CEDRE

Cycle des Evaluations Disciplinaires Réalisées sur Echantillons

Rapport technique

Compétences langagières et littératie 2015

Collège

Auteurs :
Etienne DALIBARD
Sylvie FUMEL
Saskia KESKPAIK
Marion LE CAM
Louis Marie NINNIN
Thierry ROCHER
Ronan VOURC'H

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

Février 2018

Table des matières

Introduction

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

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

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

1 Cadre d'évaluation

1.1 Objectifs

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

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

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

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

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

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

Discipline évaluée	Début du cycle	Reprises	
Maîtrise de la langue (CM2)	2003	2009	2015
Compétences langagières et littératie (3e)			2015
Langues vivantes étrangères (CM2 et 3e)	2004	2010	2016
Histoire, géographie et éducation civique (CM2 et 3e)	2006	2012	2017
Sciences expérimentales (CM2 et 3e)	2007	2013	2018
Mathématiques (CM2 et 3e)	2008	2014	2019

1.2 Connaissances et compétences visées

1.2.1 Principes généraux

CEDRE Compétences langagières et littératie 2015 s'appuie sur les programmes du collège, au-delà du cloisonnement disciplinaire scolaire, sans être une évaluation directe et exhaustive de ce qui est défini dans les programmes.

Sa perspective n'est pas exclusivement scolaire ; elle s'intéresse, en fin de collège, à évaluer les compétences en littératie dans une perspective interdisciplinaire. C'est une nouvelle orientation, aussi n'est-il ni possible ni légitime de comparer les résultats présentés ici avec ceux des années antérieures, qui concernaient des compétences dites « générales ». Le nouveau cadre proposé est plus clairement orienté vers les compétences langagières et la littératie, telles que définies dans cette section.

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

Les connaissances et compétences telles qu'elles sont définies dans le socle commun de connaissances et de compétences ainsi que les programmes officiels constituent le cadre de cette évaluation.

Un balayage exhaustif des programmes étant impossible, cette évaluation est conçue à partir de leurs finalités majeures en littératie.

1.2.2 Littératie

Le monde contemporain est marqué par une expansion considérable de l'écrit, en partie dû au développement de la communication numérique dans tous les do-

maines et à ses effets dans les relations sociales, professionnelles et personnelles. On assiste ainsi à une mutation à l'échelle planétaire. Corollairement, cette mutation exige de tous des pratiques avancées en lecture et écriture. L'école se doit de former les élèves à ces pratiques, qui font appel à des compétences générales et langagières de haut niveau, lesquelles transforment en retour les rapports avec l'écrit et le savoir qu'entretiennent ceux qui en disposent (Goody, 1979 ; Olson, 1998 [1994]). Dans des sociétés confrontées à la montée en puissance des besoins de lire et d'écrire, la notion de littératie, qu'elle soit prise dans une acception large ou restreinte (Jaffré, 2003), permet de prendre en compte la généralisation des pratiques de lecture et d'écriture par des groupes sociaux variés dans des contextes d'usage diversifiés. Elle invite à réexaminer la relation entre lecture et écriture à la lumière des contextes sociaux dans lesquels elle s'inscrit, ainsi que l'articulation des différents usages de l'écrit (en intégrant celui qui a recours aux technologies informatiques), mais aussi la relation entre oral et écrit (Barré-De Miniac, 2003 ; Barré, Brissaud, Rispail, 2004).

Elle permet en outre une approche décloisonnée, voire transdisciplinaire, des manières de lire et d'écrire, dans l'ordinaire du langage, sans coupure entre langue et littérature, entre textes fonctionnel et fictionnel, textes continu et non continu, sans réduction de la langue à une de ses parties (lecture, compréhension, écriture) (Chiss, 2003).

De la compréhension orale à la littératie multimodale

La notion de littératie, malgré tout son intérêt tant pour l'évaluation de l'écrit que pour celle les compétences liées aux technologies de l'information et de la communication, laisse de côté un pan essentiel des compétences langagières touchant le langage oral, en réception comme en production.

Si l'évaluation des langues vivantes, fondée sur le Cadre européen commun de référence pour les langues, a pris en compte cette compétence de communication qui s'inscrit dans un modèle dit « actionnel » des activités langagières, cadre dont s'inspirent également les évaluations en FLE, il n'en est pas de même pour l'évaluation des compétences dans le domaine du français langue maternelle. Pourtant, les programmes réaffirment régulièrement l'importance de l'enseignement de l'oral qui constitue un des quatre sous-domaines de la discipline « français ». Le socle commun a également introduit cette dimension orale des compétences langagières pour la maîtrise de la langue française, mais l'aborde uniquement du côté du « Dire », en présupposant acquise la compréhension à l'oral. La dimension de la réception est pourtant essentielle à l'oral et conditionne la capacité à comprendre et prendre part à des échanges. De même que les évaluations cherchent à mesurer les capacités des élèves à comprendre ce qu'ils lisent, il faut qu'elles puissent tester leurs capacités à comprendre des informations orales, qui peuvent associer le son et l'image. L'évaluation se concentrera sur la réception

et la compréhension à l'oral. La didactique du français langue maternelle offre peu de modèles théoriques pour une telle évaluation. La recherche dans ce domaine s'est davantage intéressée aux activités langagières à l'oral et donc à la production orale, qu'à la compréhension. Ainsi, J. Dolz et B. Schneuwly (1998) ont cherché à dégager des « genres formels » à l'oral, qui pourraient s'enseigner comme les genres qui servent de modèles à l'enseignement des productions écrites à l'école. Si ce modèle peut nous intéresser pour la production orale, il est moins adapté à la compréhension, sinon pour chercher à définir des genres oraux dont on peut attendre que les élèves aient une fréquentation et qui pourraient servir de support aux exercices d'évaluation (par exemple, le genre de l'interview). Un autre ensemble de recherches s'est intéressé à « l'oral pour apprendre » (E. Nonnon, 2000), en particulier à l'école primaire, en étudiant les interactions verbales et l'étayage de l'adulte, essentiellement dans leur dimension cognitive. Ces recherches sont en lien avec une pédagogie de l'oral intégré, qui s'appuie sur les conduites verbales dans la classe et se prête mal à des tests à partir de supports isolés d'un contexte d'enseignement en classe.

Faute de modèle théorique spécifique à l'oral, le choix a été fait, s'agissant de compréhension orale, d'évaluer des compétences semblables à celles de la compréhension de l'écrit. Ce choix permettra d'expérimenter des exercices d'évaluation de la compréhension orale à partir de plusieurs types de documents et également de comparer les performances des élèves selon qu'il s'agit de comprendre à partir d'un support écrit, d'un support faisant interagir la parole orale et l'image ou d'un support uniquement audio.

Plus encore, il semble que la répartition actuelle des compétences adoptée pour l'enseignement et l'évaluation des langues vivantes (compréhension de l'écrit et compréhension de l'oral) soit remise en cause par l'évolution rapide des supports d'information à l'heure du numérique. Sur Internet les informations mêlant le texte, l'image et le son dominent et c'est finalement le concept de littératie multimodale qui paraît le mieux à même de rendre compte des compétences attendues des élèves à l'heure du numérique. Mais si le supports et les modes de transmission de l'information varient ou sont mixtes, la capacité à comprendre et les différentes compétences qu'elle requiert peuvent s'analyser de la même façon que pour la compréhension en lecture. C'est pourquoi nous proposns la même grille pour les différents aspects de la compétence, qu'il s'agisse de supports écrits ou de supports mêlant image et son.

Quatre grands domaines de compétences sont évalués en compréhension de l'écrit et de l'oral.

- Prélever une information : l'information explicite à prélever peut l'être de façon immédiate ou non immédiatement repérable ;
- Traiter et intégrer des informations : on distingue trois sous-compétences : ex-

pliciter une information implicite, mettre en relation ou hiérarchiser plusieurs informations pour dégager le sens global, identifier la visée d'un texte ;
– Réfléchir et évaluer : on distingue trois sous-compétences : formuler des hypothèses pour interpréter, exprimer une opinion sur le texte en convoquant ses connaissances et son expérience, évaluer les moyens utilisés par le texte pour répondre aux objectifs visés ;
– Expliquer et raisonner : des questions de métacognition demandent aux élèves de porter un regard réflexif sur les tâches accomplies et sur leurs stratégies de lecture.

1.2.3 Production écrite

Les textes produits par les élèves sont évalués au brevet des collèges et au baccalauréat et dans le cadre des évaluations nationales pratiquées depuis la loi d'orientation de 1989. Mais la question des critères d'évaluation a toujours soulevé des problèmes et la disparité des notes obtenues en rédaction ou en dissertation est dénoncée depuis longtemps par les études docimologiques (Merle, 2007). Par ailleurs, on a relevé, également depuis longtemps, que certains aspects étaient survalorisés, comme l'orthographe ou la syntaxe, voire la propreté de la copie, dont les dysfonctionnements sont facilement repérables et vont jusqu'à occulter la qualité du contenu ou l'organisation du texte (Blanche-Benveniste, 1979) ; ou encore, comme l'originalité ou l'aisance de la rédaction, critères généraux, tous plus subjectifs et non enseignables les uns que les autres.

À partir des années 1980, des travaux issus de trois domaines de recherches vont modifier l'enseignement dans les classes, au moins dans une certaine mesure, difficile à préciser :
– les travaux sur le genre, l'énonciation et le fonctionnement des textes en linguistique (Benveniste, 1970 ; Bakhtine, 1984 ; Combettes, 1988), qui ont ouvert la voie à une appréhension plus fine et plus rigoureuse des productions écrites ;
– les travaux sur les processus rédactionnels (Hayes et Flower, 1980), qui ont permis de mettre en évidence le cout cognitif de la production d'écrit et de sa révision alors que les ressources du scripteur sont limitées, notamment chez l'élève jeune (Fayol, 1997 ; 2013) ;
– les travaux sur l'évaluation, qui ont introduit la fonction formative de l'évaluation (Allal, Cardinet, Perrenoud, 1979).

À partir de ces travaux, des notions sont élaborées (on retient de la « grammaire textuelle » les notions de cohérence textuelle, de continuité thématique, de substituts, etc.) et des outils mis au point pour l'école, qui ont pour but d'aider à

l'écriture, la réécriture, l'évaluation. La grille EVA (1991) a par exemple permis de dégrouper des approches globales et de rendre compte d'un texte d'élève à des niveaux différents (efficacité pragmatique, cohérence, morphosyntaxe, orthographe). La réécriture-révision se substitue au « brouillon à relire avant de recopier » : il s'agit d'un véritable travail qui vise à ajuster au mieux le texte aux enjeux de la situation de communication et au destinataire.

Ces avancées à l'école sont irriguées par d'autres courants de recherche centrés sur la production : la critique génétique, qui essaie de suivre pas à pas la fabrication du texte et les diverses opérations d'écriture en jeu (ITEM), les ateliers d'écriture (Bing, 1976 ; André, 1990 ; Lafont-Terranova, 2009), qui ont remis en cause l'évaluation des textes par le seul écart par rapport à la norme et ont montré la diversité des processus de production. Toutes les recherches convergent pour faire de la réécriture un processus essentiel de l'écriture experte.

Cependant, de nouvelles difficultés sont apparues, tant dans l'enseignement que dans l'évaluation :

- celle de la fragmentation en une succession de critères à respecter, issus des grilles (2 points pour le respect de la consigne, 2 points pour la conformité au genre, etc.), ce qui fait parfois perdre de vue le texte dans sa globalité ou conduit à attribuer avec réticence une bonne note à un texte jugé globalement peu satisfaisant ;
- la difficulté à faire réécrire plusieurs fois des élèves, qui ont du mal à percevoir l'intérêt de la production d'écrit et encore moins celui de la révision ;
- la déception face à des textes réécrits qui ne s'en trouvent pas nécessairement meilleurs.

On se retrouve ainsi confronté à une difficulté plus fondamentale, celle d'accepter des textes imparfaits au cours de la scolarité et qui plus est, d'évaluer des progrès dans des textes imparfaits, difficulté qui rappelle le problème de l'évaluation de l'orthographe.

Dans les programmes de 2008 du collège, la section consacrée à la production écrite est peu développée relativement à d'autres (par exemple la lecture), même si l'institution scolaire reconnait une forme de complexité à la production d'écrit et l'existence d'un travail attaché à l'écriture (section « expression écrite », page 3) : « Certains travaux d'écriture sont le fruit d'une progression, d'un projet collectif ou individuel et supposent un travail patient, continu et réfléchi, d'améliorations et de corrections, selon les critères suivants : cohérence, visée, respect des consignes, orthographe, syntaxe, lexique. Cette activité est pratiquée régulièrement tout au long de la scolarité au collège. ».

L'institution ne prodigue guère de conseils, comme si « le travail patient, continu et réfléchi » allait de soi. Tout se passe comme s'il suffisait de relire son texte pour repérer ses erreurs et pour l'améliorer. Or il est clair qu'il n'y a pas de

rédaction experte sans révision et pas de révision efficace sans apprentissage de la révision. On pourrait aller jusqu'à dire que cela n'a pas de sens d'évaluer un texte non révisé. Les travaux des premières recherches-actions ont en effet montré qu'il était possible non seulement de faire écrire les élèves, mais aussi de les amener à réviser leur texte, notamment dans un projet d'écriture avec destinataire identifié (Jolibert - Groupe d'Ecouen, 1988 ; groupe EVA, 1991). Plus récemment, d'autres recherches ont montré que, guidés dans leur révision, les élèves améliorent leurs écrits (Bucheton et Chabanne, 2002 ; Crinon, Legros & Marin, 2003). Une autre direction de recherche sur la révision concerne la production de texte en temps réel : le suivi simultané des mouvements oculaires et du curseur grâce au logiciel Eye and pen montre, tout en précisant le lien entre lecture et écriture, que les élèves révisent, plutôt en surface, mais pas seulement (Plane, Olive, Alamargot, 2010).

Évaluer les compétences d'écriture des élèves est une nécessité aujourd'hui, car la compétence à produire un texte fait partie du socle commun de la littératie. L'institution dispose de peu d'informations sur les compétences des élèves, hormis une étude conduite par la DEP (MEN-DEP, 1996) comparant les compétences des élèves des années 1920 et des collégiens de 1995 de la 6e à la 4e : ces derniers ont répondu à des consignes de certificats d'études. On a évalué le type de texte produit, sa longueur, sa cohérence, la pertinence des éléments choisis et la maitrise de la langue, et on a conclu que les élèves de 1995 étaient plutôt meilleurs en rédaction que ceux des années 1920 (à l'inverse des résultats en langue). Cette étude ne peut cependant servir de base à une nouvelle évaluation à cause des sujets utilisés, qui ne font pas toujours sens dans la société actuelle.

Une évaluation dans le cadre de CEDRE contribuera ainsi à nourrir la réflexion sur les capacités réelles des élèves et donnera des repères pour une comparaison dans le temps, d'autant plus que la production écrite n'est généralement pas évaluée dans les évaluations internationales. Elle devrait permettre de préciser ce qu'est un texte réussi à 15 ans, et à poser des jalons pour dire ce qu'on est en droit d'attendre à un âge donné.

Cependant, l'évaluation de la compétence d'écriture soulève de nombreuses difficultés de mise en œvre : elle doit s'inscrire dans le temps scolaire et éviter aux correcteurs une tâche démesurée. Le nombre de compétences testées restera donc limité. Il a semble néanmoins nécessaire de sauvegarder la possibilité d'une révision, même restreinte : on a proposé un espace brouillon.

Les items de production écrite portent sur :

– Tenir compte de la situation de communication
– Assurer l'organisation et la cohérence du texte

– Tenir compte de l'intention poursuivie (informer, raconter, décrire, persuader...)
– Élaborer des contenus adaptés à la situation de communication électionner des contenus adaptés (faits ou événements, informations, arguments...)
– Assurer l'organisation et la cohérence du texte : respecter une structure textuelle ou une organisation appropriée au texte à produire, assurer la continuité textuelle par un jeu de reprises approprié (hors orthographe)
– Maîtriser les outils de la langue : respecter la syntaxe, respecter la ponctuation forte, respecter l'orthographe

1.2.4 Orthographe en production écrite

L'orthographe est toujours évaluée dans des exercices spécifiques. Or la finalité de l'apprentissage de l'orthographe est de savoir orthographier ses propres écrits, comme il est formulé dans le socle commun de connaissances et de compétences en fin de CM2 (compétence 1, 2006) : « rédiger un texte d'une quinzaine de lignes (récit, description, dialogue, texte poétique, compte rendu) en utilisant ses connaissances en vocabulaire et en grammaire ; orthographier correctement un texte simple de dix lignes - lors de sa rédaction ou de sa dictée - en se référant aux règles connues d'orthographe et de grammaire ainsi qu'à la connaissance du vocabulaire. ».

On dispose de peu de données concernant la mise en œuvre de l'orthographe en production de texte. Toutes mettent en évidence que les élèves corrigent des erreurs à différents moments de la rédaction, à savoir en cours d'écriture, en travaillant leur premier jet ou en recopiant (Fabre, 1990 ; Rillard, Sandon, 1994 ; Cogis, 2000 ; Cogis, à paraitre ; Doquet, 2011 ; Geoffre, 2013). Même si une recherche déjà ancienne a montré que des procédures de révision orthographique peuvent s'enseigner avec succès (Blain, 1996), l'enseignement de la révision orthographique semble peu répandu dans les classes.

Les résultats de recherches accumulés depuis une vingtaine d'années sur l'acquisition de l'orthographe ont permis de préciser les points nodaux ou zones de fragilité de l'orthographe du français : accord nominal et verbal au pluriel, accord dans des structures syntaxiques complexes, nom collectif en position de sujet, participe passé détaché, etc. Dans le cadre d'une évaluation de la littératie, il est important de disposer d'une évaluation de la compétence orthographique en production d'écrit, quand l'élève est à la source de l'énonciation et maitre de ses choix lexicaux et syntaxiques, comme on l'est quand on a recours à l'écrit. Cette évaluation permet de préciser ce qu'on est en droit d'attendre dans un écrit d'élèves de 15 ans, dans les conditions contraintes de la rédaction scolaire. Elle poserait les jalons d'une comparaison dans le temps.

1.2.5 Métacognition

C'est la première fois que la Depp interroge les élèves sur les processus de lecture dans une évaluation nationale. On en sait assez peu sur les stratégies que les élèves français emploient pour comprendre des textes et sur la conscience qu'ils ont de ces stratégies. Pourtant, les recherches sur les différences entre les bons lecteurs et ceux qui sont en difficulté indiquent toutes que les bons lecteurs sont des lecteurs actifs, conscients des stratégies qu'ils emploient pour accéder au sens et pour contrôler et réguler leur compréhension.

Les stratégies de compréhension renvoient à une autre composante essentielle de la compréhension qui consiste à exercer pendant la lecture une veille permettant d'évaluer la pertinence des interprétations effectuées, de repérer les éventuelles ruptures de cohérence et de mettre en œuvre des procédures ou stratégies pour résoudre les difficultés détectées. Le lecteur doit donc contrôler ou guider sa compréhension. La mise en œuvre de ce contrôle peut varier en fonction des motivations, des objectifs et de l'expertise des lecteurs, mais aussi en fonction du contenu des textes. Les stratégies sont définies comme un ensemble de procédures que le lecteur peut mobiliser de manière délibérée et sous le contrôle de l'attention. En d'autres termes, ces stratégies touchent aux aspects métacognitifs de l'activité de compréhension.

On distingue traditionnellement 3 types de stratégies, selon le moment de la lecture auquel elles s'appliquent. Les stratégies de pré-lecture préparent la lecture (« previewing the text », poser des questions préalablement à la lecture), les stratégies liées à la construction des modèles de situation aident le lecteur dans l'élaboration de la cohérence du texte (interroger le texte en le paraphrasant, en l'auto-expliquant, en (se) posant des questions, en organisant l'information sous forme graphique) ; enfin, les stratégies postérieures à la lecture aident les lecteurs à maîtriser les habiletés de compréhension appliquées (critiquer, évaluer les sources, synthétiser l'information, organiser comparer...).

1.3 Construction du test

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

1.3.1 Elaboration des items

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

Trois formats de questions sont utilisés : question à choix multiples (QCM), série de vrai/faux et question appelant une réponse rédigée.

Les questions dites ouvertes appellent des réponses sous forme de production écrite. Elles supposent la mise en place d'un dispositif de correction experte à distance pour l'épreuve finale, nécessitant la formation technique des correcteurs et l'élaboration d'un cahier des charges strict de corrections pour éviter toute subjectivité ou la validation de réponses trop imprécises ou succinctes. Une réponse rédigée à une question ouverte peut faire l'objet de plusieurs items distinguant les différentes compétences nécessaires pour répondre.

Exemples d'items

ATTENDEZ-MOI SOUS L'ORME

DORANTE, *d'un ton de colère.*
Quoi ! J'ai eu la patience de garder huit ans un coquin comme toi !

PASQUIN
Tout autant, monsieur.

DORANTE
Un maraud !

PASQUIN
Oui, monsieur.

DORANTE
Huit ans, un valet à pendre !

PASQUIN
Ah !

DORANTE
À noyer, à écraser !

PASQUIN
Il y a du malheur à mon affaire. Vous avez été jusqu'à présent très content de mon service, et vous cessez de l'être dans le moment que je vous demande mes gages.

DORANTE, *se radoucissant*
Pasquin, ce n'est pas d'aujourd'hui que je suis dupe de ma bonté. Va, mon cher, je veux bien encore ne point te chasser de chez moi.

PASQUIN
Vraiment, monsieur, ce n'est pas vous qui me chassez ; c'est moi qui vous demande mon congé, et les six cents livres.

DORANTE
Non, mon cœur, tu ne me quitteras point. Tu ne sais ce qu'il te faut. La vie champêtre ne convient point à un intrigant, à un fourbe.

PASQUIN
Je sais bien que j'ai tous les talents pour faire fortune à la ville : mais je borne mon ambition à Lisette, à qui j'apporte en mariage les six cents livres, dont je vais vous donner quittance.
(Il tire de sa poche un papier.)

DORANTE, *lui arrêtant la main.*
Peste soit du faquin[1] ! Tu n'as que tes affaires en tête ; reparlons un peu des miennes. J'épouse demain la petite fermière Agathe. J'ai si bien fait, par mon manège, que le père est à présent aussi amoureux de moi que sa fille. Elle a dix mille écus, Pasquin.

Attendez-moi sous l'orme, Jean-François Regnard, Scène I,1684
© Gallica.fr

Exemple 1 : questions à choix multiples

Les répliques sont très courtes au début. L'auteur veut montrer que...

1 ☐ Dorante est gêné
2 ☐ chacun hésite
3 ☐ Dorante est soumis
4 ☐ personne ne veut céder

Exemple 2 : série de vrai-faux

	Vrai	Faux
Dorante veut chasser Pasquin de chez lui.	☐ 1	☐ 2
Dorante doit six cents livres à Pasquin.	☐ 1	☐ 2
Dorante va offrir dix mille écus à Pasquin pour son mariage.	☐ 1	☐ 2

Exemple 3 : question ouverte

À votre avis, à quel « changement » (ligne 30) le lecteur peut-il s'attendre dans la suite de l'histoire ?

.....

.....

.....

Les réponses sous format QCM ont été saisies de manière automatisée et les questions ouvertes ont été corrigées par des experts via une interface Internet (cf. l'application « Agate » dans la partie 3 du présent rapport). Certaines questions, notamment celles constituant un ensemble de vrai/faux, ont été regroupées afin qu'une question à deux modalités de réponse ne pèse pas autant qu'une question à quatre ou cinq propositions.

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

1.3.2 Constitution des cahiers

La méthodologie des cahiers tournants permet d'évaluer un nombre important d'items sans allonger le temps de passation. Les items sont ainsi répartis dans des blocs d'une durée de 20 minutes et les blocs sont ensuite distribués dans les cahiers tout en respectant certaines contraintes telles que chaque bloc devant se retrouver un même nombre de fois au total et chaque association de blocs devant figurer au moins une fois dans un cahier. Ce dispositif, couramment utilisé dans les évaluations bilans, notamment les évaluations internationales, permet d'estimer la probabilité de réussite de chaque élève à chaque item sans que chaque élève ait passé l'ensemble des items (cf. tableau 2).

Au final, pour l'évaluation CEDRE CLL 2015, une partie de l'échantillon a passé un cahier numéroté de 1 à 13 qui comprend deux séquences : quatre blocs de compréhension de l'écrit de 75 minutes et un questionnaire de contexte de 30 minutes, identique dans tous les cahiers, dans lequel l'élève doit renseigner plusieurs éléments concernant l'environnement familial dans lequel il évolue,

Tableau 2 – Répartition des blocs dans les cahiers pour l'évaluation CEDRE CLL collège 2015

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

ses projets scolaires et professionnels, sa perception de la discipline et de son environnement scolaire et ses stratégies de lecture.

L'autre partie de l'échantillon a passé un cahier numéroté 14 ou 15 qui comprend trois séquences : une épreuve de production écrite de 25 minutes, trois blocs de compréhension de l'écrit de 50 minutes et un questionnaire de contexte de 30 minutes.

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 papier (cahiers d'évaluation) 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).

1.3.3 Une épreuve numérique

Un sous-échantillon de 8 élèves pour chacune des classes de l'échantillon a passé une épreuve de compréhension écrite et orale sur support numérique de 50 minutes. Un module d'évaluation des compétences multimodales sur support numérique a été testé en 2014. Cette épreuve comprend des items issus de l'évaluation 2013 de la Depp, Lecture sur écran (LSE) en fin de troisième, et de nouveaux items spécifiques à ce support (site ou vidéo).

1.4 Passation des évaluations

La passation de l'évaluation finale a eu lieu en mai 2015. Cette évaluation a été précédée d'une expérimentation l'année précédente, en mai 2014, de façon à tester un grand nombre d'items auprès d'un échantillon réduit d'établissements. Dans chaque établissement, une personne a été désignée comme étant l'administrateur du test, son rôle étant de veiller au strict respect de la procédure à suivre pour que l'évaluation soit passée dans les mêmes conditions quel que soit l'établissement.

Les séquence étaient séparées par 5 minutes de pause. L'anonymat des élèves et

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

2 Sondage

2.1 Méthodes

2.1.1 Tirage équilibré de classes de 3e

De manière générale, pour le secondaire, deux options de tirage peuvent être considérées : soit un sondage par grappe en sélectionnant un échantillon de classes et tous les élèves des classes tirées au sort participent à l'évaluation ; soit un premier degré qui concerne les établissements puis un second degré où un nombre d'élèves fixe dans chaque établissement est sélectionné [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

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.

Echantillon CEDRE 2015 CLL collège

Modalités de sélection

Le tirage est à deux degrés. Le premier degré de sondage est composé de classes (et non de collèges) tirées dans chaque strate. Le deuxième degré de sondage consiste à interroger tous les élèves de la classe sélectionnée (tirage par grappe).

Dans chacune des 3 strates, le tirage est équilibré sur les variables suivantes :
- Le nombre total d'élèves de 3e
- 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

Stratification

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

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

On vise environ 8 000 élèves.

Champ et exclusions

Pour l'année 2015, nous documentons le champ de l'évaluation qui est l'ensemble des élèves de 3e générale de collèges de France métropolitaine (tableau 3).

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

Tableau 3 – Exclusions pour la base de sondage (CEDRE 2015 CLL collège)

	Etab.	Elèves
Etablissements accueillant des élèves de 3e	8 384	840 546
On retire les EREA	8 314	839 256
On retire les étab hors contrat	8 153	837 081
On retire les COM	8 116	832 817
On ne garde que les collèges	6 924	803 010
On enlève les DOM	6 687	742 209
On enlève les classes avec moins de 10 élèves	6 669	741 615
Base CEDRE CLL 3e	**6 669**	**741 615**

Base de sondage

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

Tableau 4 – Répartition dans la base de sondage (CEDRE 2015 CLL collège)

strate	collèges	classes	élèves
1. Public hors EP	4 077	18 693	479 298
2. EP	978	4 536	102 995
3. Privé	1 614	6 042	159 322
Total	**6 689**	**29 271**	**741 615**

Échantillon

Le tableau 5 présente la répartition de l'échantillon dans les différentes strates. Au total, 327 classes ont été sélectionnées dans 319 établissements, rassemblant 8 051 élèves.

Tableau 5 – Répartition dans l'échantillon (CEDRE 2015 CLL collège)

strate	collèges	classes	élèves
1. Public hors EP	115	118	3 027
2. EP	129	133	3 023
3. Privé	75	76	2 001
Total	**319**	**327**	**8 051**

2.3 Etat des lieux de la non-réponse

2.3.1 Non-réponse totale

Parmi la non-réponse totale, nous distinguons selon la non-réponse de classes entières ou la non-réponse d'élèves dans les classes participantes. Les chiffres suivants ont été observés pour 2015. Tout d'abord, 94,5 % des classes de l'échantillon ont répondu à l'évaluation (tableau 6).

Tableau 6 – Non réponse des classes (CEDRE CLL collège 2015)

strate	N classes attendues	N classes répondantes	% de classes répondantes
1- public hors EP	118	115	97,5 %
2- EP	133	125	94,0 %
3- privé	76	69	90,8 %
Total	**327**	**309**	**94,5 %**

Au final, 85,2 % des effectifs attendus ont participé (tableau 7).

Tableau 7 – Non réponse globale (classes + élèves, CEDRE CLL collège 2015)

strate	N élèves attendus	N élèves répondants	% élèves répondants
1- public hors EP	3 027	2 708	89,4 %
2- EP	3 023	2 434	80,4 %
3- privé	2 001	1 722	86,1 %
Total	**8 051**	**6 864**	**85,2 %**

2.3.2 Valeurs manquantes et imputation

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

2.3.3 Non-réponse partielle et terminale

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

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

Dans le premier cas, les non-réponses sont traitées comme des échecs (code "0"). Le second cas conduit à déterminer des règles. Nous considérons que si un élève

a passé moins de 50 % d'une séquence, il n'a pas vu la séquence, les valeurs manquantes sont donc traitées de manière structurelle (code "s").

2015

Les cahiers élèves sont composés d'une séquence. La non réponse terminale a été étudiée par cahier. Parmi les élèves ayant de la non réponse terminale, pour la 1ère séquence, il y en a en moyenne 12,3.

Au final, pour 2015, on considère que :
- 221 élèves n'ont pas vu la séquence 1 dont :
  - 132 n'ont répondu à aucun item de la séquence
  - 89 ont répondu à moins de 50 % de la séquence

Les élèves dont les trois séquences sont codées en « s » sont considérés comme de la non réponse totale. C'est le cas pour 185 élèves.

2.4 Redressement

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

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

Le tableau 8 montre que le calage concerne principalement les élèves en retard, plus souvent absents à l'évaluation et donc moins nombreux dans l'échantillon que dans la population (15,6 % contre 17,4 %).

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

	Modalité ou variable	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population
Retard	1	115371.88	128825	15.56	17.37
	2	626243.14	612790	84.44	82.63
Sexe	1	363119.15	370161	48.96	49.91
	2	378495.87	371454	51.04	50.09
Strate	1	479297.98	479298	64.63	64.63
	2	102995	102995	13.89	13.89
	3	159322.04	159322	21.48	21.48

2.5 Précision

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

Tableau 9 – Score moyen et erreur standard associée (CLL collège 2015)

Année	Score moyen	Erreur standard
2015	250	1.60

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

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

Année	Groupe < 1	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5
2015	2.9	12.1	29.2	29.5	16.3	10.0

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

Année	Groupe < 1	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5
2015	0.30	0.64	0.75	0.73	0.62	0.63

Design effect

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

$$D_{eff} = \frac{V_P(\hat{\bar{Y}})}{V_{SAS}(\hat{\bar{Y}})} \tag{3}$$

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

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

Année	Erreur Standard	Erreur SAS	*Design Effect*
2015	0.92	0.51	3.23

Cela signifie qu'en 2015, un sondage aléatoire simple avec un effectif 3 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) :

$$Y_i^j = \begin{cases} 1 & \text{si l'élève } i \text{ réussit l'item } j \\ 0 & \text{si l'élève } i \text{ échoue à l'item } j \end{cases} \tag{4}$$

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

$$p_j = \frac{1}{n} \sum_{i=1}^{n} Y_i^j \tag{5}$$

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 :

$$S_i = \sum_{j=1}^{J} Y_i^j \tag{6}$$

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 :

$$\alpha = \frac{J\bar{r}}{1 + (J-1)\bar{r}} \tag{7}$$

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

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

Indices de discrimination

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

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

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

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

En effet, on peut montrer que

$$r_{bis-point}(j) = corr(Y^j, S) = \frac{\bar{S}_{(j1)} - \bar{S}_{(j0)}}{\sigma_S}\sqrt{p_j(1-p_j)} \tag{8}$$

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 :

$$y_{ij} = \begin{cases} 0 & \text{si } z_{ij} \leq \tau_j \\ 1 & \text{si } z_{ij} > \tau_j \end{cases} \tag{9}$$

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

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

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

3.2 Codage des réponses aux items

3.2.1 Valeurs manquantes

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

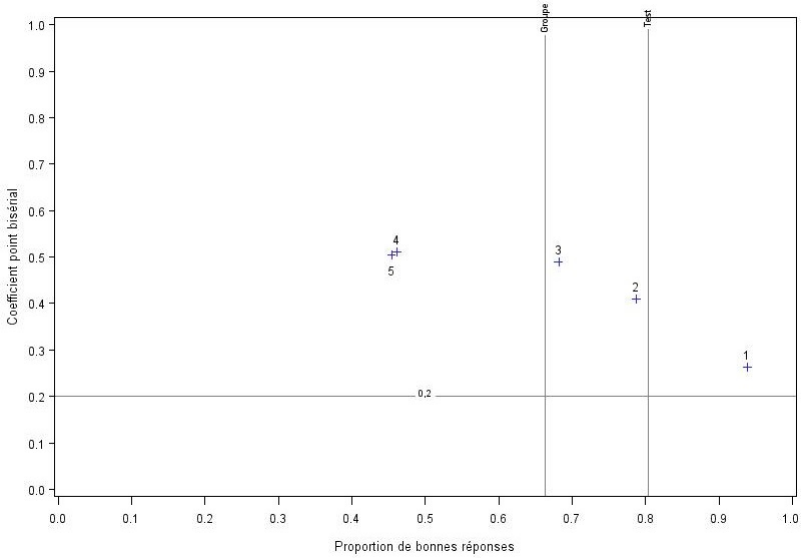
– Non-réponse terminale : l'élève s'est arrêté au cours de l'épreuve, potentiellement en raison d'un manque de temps. Des choix sont effectués pour déterminer le traitement de ces valeurs. Nous considérons que si un élève a passé moins de 50 % d'une séquence, il n'a pas vu la séquence, les valeurs manquantes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

3.2.2 Regroupement des items

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

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

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



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

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

Trois items sont apparus faiblement discriminants (i.e. *r-bis point* inférieurs à 0,2).

3.3.2 Dimensionnalité

Le tableau 13 présente les résultats de l'analyse factorielle des items effectuée sur l'année 2015.

Tableau 13 – Analyse en composantes principales (CEDRE CLL collège 2015)

	Valeur Propre	Différence	Proportion	Proportion cumulée
1	30.5	24.9	0.15	0.15
2	5.6	0.4	0.03	0.18
3	5.2	0.8	0.02	0.20

La structure des items est fortement unidimensionnelle : le « poids » de la première dimension est important (valeur propre de 30,5 contre 5,6 pour la deuxième dimension).

4 Modélisation

4.1 Méthodologie

4.1.1 Modèle de réponse à l'item

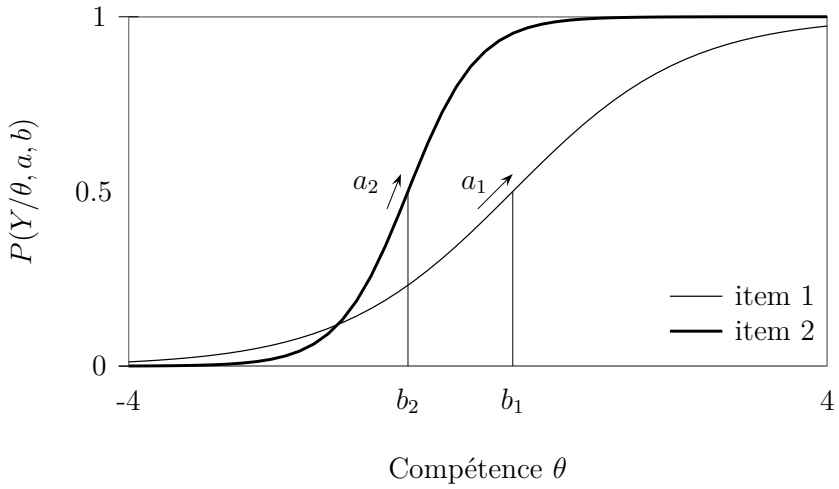
Le modèle de mesure utilisé est un modèle de réponse à l'item à deux paramètres avec une fonction de lien logistique (MRI 2PL) :

$$P_{ij} = P(Y_i^j = 1 \,|\, \theta_i, a_j, b_j) = \frac{e^{1,7a_j(\theta_i - b_j)}}{1 + e^{1,7a_j(\theta_i - b_j)}} \tag{10}$$

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 (10) 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 2015).

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

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).

s'écrit :

$$L(\mathbf{y}, \xi, \theta) = \prod_{i=1}^{n} \prod_{j=1}^{J} P_{ij}^{y_{ij}} [1 - P_{ij}]^{1-y_{ij}} \tag{11}$$

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 :

$$P(\mathbf{y} = \mathbf{y}_i) = \int_{-\infty}^{+\infty} P(\mathbf{y} = \mathbf{y}_i \,|\, \theta_i) g(\theta_i) d\theta_i \tag{12}$$

avec g la densité de θ.

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

$$L = \prod_{i=1}^{n} P(\mathbf{y} = \mathbf{y}_i) \tag{13}$$

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 (11).

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 :

$$B(\theta_i^{(ML)}) = \frac{-J}{2I^2} \tag{14}$$

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 :

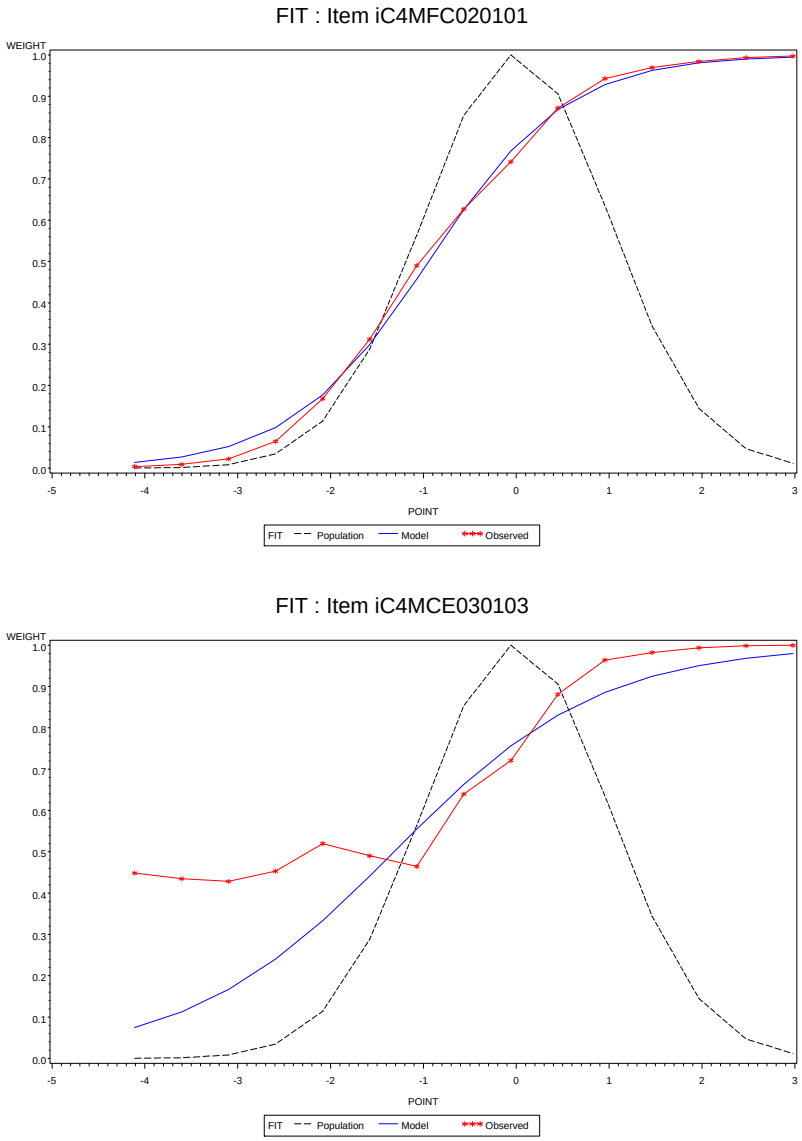
$$\frac{\partial lnL}{\partial \theta_i} + \frac{J}{2I} = 0 \tag{15}$$

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 :

$$z_{ij} = \frac{Y_i^j - P_{ij}}{\sqrt{P_{ij}(1-P_{ij})}} \tag{16}$$

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 :

$$Infit_j = \frac{1}{\sum_{i=1}^{n} w_{ij}} \sum_{i=1}^{n} w_{ij} z_{ij}^2 = \frac{1}{\sum_{i=1}^{n} w_{ij}} \sum_{i=1}^{n} (Y_i^j - P_{ij})^2 \tag{17}$$

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 :

$$P(Y \mid Z, G) \neq P(Y \mid Z) \tag{18}$$

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 :

$$I_j(\theta) = (1,7a_j)^2 P_j(\theta)(1 - P_j(\theta)) \tag{19}$$

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

4.2 Résultats

4.2.1 Identification des fonctionnements différentiels d'items (FDI)

L'évaluation des compétences langagières et de littératie portait sur une année de mesure. Par conséquent, l'analyse des FDI consistant à comparer les paramètres de difficulté des items repris, estimés de façons séparée pour les deux années, n'est donc pas applicable ici.

4.2.2 Identification des items présentant un mauvais ajustement (FIT)

L'analyse des ajustements (FIT) a permis de détecter trois items problématiques en 2015.

4.2.3 Bilan de l'analyse des items

L'analyse a conduit a retirer 3 items apparus faiblement discriminants (i.e. *r-bis point* inférieurs à 0,2) et un item problématique lors de l'analyse des ajustements (FIT). Sur les 202 items de départ, 198 ont donc été conservés.

5 Construction de l'échelle

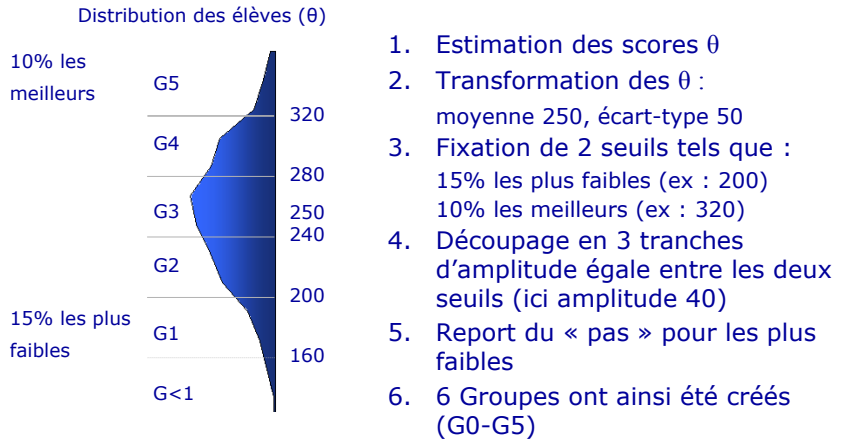
5.1 Méthode

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 CLL 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 2015. 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 (Dalibard, Fumel, & Lima, 2016).

Groupe < 1 (2,9 % des élèves)

Les élèves du groupe < 1 peuvent lire un schéma simple et y prélever une information facilement repérable. Bien que capables de répondre ponctuellement à quelques questions simples de prélèvement d'information explicite ou de compréhension à partir d'une carte ou d'un schéma simple, ces élèves, des lecteurs très en difficulté, ne maîtrisent aucune les compétences attendues.

Groupe 1 (12,1 % des élèves)

Les élèves du groupe 1 peuvent comparer plusieurs éléments dans un graphique. Bien que capables de répondre ponctuellement à des questions fermées de prélèvement d'information ou de compréhension à partir d'un texte littéraire court, les élèves ne maîtrisent pas les compétences attendues en fin de collège.

Groupe 2 (29,2 % des élèves)

Les élèves du groupe 2 sont capables de prélever une information explicite lorsque celle-ci est facilement accessible et repérable. Ils savent effectuer des inférences locales lorsqu'ils ont le choix entre plusieurs propositions (QCM). Ils répondent

ponctuellement à des questions ouvertes demandant une réponse chiffrée sans rédaction et sont capables d'exprimer une opinion personnelle sans justification sur des sujets de la vie quotidienne.

Groupe 3 (29,5 % des élèves)

Les élèves sont capables de prélever une ou plusieurs informations explicites en question ouverte y compris dans un texte long. Ils savent identifier le narrateur, les personnages d'un texte et s'appuyer sur de reprises nominales ou pronominales pour les identifier. Ils sont capables de déduire le sens d'un mot dans un contexte simple en choisissant entre plusieurs proposition d'un QCM. Ils savent identifier le genre ou la visée principale d'un texte et justifier leur opinion personnelle à partir de textes évoquant des situations de la vie quotidienne. Les élèves sont capables de lire un tableau à double entrée et de faire une déduction simple. Ils savent identifier la cause et la conséquence (récit ou compte rendu d'une expérience) dans une liste de propositions.

Groupe 4 (16,3 % des élèves)

Les élèves sont capables de prélever des informations non immédiatement repérables, d'expliciter les idées principales d'un texte en répondant à une question à réponse rédigée, de déduire le sens d'un mot rare ou spécialisé dans un contexte littéraire ou scientifique et d'argumenter leur point de vue sur un texte scientifique. Ils savent identifier un sous-genre littéraire et argumenter leur point de vue sur un texte. Les élèves sont capables de formuler une hypothèse à partir d'un tableau et d'un graphique et de comprendre une démarche expérimentale. Ils savent croiser les informations de deux documents de nature différente (texte et tableau) pour aboutir à une déduction complexe.

Groupe 5 (10,0 % des élèves)

Les élèves repèrent et comprennent les éléments implicites d'un texte. Ils perçoivent l'organisation logique et temporelle d'un texte et savent retrouver les étapes d'un raisonnement ou d'un récit. Ils sont capables de résumer ou de synthétiser un texte y-compris sous forme de carte mentale et de proposer une suite ou une conséquence. Les élèves sont capables de comparer deux tableaux à double entrée et de faire une déduction à partir de cette comparaison.

5.3 Exemples d'items

5.3.1 Item caractéristique du groupe <1

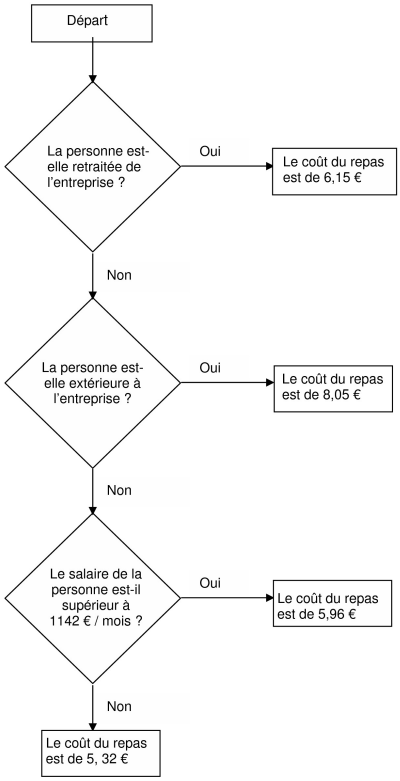
Cet item (figure 5) est réussi dès le groupe < 1. L'information à repérer est explicite sur un support qui comporte peu de texte, un organigramme, sans difficulté pour les faibles lecteurs.

Figure 5 – Exemple groupe < 1

RESTAURANT D'ENTREPRISE

Le restaurant de l'entreprise CROM applique des tarifs différents selon que la personne est retraitée de l'entreprise ou non, selon qu'elle fait partie du personnel de l'entreprise ou non, et selon son salaire.

Les différents tarifs pratiqués sont présentés dans l'organigramme ci-dessous :



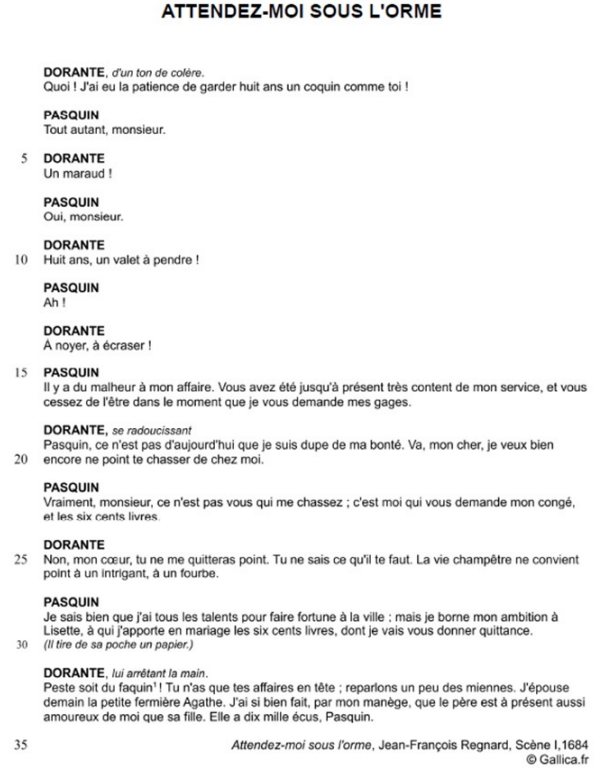
Question

Monsieur Simon est retraité de l'entreprise CROM. Quel est le prix de son repas lorsqu'il vient déjeuner au restaurant d'entreprise ?

1 ☐ 5,32 €
2 ☐ 5,96 €
3 ☐ 6,15 €
4 ☐ 8,05 €

5.3.2 Item caractéristique du groupe 1

Cet item (figure 6) est réussi à partir du groupe 1. Les 3 lignes de la série de vrai-faux doivent être réussis pour que l'item obtienne un code de réussite. La question appelle un repérage d'informations explicites dans un extrait de texte théâtral. Chaque affirmation permet de vérifier que l'élève est capable de manifester la compréhension globale de l'intrigue.



ATTENDEZ-MOI SOUS L'ORME

DORANTE, *d'un ton de colère.*
Quoi ! J'ai eu la patience de garder huit ans un coquin comme toi !

PASQUIN
Tout autant, monsieur.

DORANTE
Un maraud !

PASQUIN
Oui, monsieur.

DORANTE
Huit ans, un valet à pendre !

PASQUIN
Ah !

DORANTE
À noyer, à écraser !

PASQUIN
Il y a du malheur à mon affaire. Vous avez été jusqu'à présent très content de mon service, et vous cessez de l'être dans le moment que je vous demande mes gages.

DORANTE, *se radoucissant*
Pasquin, ce n'est pas d'aujourd'hui que je suis dupe de ma bonté. Va, mon cher, je veux bien encore ne point te chasser de chez moi.

PASQUIN
Vraiment, monsieur, ce n'est pas vous qui me chassez ; c'est moi qui vous demande mon congé, et les six cents livres.

DORANTE
Non, mon cœur, tu ne me quitteras point. Tu ne sais ce qu'il te faut. La vie champêtre ne convient point à un intrigant, à un fourbe.

PASQUIN
Je sais bien que j'ai tous les talents pour faire fortune à la ville ; mais je borne mon ambition à Lisette, à qui j'apporte en mariage les six cents livres, dont je vais vous donner quittance.
(Il tire de sa poche un papier.)

DORANTE, *lui arrêtant la main.*
Peste soit du faquin[1] ! Tu n'as que tes affaires en tête ; reparlons un peu des miennes. J'épouse demain la petite fermière Agathe. J'ai si bien fait, par mon manège, que le père est à présent aussi amoureux de moi que sa fille. Elle a dix mille écus, Pasquin.

Attendez-moi sous l'orme, Jean-François Regnard, Scène I, 1684
© Gallica.fr

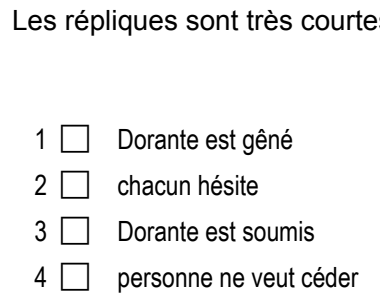
Figure 6 – Exemple groupe 1

	Vrai	Faux
Dorante veut chasser Pasquin de chez lui.	☐ 1	☐ 2
Dorante doit six cents livres à Pasquin.	☐ 1	☐ 2
Dorante va offrir dix mille écus à Pasquin pour son mariage.	☐ 1	☐ 2

5.3.3 Item caractéristique du groupe 2

Cet item (figure 7) est réussi à partir du groupe 2. Les élèves doivent faire des inférences locales et ils ont le choix entre plusieurs propositions (QCM). Ils doivent interpréter un fait stylistique en s'appuyant sur leur compréhension du texte.

Figure 7 – Exemple groupe 2



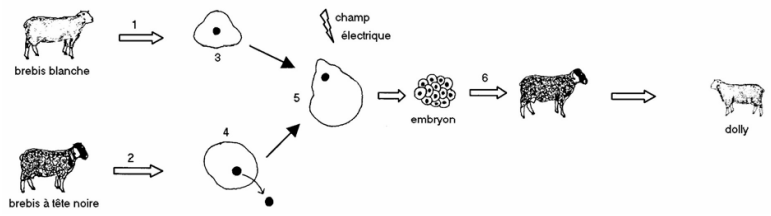
5.3.4 Item caractéristique du groupe 3

Cet item (figure 8) est réussi à partir du groupe 3. Les élèves doivent établir un lien entre une étape du croquis et une explication proposée (QCM). Ils doivent travailler à partir de deux supports : un texte documentaire court (10 lignes) qui contient des termes scientifiques et un schéma. Les connecteurs du texte facilitent la perception des différentes étapes du clonage. Le schéma reprend les informations données dans le texte mais il n'est pas possible d'identifier les cellules sans l'aide du texte.

Figure 8 – Exemple groupe 3

Dans le croquis ci dessous (document 2), la cellule 5 est :

Document 2 : les différentes étapes du clonage

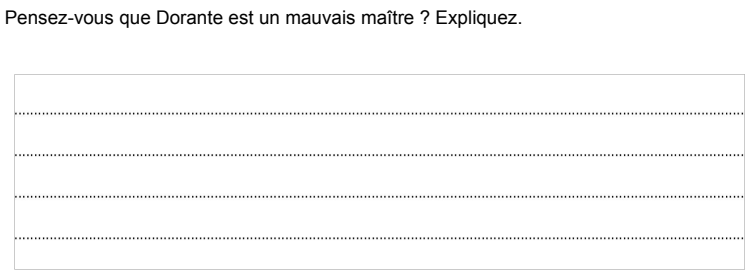


1 ☐ l'ovocyte prélevé chez la brebis à tête noire
2 ☐ la cellule de glande mammaire prélevée chez la brebis blanche
3 ☐ la cellule issue de la fusion entre la cellule de glande mammaire et l'ovocyte énucléé
4 ☐ la cellule qui contient le noyau de l'ovocyte

5.3.5 Item caractéristique du groupe 4

Cet item (figure 9) est réussi à partir du groupe 4. A ce niveau les élèves savent argumenter un point de vue sur un texte. Ils doivent rédiger un avis personnel sur l'attitude d'un personnage et le justifier.

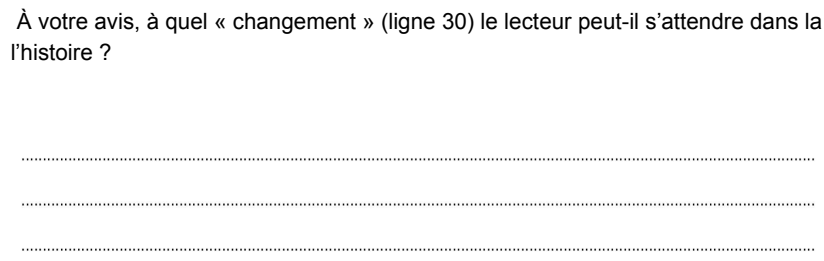
Figure 9 – Exemple groupe 4



5.3.6 Item caractéristique du groupe 5

Cet item (figure 10) est réussi à partir du groupe 5. Les élèves doivent repérer et comprendre les éléments implicites d'un texte afin de proposer une suite. Les élèvent doivent formuler à l'écrit une hypothèse sur la suite du récit en s'appuyant sur les indices du texte.

Figure 10 – Exemple groupe 5



6 Contexte et dimensions conatives

6.1 Variables sociodémographiques et indice de position sociale

Un certain nombre de variables sociodémographiques permettent d'enrichir l'analyse des résultats. Le score moyen des élèves est ainsi analysé en fonction du genre, du retard scolaire et quand les effectifs le permettent en fonction du secteur d'enseignement. Le lecteur est invité à consulter la Note d'Information pour plus de détails (Dalibard et al., 2016).

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

Pour l'échantillon de 2015, l'indice a été calculé pour chaque élève évalué. On obtient ainsi une distribution de cet indice qu'on découpe en cinq tranches égales, la première représentant les 20 % d'élèves les plus défavorisés. Pour chaque tranche, on calcule le score moyen en compétences langagières et littératie obtenu par les élèves correspondants (tableau 14).

Tableau 14 – Score moyen selon l'indice de position sociale des élèves (CEDRE compétences langagières et littératie 2015)

Élèves	Score moyen	Écart type
1re tranche	235	49
2e tranche	238	43
3e tranche	250	46
4e tranche	257	50
5e tranche	273	52

Note de lecture : en 2015, le score moyen des élèves les plus défavorisés (1ère tranche) est de 235.

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 statut (certifié, agrégé, contractuel ou autre). Ce questionnaire inclut aussi des questions sur les pratiques pédagogiques, les stratégies d'enseignement, le sentiment d'efficacité personnelle etc.

Le questionnaire élève interroge des dimensions dites conatives interessantes à mettre en lien avec le score obtenu à l'épreuve - les stratégies de lecture, la motivation, la perception de soi, l'anxiété scolaire etc. De plus, on demande aux élèves d'évaluer la difficulté de l'épreuve et leur degré d'implication à faire le test.

6.3 Construction des scores factoriels et des indicateurs

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

Le tableau 15 présente en guise d'illustration les items d'une de ces dimensions, la perception de soi.

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

Tableau 15 – Exemple de variable conative - perception de soi (CEDRE compétences langagières et littératie 2015)

Item	**1er axe ACP**
Je suis assez satisfait(e) de moi	0,77
J'ai l'impression que je suis très bon(ne) dans mon travail scolaire	0,72
En général, j'aime bien la façon dont je mène ma vie	0,72
J'aime bien le genre de personne que je suis	0,68
Je suis très content(e) d'être comme je suis	0,67
Je me sens aussi intelligent(e) que ceux de mon âge	0,64
Je fais très bien mon travail en classe	0,54
Je suis capable de me rappeler facilement des choses	0,51

Note de lecture : Les élèves devaient répondre à ces items sur échelle dite de Lickert, de Pas du tout d'accord à Tout à fait d'accord. Plus la valeur de l'indicateur est élevé, et plus grande est « l'adhésion » de l'élève à la dimension correspondante.

6.4 Motivation des élèves face à la situation d'évaluation

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

Plusieurs questions pour mesurer la motivation ont été proposées aux élèves. Une question concernant la motivation leur a été posée au début du cahier d'évaluation (cf. figure 11). De plus, un instrument pour mesurer la motivation a été ajouté à la fin du cahier élève. Cet instrument (cf. figure 12) a été adapté à partir du « thermomètre d'effort » proposé dans PISA (Keskpaik. & Rocher, 2015). Les données recueillies permettent d'évaluer le niveau de motivation initial de l'élève, 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 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.

Les tableaux 16 et 17 présentent les grands résultats de ces instruments.

Tableau 16 – Quel est votre degré de motivation pour faire cette évaluation ? (CEDRE compétences langagières et littératie 2015)

Avant le test :	**%**
Je suis très motivé(e)	7,9
Je suis motivé(e)	42,0
Je suis peu motivé(e)	36,1
Je ne suis pas du tout motivé(e)	14,0

Tableau 17 – Résultats de l'instrument de mesure de la motivation au test (CEDRE compétences langagières et littératie 2015)

Après le test :	**Moyenne**	**Erreur standard**
Difficulté perçue du test	5,1	0,05
Motivation au test	6,3	0,05
Motivation au test si les résultats comptaient pour le bulletin scolaire	8,8	0,03

Figure 11 – Question sur la motivation « initiale » (avant le test)

[Q1]

Quel est votre degré de motivation pour faire cette évaluation ?

(Cochez une seule case)

1 ☐ Je suis très motivé(e)
2 ☐ Je suis motivé(e)
3 ☐ Je suis peu motivé(e)
4 ☐ Je ne suis pas du tout motivé(e)

Figure 12 – Instrument de mesure de la motivation au test (après le test)

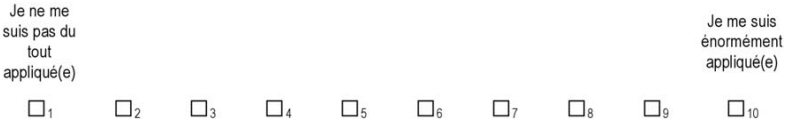
[Q1]

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

[Q2]

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

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



[Q3]

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)

7 Annexe

Certification AFNOR pour les évaluations CEDRE

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

Les finalités de la certification

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

Les enjeux pour la DEPP

Il y a deux enjeux forts pour la DEPP, l'un interne, l'autre externe :
- améliorer les processus de construction des instruments d'évaluation des acquis des élèves, fiabiliser ces processus par une démarche de contrôle-qualité ;
- valoriser l'enquête CEDRE comme un standard de qualité procédurale dans le domaine de l'évaluation.

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

La démarche qualité

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

Les engagements de service

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

Les engagements de service de la DEPP

Des objectifs clairs et partagés

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

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

Des évaluations fondées sur l'expertise pédagogique

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

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

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

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

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

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

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

Une mesure fiable et des comparaisons temporelles pertinentes

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

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

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

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

Des analyses enrichies par des données de contexte

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

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

Transparence des méthodes et partage des résultats

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

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

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

Références

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

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

Dalibard, E., Fumel, S., & Lima, L. (2016). CEDRE 2015 - nouvelle évaluation en fin de collège : compétences langagières et littératie. *Note d'information*, *21*.

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

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

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

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

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

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

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

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

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

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

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

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

Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54 n°3*, 427-450.

Liste des tableaux

Table des figures