluation par les élèves et le stockage des données qui sont ensuite analysées.

1.3.1 Elaboration des items

1.3.1.a Conception

Les unités composées d'un ensemble d'items sont le fruit d'un travail collectif des concepteurs, encadré par le chargé d'étude, 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 jusqu'à aboutir à un consensus. Une fois validé par le chargé d'étude et l'inspection, l'item fait l'objet d'un cobayage, c'est-à-dire d'une passation auprès d'une ou plusieurs classes pour estimer sa difficulté, les durées de passation minimale, maximale et moyenne, et receuillir les réactions éventuelles des élèves.

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).

1.3.1.b Formats d'items

Deux types de formats de questions sont utilisés : les questions fermées (QCM, glisser/déposer, ordonner, associer ...) et les questions ouvertes appelant une réponse écrite (réponse courte : un mot, un groupe de mot, un nombre - réponse longue : réponse rédigée, argumentée). Un entrainement est prévu au début de la passation afin que l'élève se familiarise avec les différents types de questions rencontrées et à la plateforme de passation.

Les items aux formats "tableau série" attendent une réponse de l'élève, pour chaque ligne du tableau. Les réponses aux différentes propositions relevant d'un mÃªme item font alors l'objet d'un regroupement pour lequel il faut définir un seuil de validation de l'item avec les concepteurs. Autrement dit, selon le niveau de difficulté voulu pour l'item concerné, on détermine pour sa validation, au cas par cas, qu'il faut des réponses correctes pour l'ensemble des propositions ou pour un nombre précis d'entre elles. Pour ce qui est des items d'ancrage, ce seuil reste identique.

1.3.1.c Correction automatisée ou experte

Les réponses aux formats "questions fermées" sont corrigées de manière automatisée, alors que les réponses aux formats "questions ouvertes" sont corrigées par des experts. Cela suppose la mise en place d'un dispositif de correction, nécessitant la formation technique des correcteurs et l'élaboration de grilles de correction précises déclinant les critères de réussite pour éviter toute subjectivité ou la validation de réponses trop imprécises ou trop succinctes. Ce dispositif de correction s'appuie sur le logiciel AGATE (cf encadré "AGATE" p.32).

1.3.1.d Capacités expérimentales

La spécificité de l'évaluation des sciences dans CEDRE, par rapport aux évaluations internationales et aux évaluations nationales d'autres pays, est la prise en compte des capacités expérimentales des élèves. Une épreuve de travaux pratiques est proposée à huit élèves par classe échantillonnée en plus de l'épreuve numérique.

1.3.2 Constitution des modules

Pour la première fois, l'évaluation CEDRE a été réalisée sur ordinateur, en ligne. Elle était constituée de 262 items au total. Parmi eux, 43 items ont été repris

à l'identique de 2007 et 31 de 2013, soit 28 % des items proposés. Le reste de l'évaluation était constitué de 188 nouveaux items.

Afin de pouvoir évaluer un nombre important d'items sans allonger le temps de passation pour l'élève, CEDRE utilise la méthodologie des modules tournants. Les items sont ainsi répartis dans des blocs qui sont ensuite distribués dans les modules tout en respectant certaines contraintes : chaque bloc doit se retrouver un même nombre de fois au total et chaque association de blocs doit figurer au moins une fois dans un module. Chaque éléve ne passe qu'un seul module. 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 qu'il ait à répondre à l'ensemble de ceux-ci.

Tableau 2 – Exemple de répartition des blocs dans les modules

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

1.4 Passation des évaluations

La passation de l'évaluation finale a eu lieu en mai 2018. Comme en 2007 et 2013, cette évaluation a été précédée d'une expérimentation l'année n-1 de façon à tester un grand nombre d'items auprés d'un échantillon réduit d'établissement. Durant l'expérimentation en 2017, une étude de comparabilité entre l'évaluation au format papier et celle au format numérique ("bridge-study") a permis de mesurer l'écart de difficulté entre ces deux modes de passation et de le reporter sur l'évaluation finale.)

Dans chaque établissement, une personne a été désignée comme étant le coordinateur, son rôle étant de veiller au strict respect de la procèdure à suivre pour que l'évaluation soit passée dans les mêmes conditions quelque soit l'établisse-

ment. Il est l'interlocuteur privilégié de la DEPP.

La séquence d'évaluation pour l'élève, sur la plateforme TAO, se déroulait en trois temps :

- Un entraînement de 10 minutes, permettant aux élèves de se familiariser avec la plateforme et de découvrir les différents types d'exercices qu'ils auront à effectuer.
- Une évaluation des connaissances et des compétences en sciences d'une durée de 1 heure.
- Un questionnaire de contexte, de 30 minutes, permettant de recueillir leurs avis sur différents aspects notamment l'évaluation CEDRE, leur travail scolaire dans les matières scientifiques et leurs usages numériques en classe et à la maison.

Pour les huit élèves sélectionnés par classe échantillonnée, une deuxième séquence d'évaluation d'une durée d'une heure était consacrée à une épreuve pratique. Cette épreuve s'est déroulée dans une salle de sciences et avec un cahier pour l'élève. Pour préparer la séquence de "travaux pratiques", le professeur de SVT ou de physique-chimie a reçu un manuel de passation contenant la grille d'évaluation ainsi qu'une description exhaustive :

- Du rôle du professeur lors de la séquence,
- Du matériel nécessaire à la préparation de la séquence ;
- De l'installation de la salle et du matériel. Le professeur en charge de l'épreuve pratique a aussi dû renseigner un questionnaire de contexte portant sur son métier et ses pratiques pédagogiques.

2 Sondage

2.1 Méthodes

2.1.1 Tirage équilibré de classes de 3e

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

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

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

Cette propriété s'écrit :

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

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

Tirage d'établissements *versus* Tirage de classes

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

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

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

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

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

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

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

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

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

2.1.2 Redressement de la non réponse : calage sur marges

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

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

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

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

Tirage équilibré après élimination de la base des échantillons précédemment tirés

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

Formulation du problème et notations

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

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

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

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

Équilibrage

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

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

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

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

Ainsi, on aura :

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

Or, en espérance on a

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

La condition d'équilibrage initiale est donc remplie.

Condition fondamentale

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

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

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

2.1.3 Calcul de précision : méthode

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

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

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

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

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

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

2.2 Echantillonnage

Champ

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

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

Stratification

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

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

Modalités de sélection

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

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

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

Echantillon 2018

L'échantillon vise 6 000 élèves répartis proportionnellement selon les trois strates.

Base de sondage

Le tableau 3 présente les exclusions dans la population ciblée.

Tableau 3 – Exclusions pour la base de sondage - CEDRE 2018 Sciences expérimentales Collège

	Établissements	**Elèves**
Etab. accueillant des élèves de 3e	8 456	831 240
On retire les COM	8 417	826 889
On retire les étab hors contrat	8 223	824 225
On retire les EREA	8 155	822 948
On retire les UPE2A	8 145	821 865
On retire les ULIS	8 131	820 076
On ne garde que les collèges	6 935	791 891
On ne garde que les 3ème générales	6 929	765 985
Base CEDRE 3e	6 929	765 985
On retire ICILS, TIMSS8, PISA, PISA SUP, SOCLE6e, CEDRE MC1 (TAO) et CEDRE MC2 (Bridge Study)	5 677	606 667
Base de tirage CEDRE SCIENCES	**5 677**	**606 667**

Le tableau 4 présente la répartition de la population ciblée selon le secteur d'enseignement.

Tableau 4 – Répartition dans la base de sondage - CEDRE 2018 Sciences expérimentales Collège

Strate	**Établissements**	**Élèves**
1. Public hors EP	4 190	481 182
2. EP	1 092	120 979
3. Privé	1 647	163 824
Total	**6 929**	**765 985**

Échantillon

Le tableau 5 présente la répartition de l'échantillon selon le secteur d'enseignement. Au total, 236 écoles ont été sélectionnées.

Tableau 5 – Répartition dans l'échantillon - CEDRE 2018 Sciences expérimentales Collège

Strate	**Établissements**	**Élèves**
1. Public hors EP	147	3 780
2. EP	41	965
3. Privé	48	1 315
Total	**236**	**6 060**

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 2018.

80.1 % des établissements de l'échantillon ont répondu à l'évaluation (tableau 6). 77.3 % des éffectifs attendus ont participé (tableau 7).

Tableau 6 – Non-réponse des établissements - CEDRE 2018 Sciences expérimentales Collège

Strate	Nb établissements attendus	Nb établissements répondants	% d'établissements répondants
1. Public hors EP	147	133	90.5 %
2. EP	41	34	82.9 %
3. Privé	48	37	77.1 %
Total	**236**	**189**	**80.1 %**

Tableau 7 – Non-réponse des élèves - CEDRE 2018 Sciences expérimentales Collège

Strate	Nb élèves attendus	Nb élèves répondants	% d'élèves répondants
1. Public hors EP	3 780	3 051	80.7 %
2. EP	965	716	74.2 %
3. Privé	1 315	918	69.8 %
Total	**6 060**	**4 685**	**77.3 %**

2.3.2 Valeurs manquantes et imputation

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

2.3.3 Non-réponse partielle et terminale

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

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

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

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

- 7.2 items pour la séquence 1

On considère que :

- 51 élèves n'ont pas vu la séquence 1, dont :
  - 0 n'ont répondu à aucun items de la séquence
  - 51 ont répondu à moins de 50 % de la séquence

Les élèves dont toutes les séquences sont codées en "s" sont classés en non réponse totale. C'est le cas pour 51 élèves.

2.4 Redressement

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

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

Tableau 8 – Comparaison entre les marges de l'échantillon et les marges dans la population - CEDRE 2018 Sciences expérimentales Collège

Modalité	Variable	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population
Retard	1	87 169.09	105 935.73	11.38	13.83
	2	678 815.91	660 049.27	88.62	86.17
Sexe	1	362 540.7	383 145.7	47.33	50.02
	2	403 444.3	382 839.3	52.67	49.98
Strate	1	481 191.78	481 191.78	62.82	62.82
	2	120 949.03	120 949.03	15.79	15.79

2.5 Précision

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

Tableau 9 – Scores moyens et erreurs standard associées - CEDRE 2018 Sciences expérimentales Collège

Année	Score moyen	Erreur standard
2007	250	1.84
2013	250.3	1.07
2018	237.8	1.28

Pour savoir par exemple si l'évolution entre 2013 et 2018 est significative , il faut calculer la valeur suivante :

$$\frac{\left|\hat{\bar{Y}}_{2018} - \hat{\bar{Y}}_{2013}\right|}{\sqrt{se^2_{\hat{\bar{Y}}_{2018}} + se^2_{\hat{\bar{Y}}_{2013}}}} \tag{3}$$

Entre 2013 et 2018, on obtient une valeur de 7.5 (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 10 et 11).

Tableau 10 – Répartitions en % dans les groupes de niveaux - CEDRE 2018 Sciences expérimentales Collège

Année	Groupe <1	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5
2007	2.2	12.8	29.1	29	16.9	10
2013	2.6	12.8	27	29.5	18.9	9.2
2018	5.7	15.9	29.2	29.3	14.6	5.3

Tableau 11 – Erreurs standards des répartitions en % dans les groupes de niveaux - CEDRE 2018 Sciences expérimentales Collège

Année	Groupe <1	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5
2007	0.4	0.9	0.8	0.8	0.8	0.7
2013	0.2	0.5	0.6	0.6	0.5	0.5
2018	0.5	0.7	0.8	0.8	0.7	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 12 – Effet du plan de sondage - CEDRE 2018 Sciences expérimentales Collège

Année	Erreur Standard	Erreur SAS	*Design Effect*
2006	1.84	0.65	2.84
2012	1.07	0.54	1.98
2017	1.28	0.72	1.78

Dans le cas d'un sondage en grappes, la précision est dégradée en comparaison d'un sondage aléatoire simple. Cela signifie qu'en 2018, un sondage aléatoire simple avec un effectif 1.78 fois moins important aurait conduit au même niveau de précision.

3 Analyse des items

3.1 Méthodologie

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

3.1.1 Approche classique

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

Réussite et score

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

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

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

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

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

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

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

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

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

Fidélité

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

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

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

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

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

Indices de discrimination

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

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

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

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

En effet, on peut montrer que

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

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

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

3.1.2 Analyse factorielle des items

L'analyse factorielle permet d'étudier la structure des données et, plus particulièrement, la structure des corrélations entre les variables observées (ou manifestes) [4]. Il s'agit d'identifier les différentes dimensions sous-jacentes aux réussites observées et surtout d'évaluer le poids de la dimension principale, dans la mesure où c'est une optique unidimensionnelle qui sera envisagée lors de la modélisation.

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

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

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

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

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

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

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

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

3.2 Codage des réponses aux items

3.2.1 Valeurs manquantes

Trois types de valeurs manquantes sont distinguées :

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

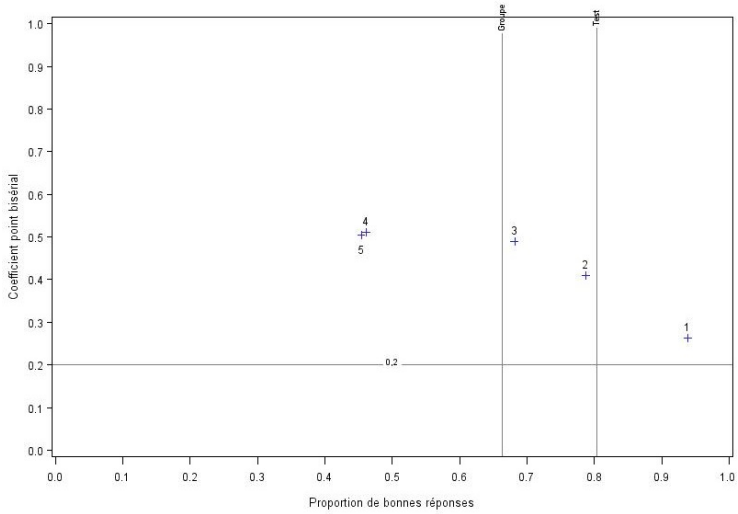
— Non-réponse terminale : l'élève s'est arrêté au cours de l'épreuve, potentiellement en raison d'un manque de temps. Des choix sont effectués pour déterminer le traitement de ces valeurs. Nous considérons que si un élève a passé moins de 50 % d'une séquence, il n'a pas vu la séquence, les valeurs manquantes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

3.2.2 Regroupement des items

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

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

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



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

3.2.3 Traitement des données et correction des questions ouvertes

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

AGATE : un outil de correction à distance des questions ouvertes

Objectifs

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

Principes fonctionnels

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

3.3 Résultats

3.3.1 Pouvoir discriminant des items

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

- 2 items de 2007
- 1 item de 2013
- 14 items de 2018
- 3 items de 2013-2018
- 2 items de 2007-2013-2018

4 Modélisation

4.1 Méthodologie

4.1.1 Modèle de réponse à l'item

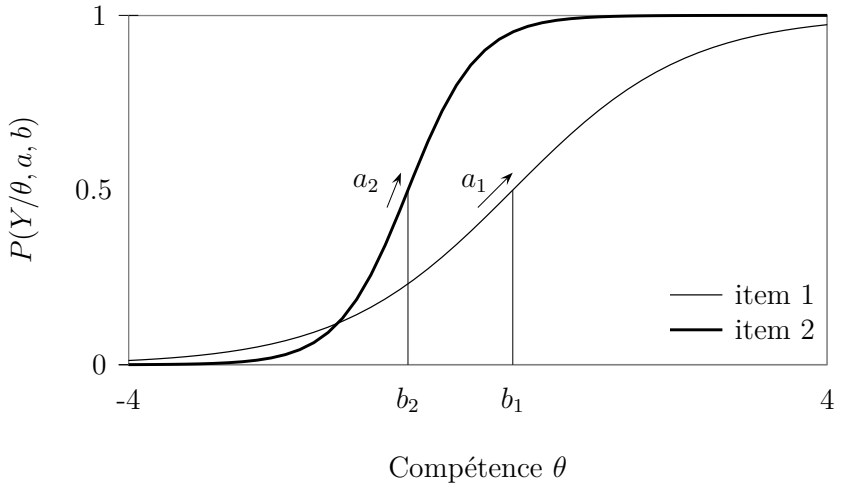
Le modèle de mesure utilisé est un modèle de réponse à l'item à deux paramètres avec une fonction de lien logistique (MRI 2PL) :

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

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

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

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



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

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

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

4.1.2 Procédures d'estimation

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

Estimation des paramètres des items

Nous reprenons les notations de l'équation (11) qui formule la probabilité P_{ij} d'un élève i de répondre correctement à un item j dans le cadre d'un modèle de réponse à l'item, avec les items sont dichotomiques.

Notons tout d'abord que les modèles présentés ne sont pas identifiables. En effet, les transformations $\theta_i^* = A\theta_i + B$, $b_j^* = Ab_j + B$ et $a_j^* = a_j/A$ avec A et B deux constantes $(A > 0)$, conduisent aux mêmes valeurs des probabilités. Dans CEDRE, nous levons l'indétermination en standardisant la distribution des θ pour les données du premier cycle (en l'occurrence, moyenne de 250 et écart-type de 50 pour l'année 2007).

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

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

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

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

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

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

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

avec g la densité de θ.

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

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

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

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

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

Estimation des niveaux de compétence

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

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

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

avec

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

et

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

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

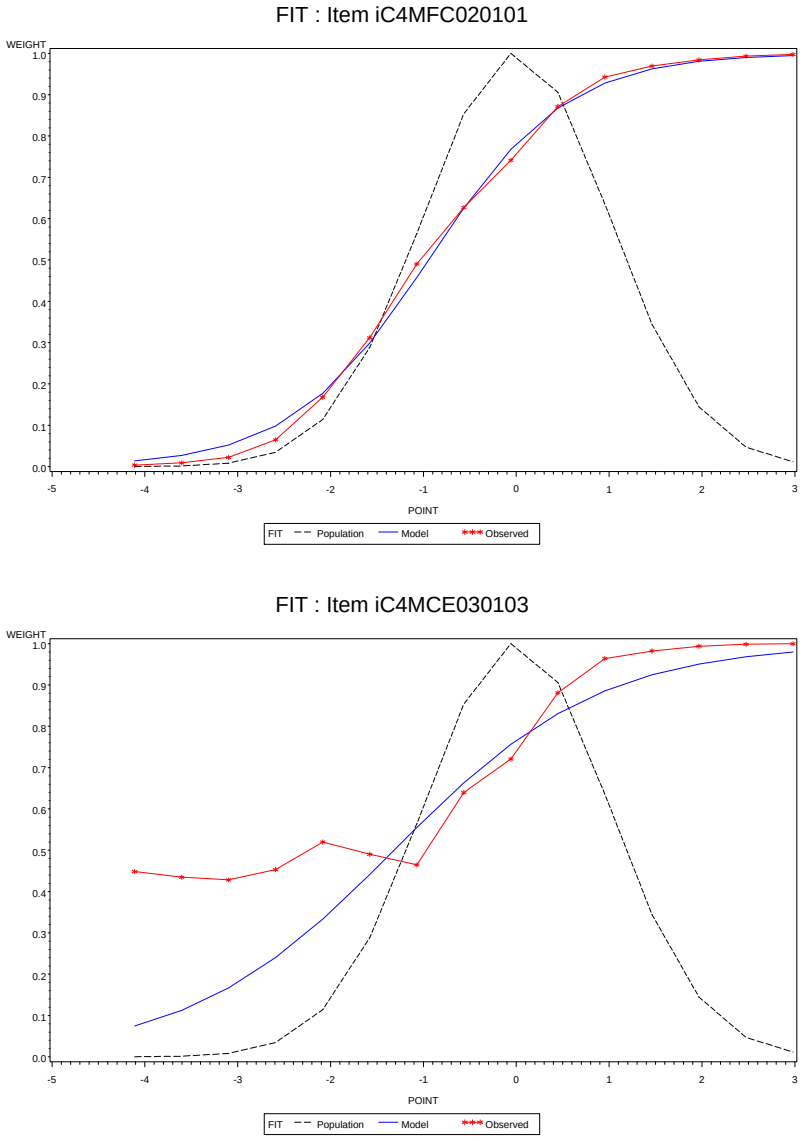
$$\frac{\partial lnL}{\partial \theta_i} + \frac{J}{2I} = 0 \tag{16}$$

4.1.3 Indice d'ajustement (FIT)

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

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

Figure 3 – Exemples d'ajustements (FIT)



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

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

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

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

4.1.4 Fonctionnement Différentiel d'Item (FDI)

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

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

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

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

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

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

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

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

4.1.5 L'information du test

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

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

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

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

4.1.6 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 2018 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 2017 était composée d'items de 2013, 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 2018, 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 égale-ment 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)

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

— 4 items d'ancrage 2007-2013
— 12 items d'ancrage 2007-2013-2018

4.2.2 Bilan de l'analyse des items

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

— 85 items de 2007
— 40 items de 2013
— 188 items de 2018
— 42 items d'ancrage 2007-2013
— 31 items d'ancrage 2013-2018
— 43 items d'ancrage 2007-2013-2018

Cela représente 429 items passés par les élèves en tout, dont 262 en 2018.

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

— 83 items de 2007
— 39 items de 2013
— 174 items de 2018
— 38 items d'ancrage 2007-2013
— 28 items d'ancrage 2013-2018
— 29 items d'ancrage 2007-2013-2018

391 items sont donc conservés dans l'analyse, dont 231 utilisés dans l'évaluation 2018.

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 2007. Le tableau 13 présente les résultats obtenus.

Tableau 13 – Niveaux de compétences (moyennes des scores et écarts-types) - CEDRE 2018 Sciences expérimentales Collège

Année	Score moyen	Écart-type
2007	250	50
2013	250.3	50.1
2018	237.8	48.9

5 Construction de l'échelle

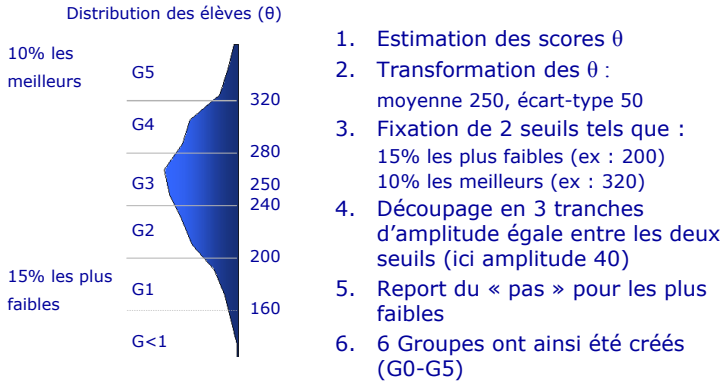
5.1 Méthode

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 Sciences expérimentales 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 2007. Puis, comme le montre la figure 4, 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 4 – 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 (Bret, Dos Santos, Ninnin, & Roussel, 2019).

Groupe < 1 (5,7 % des élèves)

Les élèves du groupe inférieur à 1 restituent des connaissances simples en relation avec leur vécu ou liées à l'éducation à la santé. Ils connaissent les gestes manipulatoires de base. À partir d'un tableau, ils prélèvent des données et observent une tendance. Ils extraient des informations simples d'un schéma (par exemple électrique).

Groupe 1 (15,9 % des élèves)

Les élèves du groupe 1 font preuve de bon sens, de comportement responsable lorsqu'il s'agit d'une situation liée à leur vécu. Ils sélectionnent des informations dans des documents divers : tableau à double entrée, graphique, photographie, carte. Ils reconnaissent l'évolution d'une grandeur dans un graphique. Ils passent d'un texte simple ou d'une photographie à un schéma simple.

Groupe 2 (29,2 % des élèves)

C'est à partir du groupe 2 que la compétence "adopter un comportement éthique et responsable" est maîtrisée. Les élèves du groupe 2 ont des connaissances plus abstraites non liées à la vie quotidienne (définition d'une planète, structure d'une cellule, brassage chromosomique lors de la fécondation). Ils identifient des questions scientifiques de la vie de tous les jours et y répondent. Ils choisissent une

hypothèse, un dispositif expérimental simple ou une conclusion parmi plusieurs propositions. Ils interprètent des résultats d'expérience. Ils utilisent un modèle simple pour répondre à un problème (forces, rayon lumineux), une simulation pour distinguer cause et conséquence. Ils extraient des informations apportées par un texte long ou par un graphique complexe à deux courbes et peuvent également comparer l'allure de deux courbes. Ils transposent un texte avec un vocabulaire spécifique en un schéma d'un mécanisme biologique complexe. Ils commencent à rédiger des réponses descriptives.

Groupe 3 (29,3 % des élèves)

C'est à partir du groupe 3 que les compétences "maîtriser les connaissances attendues", "pratiquer des langages" ou "pratiquer des démarches scientifiques" sont maîtrisées quel que soit le type de connaissances (notionnelles, procédurales ou épistémiques) en jeu. Ils sont sensibilisés aux questions environnementales. Les élèves du groupe 3 maîtrisent des connaissances scientifiques générales du cycle 4 (caryotype, fécondation et combinaison allélique, ressources d'énergies renouvelables et non renouvelables, différence entre transformation chimique et physique). Ils utilisent leurs connaissances pour exploiter un schéma, un tableau ou une clé de détermination. Ils mettent en relation des informations issues de différentes sources pour répondre à une question. Ils connaissent certains protocoles expérimentaux et peuvent les rédiger. Ils prévoient un résultat expérimental dans des cas simples. Ils utilisent une animation pour déterminer les paramètres influençant un phénomène physique. Ils associent les éléments du réel avec les éléments d'un modèle. Ils utilisent la notion de proportionnalité et la reconnaissent.

Groupe 4 (14,6 % des élèves)

C'est à partir du groupe 4 que la compétence "se situer dans l'espace et dans le temps" est maîtrisée. Les élèves ont des connaissances pointues dans des domaines variés du cycle 4 (nombre de chromosomes dans différentes cellules, conservation de la masse lors d'une transformation chimique ou encore description microscopique de la matière). Ils mobilisent une connaissance précise pour effectuer un calcul et savent associer grandeur et unité de mesure. Quel que soit le domaine de connaissances, ils passent facilement d'un langage à un autre (par exemple, d'un texte long à un schéma fonctionnel). Ils mettent en relation des documents de nature et de représentations variées avec des données complexes (graphiques dont les paramètres ne varient pas dans le même sens, par exemple). Ils choisissent ou proposent des dispositifs expérimentaux complexes, pour répondre à un problème scientifique et peuvent également formuler la question scientifique associée à un dispositif expérimental. Ils confrontent les résultats expérimentaux pour conclure et sont critiques face à une expérience. Ils maîtrisent les étapes de la démarche scientifique. Ces élèves peuvent rédiger

des réponses longues pour expliquer et justifier leur propos.

Groupe 5 (5,3 % des élèves)

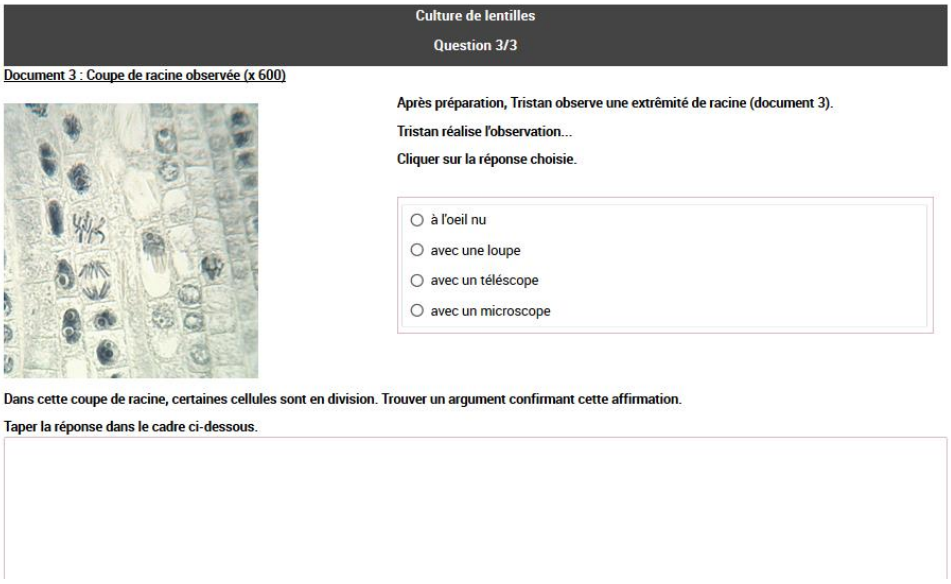
Les élèves du groupe 5 manient avec rigueur le vocabulaire et le formalisme scientifiques. Ils maîtrisent l'utilisation des nombres ainsi que le calcul littéral pour répondre à une question scientifique. Ils savent écrire un résultat avec la bonne unité. Ils sont capables de prévoir l'évolution d'une grandeur. Leur raisonnement est rigoureux et exposé de façon structurée. Ils font preuve d'esprit critique dans l'analyse de situations complexes, de modèles ou de documents dans des situations différentes de celles vues en classe usuellement. Ils utilisent un support numérique pour construire un tableau de données. Ils éprouvent encore des difficultés à travailler sur l'erreur, à appréhender l'échelle d'espace pour les plus petits éléments, et à argumenter dans une situation non étudiée en classe.

5.3 Exemples d'items

5.3.1 Item caractéristique des groupes < 1 et 1

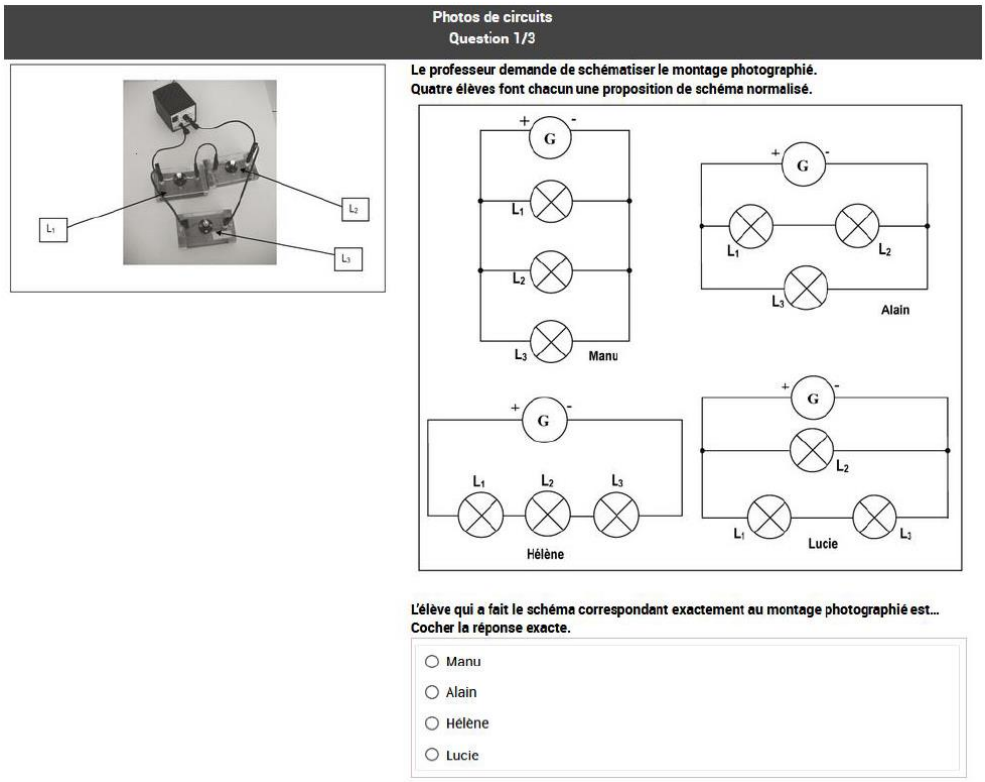
Un exemple en SVT :

Figure 5 – Exemple groupe < à 1



Un exemple en physique-chimie :

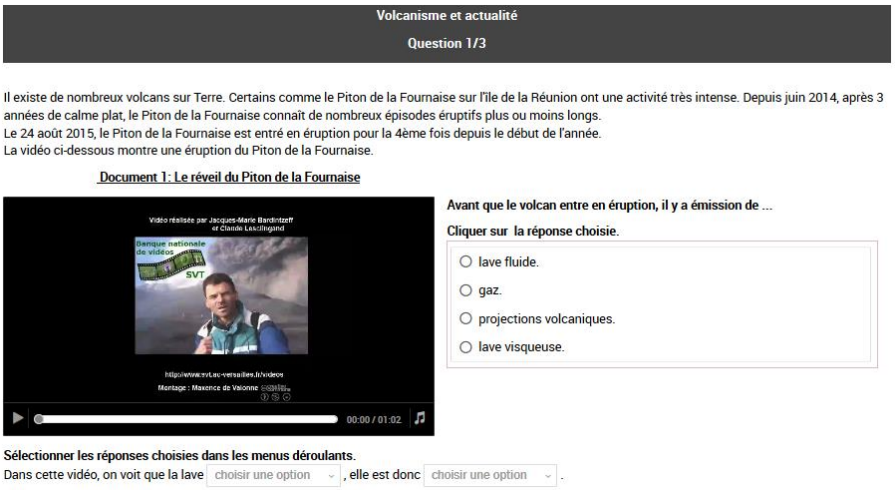
Figure 6 – Exemple groupe < à 1



5.3.2 Item caractéristique du groupe 2

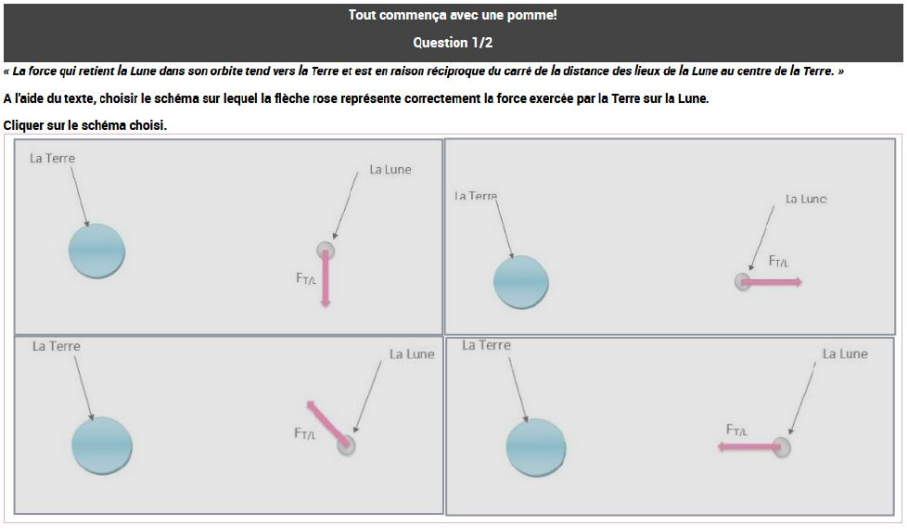
Un exemple en SVT :

Figure 7 – Exemple groupe 2



Un exemple en physique-chimie :

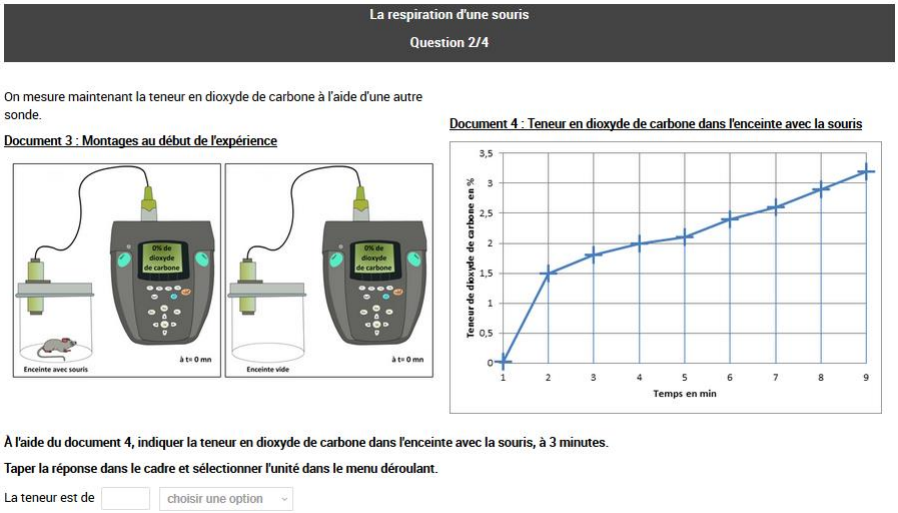
Figure 8 – Exemple groupe 2



5.3.3 Item caractéristique du groupe 3

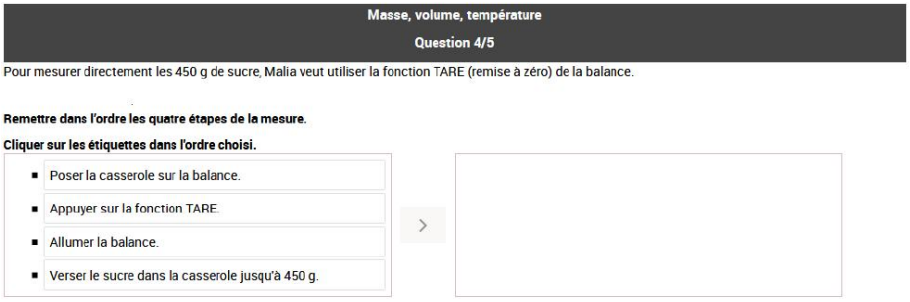
Un exemple en SVT :

Figure 9 – Exemple groupe 3



Un exemple en physique-chimie :

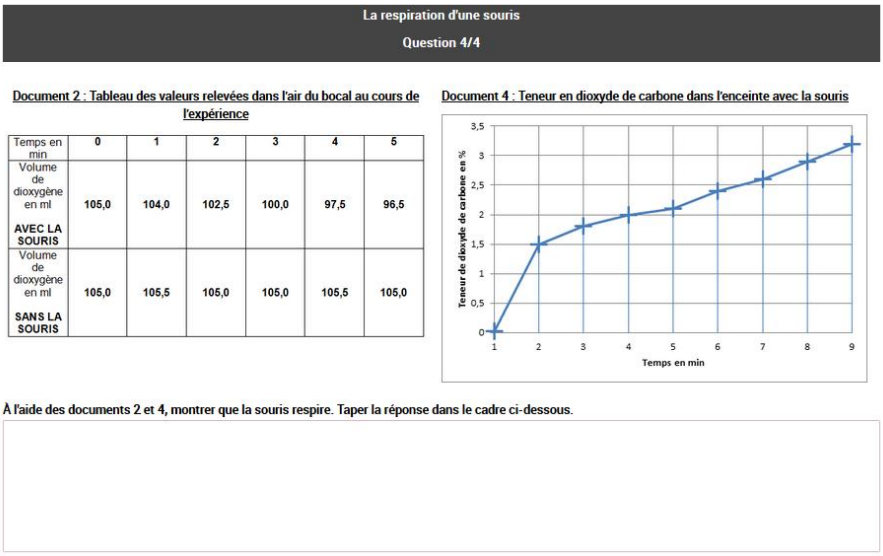
Figure 10 – Exemple groupe 3



5.3.4 Item caractéristique du groupe 4

Un exemple en SVT :

Figure 11 – Exemple groupe 4 - 1



Un exemple en physique-chimie :

Figure 12 – Exemple groupe 4 - 2

Une nouvelle Terre

Question 1/4

En Juillet 2015, le télescope Kepler a détecté une planète située dans un autre système solaire à $1{,}33 \times 10^{15}$ km de notre Terre. Baptisée Kepler-452b, elle a des caractéristiques très voisines de celles de la Terre car elle se situe aussi à environ 150 millions de kilomètres de son étoile. Elle est en revanche un peu plus grosse car elle a un rayon de 10 000 kilomètres. Sa masse, estimée à 30×10^{23} kg, est aussi plus élevée.

A l'aide de l'animation, donner la valeur de la force de gravitation exercée par l'étoile sur la planète Kepler-452b.

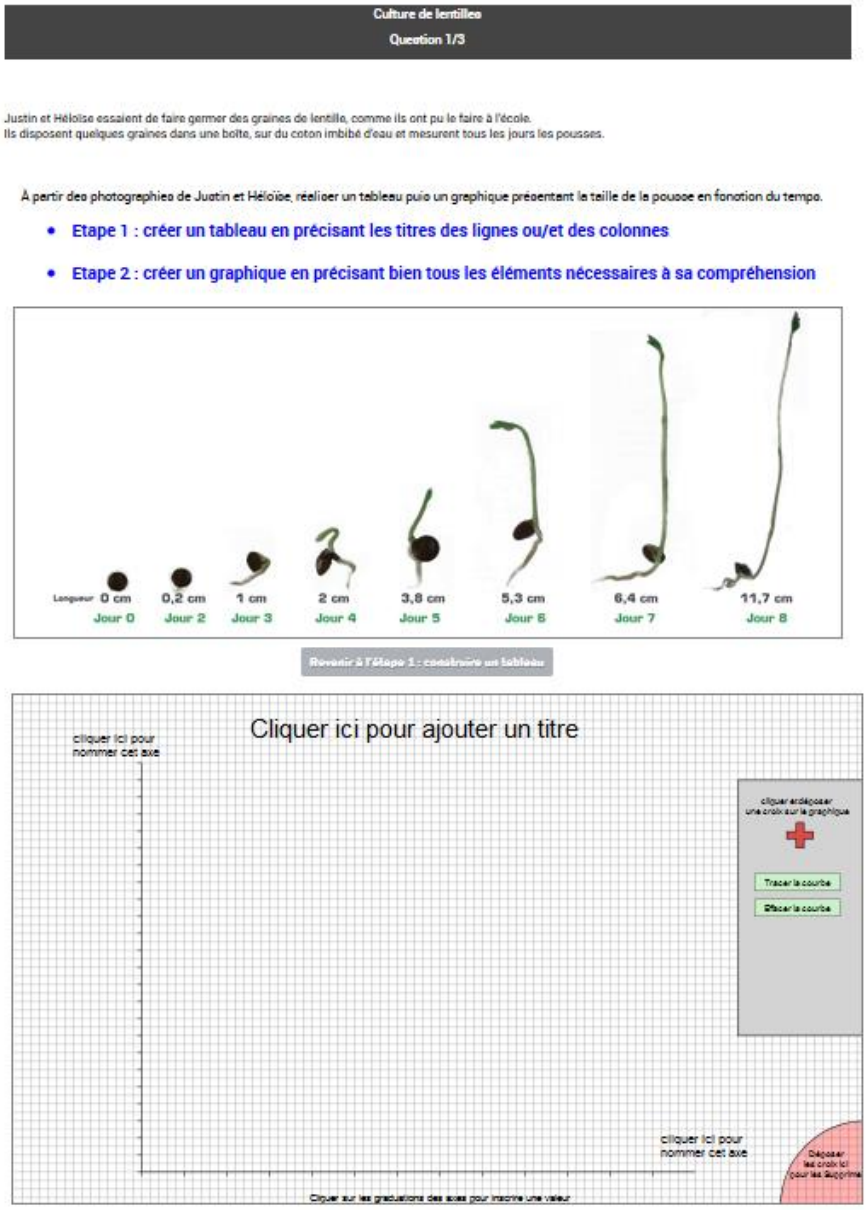
Taper la réponse dans le cadre ci-dessous.

$F =$ ____ $\times 10^{27} N$

5.3.5 Item caractéristique du groupe 5

Un exemple en SVT :

Figure 13 – Exemple groupe 5 - 1



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 (Bret et al., 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 2018, 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 2007, 2013 et 2018, 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 14).

Tableau 14 – Score moyen selon l'indice de position sociale moyen de l'établissement (CEDRE 2007 - 2013 - 2018)

Indice moyen école	Année	Répartition (%)	Score moyen	Écart type
1er quart	2007	24.6	236.3	46.8
1er quart	2013	25.0	233.2	48.4
1er quart	2018	24.9	**220.4**	48.9
2e quart	2007	25.1	250.6	46.8
2e quart	2013	24.8	248.4	49.8
2e quart	2018	24.9	**235.6**	46.9
3e quart	2007	25.1	248.0	52.4
3e quart	2013	25.2	252.4	**45.2**
3e quart	2018	25.0	**244.7**	45.2
4e quart	2007	25.3	264.7	49.6
4e quart	2013	25.0	267.0	51.2
4e quart	2018	25.2	**250.2**	49.1

Note de lecture : en 2018, le score moyen des élèves appartenant au quart des collèges les plus défavorisées (1er quart) diminue de 12.8 points par rapport à 2013. 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

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

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

Le tableau 15 présente les grands résultats de cet instrument.

Tableau 15 – Résultats de l'instrument de mesure de la motivation au test (CEDRE 2018)

	Moyenne	Erreur standard
Comment avez-vous trouvé les exercices de cette évaluation ?	5,4	2,26
Comment vous êtes-vous appliqué(e) pour faire cette évaluation ?	6,2	2,65
Si les résultats de cette évaluation comptaient pour votre bulletin scolaire, comment vous seriez-vous appliqué(e) ?	8,5	3,93

Figure 14 – Instrument de mesure de la motivation au test

Question 16

Sur une échelle de difficulté allant de 1 à 10, comment avez-vous trouvé les exercices de cette évaluation ?

Très faciles ☐1 ☐2 ☐3 ☐4 ☐5 ☐6 ☐7 ☐8 ☐9 ☐10 Très difficiles

C3HQQ1290201
247

Question 17

Comment vous êtes-vous appliqué(e) pour faire cette évaluation ?

(Indiquez votre niveau d'application sur une échelle allant de 1 à 10)

Je ne me suis pas du tout appliqué(e) ☐1 ☐2 ☐3 ☐4 ☐5 ☐6 ☐7 ☐8 ☐9 ☐10 Je me suis énormément appliqué(e)

C3HQQ1290401
248

Question 18

Si les résultats de cette évaluation comptaient pour votre bulletin scolaire, comment vous seriez-vous appliqué(e) ?

(Indiquez votre niveau d'application sur une échelle allant de 1 à 10)

Je ne me serais pas du tout appliqué(e) ☐1 ☐2 ☐3 ☐4 ☐5 ☐6 ☐7 ☐8 ☐9 ☐10 Je me serais énormément appliqué(e)

C3HQQ1290301
249

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.

Bret, A., Dos Santos, R., Ninnin, L.-M., & Roussel, L. (2019). CEDRE 2007-2013-2018 - sciences en fin de collège : des résultats en baisse. *Note d'information*, *16*.

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

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

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

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

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

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

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

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

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

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

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

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

Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54 nº3*, 427-450.

Liste des tableaux

Table des figures