CEDRE

Cycle des Évaluations Disciplinaires Réalisées sur Échantillons

Rapport technique

Langues vivantes étrangères 2022

École et Collège

Auteurs :

Stéphane BOUCÉ
Isabelle CIOLDI
Yann ETEVE
Marguerite GARNERO
Colin JOURDE
Damien LAFLAQUIÈRE
Florence LOZACHMEUR
Louis PHILBERT

Bureau de la conception et du pilotage des évaluations des élèves du premier degré
Bureau des études statistiques et psychométriques sur les évaluations des élèves
DEPP - Direction de l'évaluation, de la prospective et de la performance
Ministère de l'éducation nationale

Mai 2026.

Contents

Introduction

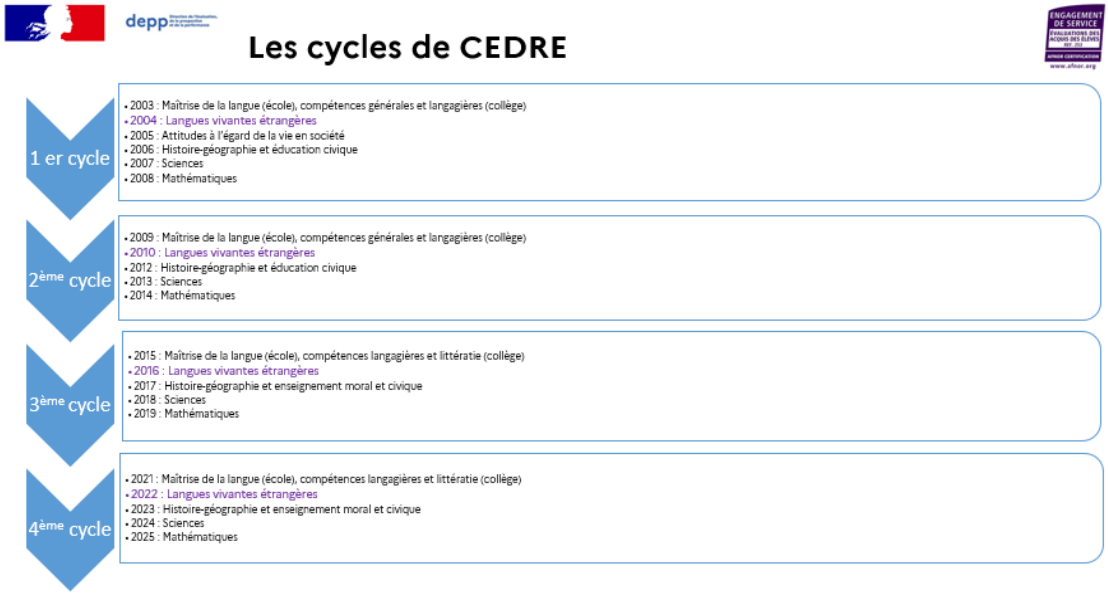
La Direction de l'Evaluation, de la Prospective et de la Performance (DEPP) met en place des dispositifs d'evaluation des acquis des eleves reposant sur des epreuves standardisees.

Elle est egalement maître d'oeuvre pour la France des evaluations internationales telles que PIRLS ou PISA. Ces programmes d'evaluations sont des outils d'observation des acquis des eleves pour le pilotage d'ensemble du systeme educatif (Trosseille Rocher, 2015).

Les evaluations du CEDRE (Cycle d'Evaluations Disciplinaires Realisees sur Echantillons) revelent ainsi, en reference aux programmes scolaires, les objectifs atteints et ceux qui ne le sont pas. Ces evaluations 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 caracterisees.

Leur methodologie de construction s'appuie sur les methodes de la mesure en education et sur des modelisations psychometriques. Ces evaluations concernent de larges echantillons representatifs d'etablissements, de classes et d'eleves. Elles permettent d'etablir des comparaisons temporelles an de suivre l'evolution des performances du systeme educatif.

Ce rapport presente l'ensemble des methodes qui sont employees pour realiser les evaluations du cycle CEDRE, en balayant des aspects aussi divers que la construction des epreuves, la selection des echantillons ou bien la modelisation des resultats. L'objectif est de rendre accessible les fondements methodologiques de ces evaluations, 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 certication des evaluations du cycle CEDRE


depp
Les cycles de CEDRE
1 er cycle
• 2003 : Maîtrise de la langue (école), compétences générales et langagières (collège)
• 2004 : Langues vivantes étrangères
• 2005 : Attitudes à l'égard de la vie en société
• 2006 : Histoire-géographie et éducation civique
• 2007 : Sciences
• 2008 : Mathématiques
2ème cycle
• 2009 : Maîtrise de la langue (école), compétences générales et langagières (collège)
• 2010 : Langues vivantes étrangères
• 2012 : Histoire-géographie et éducation civique
• 2013 : Sciences
• 2014 : Mathématiques
3ème cycle
• 2015 : Maîtrise de la langue (école), compétences langagières et littératie (collège)
• 2016 : Langues vivantes étrangères
• 2017 : Histoire-géographie et enseignement moral et civique
• 2018 : Sciences
• 2019 : Mathématiques
4ème cycle
• 2021 : Maîtrise de la langue (école), compétences langagières et littératie (collège)
• 2022 : Langues vivantes étrangères
• 2023 : Histoire-géographie et enseignement moral et civique
• 2024 : Sciences
• 2025 : Mathématiques

1 Cadre d'évaluation

1.1 Objectifs

Les évaluations CEDRE Langues vivantes Allemand et Anglais réalisées en fin d'école élémentaire ont pour objectif de faire le point sur les connaissances et compétences des élèves de CM2. Les évaluations CEDRE Langues vivantes Allemand, Anglais, Espagnol, réalisées en fin de collège, ont pour objectif de faire le point sur les connaissances et compétences des élèves de troisième.

Le renouvellement périodique de ces évaluations permet une analyse de l'évolution des acquis des élèves dans le temps. Il s'agit ici de la troisième prise d'informations. Les précédentes évaluations (2010 et 2016) ont été administrées sur format papier.

Les évaluations de 2022 présentent la particularité d'avoir été passées à la fois sous format papier et sous format numérique. Les items repris des évaluations antérieures ont été passés sur cahier. Tous les nouveaux items ont été conçus pour une passation numérique, sur ordinateur au collège et sur tablette à l'école.

Ces évaluations portent sur les acquis des élèves en compréhension de l'écrit et de l'oral. L'ensemble des items ont été respectivement utilisés pour la constitution de ces deux échelles de compréhension.

L'éclairage apporté par l'enquête peut concerner tous les acteurs ayant un intérêt dans le système éducatif français : décideurs, enseignants sur le terrain, formateurs et chercheurs.

1.2 Connaissances et compétences visées en fin d'école

En 2022, l'évaluation a été proposée dans deux activités langagières, la compréhension de l'oral et la compréhension de l'écrit. Cette évaluation a été élaborée à partir des objectifs fixés par les programmes nationaux (BOEN n° 11 du 26 novembre 2015), programmes adossés au cadre européen commun de référence pour les langues (CECRL). Les situations d'évaluation relèvent, pour la plupart d'entre elles, du niveau A1 (découverte) mais, afin d'apprécier au mieux les différentes performances des élèves, des items de niveau pré-A1 et d'autres d'un niveau tendant vers A2 (intermédiaire) ont également été proposés.

Le tableau 1 restitue les attendus en compréhension de l'oral et de l'écrit.

Table 1: Compétences évaluées en compréhension de l'oral et de l'écrit

Compétences	
Reconnaitre	- les lettres de l'aphabet (à l'oral uniquement) - des mots élémentaires et familiers - des expressions très courantes, des blocs lexicalisés
Identifier	- des éléments culturels simples - des informations explicites ciblées, des instructions - la situation d'énonciation, le thème, l'information essentielle, le type de document avec ou sans visuel
Construire le sens	- en mettant en relation des informations explicites - en faisant une inférence simple

L'évaluation de la compréhension de l'oral à l'école mesure avant tout la capacité des élèves à reconnaître des mots familiers ou expressions courantes et à dégager le sujet principal d'un bref message sonore. Les supports sélectionnés sont adaptés à l'âge des élèves (présence d'un grand nombre de comptines et de chansons).

En compréhension de l'écrit, les élèves sont évalués sur des compétences allant de la simple reconnaissance d'éléments lexicaux jusqu'à la construction du sens, en passant par l'identification d'informations spécifiques. Certaines situations sont accompagnées de visuels. Les élèves travaillent à partir de textes courts essentiellement descriptifs et narratifs. Des supports tels qu'une recette de cuisine, un extrait de site Internet, un poème assurent une certaine variété des épreuves et contribuent à la recherche d'authenticité des situations retenues.

1.3 Connaissances et compétences visées en fin de collège

Au collège, en 2022 l'évaluation a également été proposée dans deux activités langagières, la compréhension de l'oral et la compréhension de l'écrit. Elle a été construite à partir des objectifs fixés par les programmes nationaux (BOEN n°11 du 26 novembre 20215), adossés au CECRL. Les situations d'évaluation relèvent pour la plupart du niveau A2, mais certaines correspondent au niveau A1 et d'autres tendent vers B1 (indépendant), afin d'apprécier au mieux les différents niveaux des élèves.

Le tableau 2 restitue les attendus en compréhension de l'oral et de l'écrit

Table 2: Compétences évaluées en compréhension de l'oral et de l'écrit

Compétences	
Reconnaitre	- des mots élémentaires et familiers - des expressions très courantes, des blocs lexicalisés
Identifier	- des éléments culturels - des informations explicites ciblées, des instructions - la situation d'énonciation, le thème, l'information essentielle
Construire le sens	- déduire, inférer à partir de la mise en relation d'informations explicites - accéder à l'implicite, inférer le sens d'une expression simple - synthétiser à partir de la mise en relation d'indices de nature différente

En compréhension de l'oral, l'évaluation vise à vérifier que les élèves sont capables de repérer, dans un message sonore, des informations explicites (lexique de la vie quotidienne, éléments culturels simples, repères temporels, spatiaux) et de construire du sens en mettant ces informations en relation, en inférant à partir de l'explicite, en accédant à l'implicite, en synthétisant. Pour évaluer cette activité langagière, des enregistrements de différentes longueurs ont été proposés. Les supports sélectionnés étaient également de nature variée (utilisation d'extraits d'interviews, d'émissions de télévision, d'informations diffusées à la radio, de publicités, d'annonces destinées aux voyageurs, de conversations téléphoniques ainsi que de vidéos).

En compréhension de l'écrit sont mesurées les aptitudes des élèves à reconnaître dans un support écrit un lexique de la vie quotidienne et des expressions mémorisées, à identifier l'information pertinente (repères culturels, thème, repères temporels et spatiaux) et à construire le sens en mettant en relation des informations explicites, en inférant le sens d'une expression, en synthétisant. Les élèves ont été évalués à partir de textes littéraires (poèmes, extraits de romans et de contes), de textes informatifs issus de la presse ou de sites Internet (sommaire de magazine, annonce d'emploi), de textes argumentatifs (extraits de blogs, de forums). Les supports proposés visaient tous l'authenticité.

1.4 Construction du test

Le cycle complet de cette évaluation s'est déroulé sur six années au lieu de cinq en raison de la pandémie de COVID19.

Table 3: Déroulé de l'évaluation Langues vivantes étrangères 2022

Année 1 (2018-2019) : • Création d'items	• Définition du cadre de l'évaluation à partir des programmes. • Conception d'items au format numérique selon le cadre de l'évaluation.
Année 2 (2019-2020) : • Création d'items • Expérimentation reportée en raison de la pandémie de COVID19	• Poursuite de la conception d'items
Année 3 (2020-2021) : Évaluation • Création d'items • Études de comparabilité • Expérimentation des items au format numérique	• Poursuite de la conception d'items • Étude de comparabilité (Bridge Study) entre le format papier et le format numérique (mars-avril 2021). • Construction des expérimentations à partir des items créés. • Expérimentations des nouveaux items en mai 2021 auprès d'échantillons représentatifs d'élèves.
Année 4 (2021-2022) • Évaluation	• Analyse des résultats des expérimentations. • Montage des évaluations numériques à partir d'une sélection d'items validés après les expérimentations. • Montage des évaluations numériques à partir d'une sélection d'items validés après les expérimentations. • Passations des évaluations sur cahier et des évaluations numériques (sur ordinateur pour le collège et sur tablettes pour l'école) en mai-juin 2022 auprès d'échantillons représentatifs d'élèves.
Année 5 et 6 (2022-2023-2024) • Analyse et valorisation des résultats	• Analyse des résultats de l'évaluation. • Publication des Note d'Information . • Présentation des résultats lors d'un jeudi de la DEPP. • Préparation des documents d'analyse de situations d'évaluation.

1.4.1 Conception des items

Les items produits dans une évaluation CEDRE sont le fruit d'un travail collectif de concepteurs, encadrés par un chargé d'étude, personnel du Bureau de la conception et du pilotage des évaluations des élèves (DEPP B5), sous la responsabilité du chef de bureau. Pour le collège, les concepteurs sont

des professeurs d'allemand, d'anglais ou d'espagnol exerçant en collège ou en lycée dans différentes académies. Certains ont des fonctions de formateurs. Des experts issus de l'Inspection générale de l'Éducation, du Sport et de la Recherche (IGÉSR) peuvent participer aux réunions. Pour l'école, les concepteurs proviennent également d'académies variées. Ils sont généralement des professeurs des écoles, parfois des professeurs de collège enseignant l'allemand ou l'anglais en cycle 3, ainsi que des conseillers pédagogiques en langues vivantes.

Les concepteurs recherchent des supports selon le cahier des charges qui leur est fourni et proposent des items aux différents formats de questions demandés. Un équilibre de proportion entre les items considérés comme étant "faciles", "moyennement faciles" ou "difficiles" est recherché (correspondant pour l'allemand et l'anglais essentiellement au niveau A1 à l'école et pour l'ensemble des langues aux niveaux A1, A2 et tendant vers B1 du Cadre Européen au collège). Ces items font l'objet d'une relecture au sein du groupe et sont soumis à un cobayage dans les classes des concepteurs. Une expérimentation organisée auprès d'un échantillon national dans des conditions réelles de passation permet de valider les items sur le plan statistique. Les items retenus peuvent être intégrés à l'évaluation.

Une application ad hoc (GEODE) est utilisée en interne pour faciliter la production des items, les standardiser, les éditer, les stocker et gérer l'évaluation dans son ensemble.

1.4.2 Constitution du test

Le test est constitué d'un ensemble d'items organisés en blocs répartis dans des cahiers pour la séquence passée sur papier ou dans des modules pour la séquence proposée sur support numérique.

La séquence papier CEDRE 2022 est intégralement composée d'items repris à l'identique par rapport aux évaluations antérieures : ce sont les items d'ancrage.

L'évaluation numérique est entièrement constituée d'items nouveaux qui ont fait l'objet d'une expérimentation en 2021. Les items ont été créés pour une passation sur tablette à l'école et sur ordinateur au collège.

Chaque élève passe une séquence de 50 minutes organisée sur support papier et une séquence de 50 minutes sur support numérique. Dans le cadre de la séquence papier, tous les élèves d'une même classe passent la même séquence de compréhension de l'oral, mais des séquences différentes de compréhension de l'écrit. Dans le cadre de la séquence numérique, les élèves sont évalués à partir de modules différents tant en compréhension de l'oral que de l'écrit.

1.4.2.a À l'école

En allemand

L'évaluation sur papier CEDRE école allemand 2022 est constituée de 57 items pour la compréhension de l'oral, répartis dans 6 cahiers, et de 66 autres pour la compréhension de l'écrit, répartis dans 20 cahiers.

La séquence numérique comporte au total 105 items retenus sur la base de leur validité statistique après expérimentation en 2021. Elle est constituée de 38 items pour la compréhension de l'oral, répartis dans 4 blocs, et de 67 autres pour la compréhension de l'écrit, répartis dans 16 blocs.

Table 4: Cedre Allemand Ecole : Nombre d'items par activité langagières et année de conception

	Compréhension de l'oral						Compréhension de l'écrit					
	Papier		Numérique		Ensemble		Papier		Numérique		Ensemble	
	Eff	(%)	Eff	(%)	Eff	(%)	Eff	(%)	Eff	(%)	Eff	(%)
Ancien	57	100	28	74	85	89	66	100	18	27	84	63
Nouveaux	0	0	10	26	10	11	0	0	49	73	49	37
Ensemble	57	100	38	100	95	100	66	100	67	100	133	100

En anglais

L'évaluation sur papier CEDRE école anglais 2022 est constituée de 70 items pour la compréhension de l'oral, répartis dans 15 cahiers, et de 59 autres pour la compréhension de l'écrit, répartis dans 12 cahiers.

La séquence numérique comporte au total 139 items retenus sur la base de leur validité statistique après expérimentation en 2021. Elle est constituée de 67 items pour la compréhension de l'oral, répartis dans 7 blocs, et de 72 autres pour la compréhension de l'écrit, répartis dans 8 blocs.

Table 5: Cedre Anglais Ecole : Nombre d'items par activité langagières et année de conception

	Compréhension de l'oral						Compréhension de l'écrit					
	Papier		Numérique		Ensemble		Papier		Numérique		Ensemble	
	Eff	(%)	Eff	(%)	Eff	(%)	Eff	(%)	Eff	(%)	Eff	(%)
Ancien	70	100	23	34	93	68	59	100	22	31	81	62
Nouveaux	0	0	44	66	44	32	0	0	50	69	50	38
Ensemble	70	100	67	100	137	100	59	100	72	100	131	100

1.4.2.b Au collège

En allemand

L'évaluation sur papier CEDRE collège allemand 2022 est constituée de 64 items pour la compréhension de l'oral, répartis dans 10 cahiers, et de 60 autres pour la compréhension de l'écrit, répartis dans 12 cahiers.

La séquence numérique comporte au total 82 items. Elle est constituée de 45 items pour la compréhension de l'oral, répartis dans 5 blocs, et de 37 autres pour la compréhension de l'écrit, répartis dans 9 blocs.

Table 6: Cedre Allemand Collège : Nombre d'items par activité langagières et année de conception

	Compréhension de l'oral						Compréhension de l'écrit					
	Papier		Numérique		Ensemble		Papier		Numérique		Ensemble	
	Eff	(%)	Eff	(%)	Eff	(%)	Eff	(%)	Eff	(%)	Eff	(%)
Ancien	64	100	15	33	79	72	60	100	11	30	71	73
Nouveaux	0	0	30	67	30	28	0	0	26	70	26	27
Ensemble	64	100	45	100	109	100	60	100	37	100	97	100

En anglais

L'évaluation sur papier CEDRE collège anglais 2022 est constituée de 65 items pour la compréhension de l'oral, répartis dans 15 cahiers, et de 30 autres pour la compréhension de l'écrit, répartis dans 12 cahiers.

La séquence numérique comporte au total 171 items. Elle est constituée de 89 items pour la compréhension de l'oral, répartis dans 7 blocs, et de 82 autres pour la compréhension de l'écrit, répartis dans 8 blocs.

Table 7: Cedre Anglais Collège : Nombre d'items par activité langagières et année de conception

	Compréhension de l'oral						Compréhension de l'écrit					
	Papier		Numérique		Ensemble		Papier		Numérique		Ensemble	
	Eff	(%)	Eff	(%)	Eff	(%)	Eff	(%)	Eff	(%)	Eff	(%)
Ancien	65	100	16	18	81	53	30	100	9	11	39	35
Nouveaux	0	0	73	82	73	47	0	0	73	89	73	65
Ensemble	65	100	89	100	154	100	30	100	82	100	112	100

En espagnol

L'évaluation papier CEDRE espagnol 2022 est constituée de 75 items pour la compréhension de l'oral, répartis dans 10 cahiers, et de 121 autres pour la

compréhension de l'écrit, répartis dans 20 cahiers.

L'évaluation numérique CEDRE espagnol 2022 comporte au total 128 items. Elle est constituée de 55 items pour la compréhension de l'oral, répartis dans 5 modules, et de 73 autres pour la compréhension de l'écrit, répartis dans 16 modules.

Table 8: Cedre Espagnol Collège : Nombre d'items par activité langagières et année de conception

	Compréhension de l'oral						Compréhension de l'écrit					
	Papier		Numérique		Ensemble		Papier		Numérique		Ensemble	
	Eff	(%)	Eff	(%)	Eff	(%)	Eff	(%)	Eff	(%)	Eff	(%)
Ancien	75	100	22	40	97	75	121	100	28	38	149	77
Nouveaux	0	0	33	60	33	25	0	0	45	62	45	23
Ensemble	75	100	55	100	130	100	121	100	77	100	194	100

1.4.3 Formats d'items

Un type de format de questions est utilisé : les questions fermées (principalement QCM et tableaux séries, glisser-déposer, mise en ordre, menu déroulant)

Table 9: Cedre allemand : Répartition des items selon leur format en compréhension de l'écrit et de l'oral

		Compréhension de l'oral		Compréhension de l'écrit		Ensemble	
		Eff	(%)	Eff	(%)	Eff	(%)
Fin de CM2	QCM	60	63	68	51	128	56
	Tableau série	34	36	51	38	85	37
	Glisser-déposer	1	1	14	11	15	7
	Ensemble	95	100	133	100	228	100
Fin de 3e	QCM	59	54	39	40	98	48
	Tableau série	50	46	56	58	106	52
	Glisser-déposer	0	0	1	1	1	<1
	Menu déroulant	0	0	1	1	1	<1
	Ensemble	109	100	97	100	206	100

Table 10: Cedre anglais : Répartition des items selon leur format en compréhension de l'écrit et de l'oral

		Compréhension de l'oral		Compréhension de l'écrit		Ensemble	
		Eff	(%)	Eff	(%)	Eff	(%)
Fin de CM2	QCM	74	54	60	46	134	50
	Tableau série	43	31	59	45	102	38
	Menu déroulant	16	12	2	2	18	7
	Glisser-déposer	4	3	10	8	14	5
	Ensemble	137	100	131	100	268	100
Fin de 3e	QCM	98	64	47	42	145	55
	Tableau série	38	25	17	15	55	21
	Menu déroulant	18	12	44	39	62	23
	Glisser-déposer	0	0	4	4	4	2
	Ensemble	154	100	112	100	266	100

Table 11: Cedre espagnol : Répartition des items selon leur format en compréhension de l'écrit et de l'oral

		Compréhension de l'oral		Compréhension de l'écrit		Ensemble	
		Eff	(%)	Eff	(%)	Eff	(%)
Fin de 3e	QCM	92	71	98	54	190	61
	Tableau série	38	29	85	46	123	39
	Ensemble	130	100	183	100	313	100

1.4.3.a Le questionnaire à choix multiles (QCM)

Le QCM (questionnaire à choix multiples) permet d'investiguer un large champ de connaissances et de compétences dans un temps restreint et de ne pas conditionner la réussite de l'item à la subjectivité du correcteur. Les réponses choisies par les élèves ne sont pas ambiguës, puisqu'elles résultent d'un choix unique. Les compétences rédactionnelles de l'élève ne sont pas mises en jeu dans la validation de sa réponse. En multipliant les questions relatives à une même compétence, on fiabilise la prise d'information.

La question à choix multiples est constituée de deux composantes :

- l'amorce, qui pose la question ou définit le problème. L'amorce doit être simple et facilement compréhensible ;
- les 4 propositions, parmi lesquelles figure une seule réponse correcte ; les trois autres propositions sont appelées distracteurs.

1.4.3.b Le tableau série

Le format tableau série permet de varier les processus cognitifs attendus des élèves et parfois d'interroger un niveau plus détaillé de compréhension. Le score n'est pas calculé sur une seule réponse mais sur l'ensemble des propositions. Pour chaque tableau, un seuil de réponses correctes à partir duquel un crédit complet est accordé est défini. Ce seuil peut exiger une réponse correcte à chaque ligne.

1.4.3.c Le glisser-déposer

Format exclusivement réservé à la passation numérique, le glisser-déposer permet notamment l'évaluation des compétences d'ordonnancement ou de hiérarchisation des élèves.

1.4.3.d Le menu déroulant

Tout comme le QCM, le menu déroulant propose plusieurs options parmi lesquelles une seule est correcte. Son utilisation dans l'évaluation numérique offre la possibilité d'accéder au contenu sur une page restant claire et dégagée.

Pour des exemples d'items le lecteur est invité à consulter le document d'analyse des items libéres.

2 Sondage

2.1 Méthodes

2.1.1 Sondage par grappes stratifié pour le premier degré

Le tirage consiste en un sondage par grappes stratifié. La stratification porte sur le secteur de scolarisation et le classement en éducation prioritaire. Trois strates sont ainsi définies : écoles privées, publiques hors éducation prioritaire,et publiques en éducation prioritaire. Tous les élèves de CM2 des écoles sélectionnés participent. Le choix du sondage par grappes est motivé par la facilité de gestion. En effet, le fait de sélectionner tous les élèves d'une école permet d'éviter de mettre en place des procédures de tirage au sort d'élèves une fois les écoles tirées.

Par ailleurs, au moment du tirage de l'échantillon, les écoles ayant déjà été sélectionnées pour une autre évaluation la même année sont exclues 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 après élimination de la base des échantillons précédemment tirés).

2.1.2 Tirage équilibré de classes de 3e pour le second degré

De manière générale, dans l'enseignement secondaire, deux modalités de sondage stratifié peuvent être envisagées. La première correspond à un plan par grappes : au sein de chaque strate, un échantillon de classes est sélectionné et l'ensemble des élèves de ces classes participe à l'évaluation. La seconde correspond à un plan à deux degrés : les établissements constituent les unités primaires de sondage, tirées dans chaque strate, puis un nombre fixe d'élèves est sélectionné au second degré dans chacun des établissements retenus[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 ; cela facilite également l'organisation de l'évaluation pour le personnel de l'établissement.

On note U la population visée par une évaluation donnée, Y la variable d'intérêt

[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.

(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 :

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

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 à la fonction samplecube du package sampling sur R. 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).

2.1.3 Redressement de la non réponse : calage sur marges

Comme toute enquête réalisée par sondage, les évaluations des élèves sont exposées à la non-réponse. Bien que les taux de retour soient élevés, il est nécessaire de tenir compte de la non-réponse dans les estimations car celle-ci n'est pas purement aléatoire (par exemple, la non-réponse est plus élevée chez

les élèves en retard). Afin de la prendre en compte, un calage sur marges est effectué. La méthode de calage sur marges consiste à modifier les poids de sondage d_i des répondants de manière à ce que l'échantillon ainsi repondéré soit représentatif de certaines variables auxiliaires dont on connait les totaux sur la population (Sautory, 1993). C'est une méthode qui permet de corriger la non-réponse mais également d'améliorer la précision des estimateurs. En outre, elle a pour avantage de rendre cohérents les résultats observés sur l'échantillon pour ce qui concerne des informations connues sur l'ensemble de la population.

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

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

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

Tirage équilibré après élimination de la base des échantillons précédemment tirés

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

Formulation du problème et notations

Un échantillon S_1 a été tiré. Il est connu et les probabilités d'inclusion des établissements π_j^1 sont également connues. On souhaite alors tirer un échantillon S_2 dans la population U avec les probabilités π_j^2, mais sans aucun recouvrement avec l'échantillon S_1. On va donc tirer l'échantillon S_2 dans la population $U(S_1)$, c'est-à-dire la population U privée des

établissements de l'échantillon S_1 qui appartiennent à U. Notons d'emblée que S_1 n'a pas nécessairement été tiré dans U, mais potentiellement dans une autre population, plus large ou plus réduite ; cela n'affecte en rien la formulation envisagée ici. Notons également que l'indice j est utilisé ici : il concerne les établissements et non les élèves, représentés par l'indice i.

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

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

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

Équilibrage

On souhaite maintenant que l'échantillon S_2 soit équilibré selon certaines variables (nombre d'élèves en retard, etc.). Soit X une variable d'équilibrage, la condition s'écrit :

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

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

Ainsi, on aura :

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

Or, en espérance on a

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

La condition d'équilibrage initiale est donc remplie.

Condition fondamentale

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

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

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

2.1.4 Calcul de précision : méthode

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

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

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

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

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

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

2.2 Echantillonnage

Champ dans le premier degré
Le champ des évaluations CEDRE à l'école est celui des élèves de CM2 scolarisés dans des écoles publiques et privées sous contrat en France (hors Mayotte). Pour des raisons de coût, les écoles ayant moins de 6 élèves de CM2 sont exclues du champ.

Sont donc exclus du champ :

- Les COM.
- Les écoles hors contrat.
- Les écoles à l'étranger.
- Les écoles spécialisées.
- Les écoles de moins de 6 élèves de CM2.
- Mayotte.

Dans le premier degré, nous disposons d'informations sur les établissements scolaires, notamment le secteur d'enseignement et la langue enseignée.

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

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

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

1. Etablissement public hors éducation prioritaire rural
2. Etablissement public hors éducation prioritaire urbain
3. Etablissement public situé en rep

4. Etablissement public situé en rep+
5. Etablissement privé (PR)

Modalités de sélection
Dans le premier degré on effectue un tirage d'écoles, stratifié selon trois strates (variable strate2). Ensuite, tous les élèves de CM2 des écoles sélectionnées sont interrogés. C'est un sondage par grappes. Les écoles des échantillons des autres opérations de la DEPP ont été préalablement retirées de la base de tirage pour qu'il n'y ait pas de recouvrement.

Dans le second degré, il s'agit d'un sondage stratifié à deux degrés. Au premier degré, l'unité primaire de sondage correspond à la classe (et non à l'établissement), tirée dans chaque strate selon une allocation proportionnelle. Le deuxième degré de sondage consiste à interroger tous les élèves de la classe sélectionnée (tirage par grappe). La fonction samplecube du package sampling est utilisée pour garantir des échantillons équilibrés sur la base de sondage selon certaines variables

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

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

Dans le premier degré, la base de sondage du premier degré ne comporte pas d'informations auxiliaires telles que la PCS des parents par exemple. Il n'est donc pas possible de réaliser un tirage équilibré comme c'est le cas pour les évaluations CEDRE en 3e.

Echantillons 2022

Les échantillons tirent des élèves répartis proportionnellement selon les trois strates distribuées dans les trois matières et les deux niveaux comme ceci :

- 4 000 élèves pour l'allemand à l'école
- 4 000 élèves pour l'anglais à l'école
- 6 000 élèves pour l'allemand au collège
- 6 000 élèves pour l'anglais au collège
- 8 000 élèves pour l'espagnol au collège

Base de sondage

Le tableau 12 présente les exclusions dans la population ciblée dans le premier degré.

Table 12: Exclusions pour la base de sondage - Langues vivantes étrangères École

	Établissements	**Elèves**
Base initiale	49 644	831 684
On conserve les écoles avec des CM2	32 513	831 684
On retire les COM	32 513	831 684
On retire les écoles hors contrat	31 771	826 017
On retire les écoles spécialisées	31 750	825 541
On retire Mayotte	31 632	818 250
On retire les petites écoles < 6 CM2	29 456	810 498
Base CM2 CEDRE langues vivantes étrangères	**29 456**	**810 498**
Allemand		
Ecoles où tous les CM2 font allemand	939	24 270
On exclut les écoles déjà sélectionnées dans d'autres opérations	902	23 312
Base CM2 CEDRE allemand école	**902**	**23 312**
Anglais		
On retire les académies de Strasbourg et Nancy-Metz	27 555	762 856
On retire les écoles où tous les CM2 font allemand	27 518	761 552
On exclut les écoles sélectionnées dans d'autres opérations	26 217	720 407
Base CM2 CEDRE anglais école	**26 217**	**720 407**

Le tableau 13 présente les exclusions dans la population ciblée dans le second degré.

Table 13: Exclusions pour la base de sondage - Langues vivantes étrangères École et Collège

	Établissements	**Elèves**
Base initiale	11 766	6 114 224
On retire les COM	11 704	6 082 949
On retire les établissements hors contrat	10 717	6 021 915
On retire les EREA et les établissements hospitaliers	10 640	6 012 421
On retire les UPE2A pures	10 640	6 008 856
On retire les ULIS pures	10 639	6 004 457
On ne garde que les classes de 3e	6 940	808 852
On ne garde que les collèges	6 942	808 439
On ne garde que les 3e générales	6 938	786 309
On retire Mayotte	6 916	779 497
Base troisième CEDRE langues vivantes étrangères	**6 916**	**779 497**
Allemand		
Classes avec plus de 5 élèves en LV1 et/ou LV2 allemand	5 143	285 283
On exclut les établissements déjà sélectionnés dans d'autres opérations	4 451	111 785
Base troisième CEDRE allemand collège	**4 451**	**111 785**
Anglais		
Classes avec plus de 10 élèves en LV1 anglais	6 844	763 274
On exclut les établissements sélectionnés dans d'autres opérations	6 209	663 116
Base troisième CEDRE anglais collège	**6 209**	**663 116**
Espagnol		
Classes avec plus de 10 élèves en LV2 espagnol	6 470	685 230
On exlut les établissements déjà sélectionnés dans d'autres opérations	5 183	439 839
Base troisième CEDRE espagnol collège	**5 183**	**439 839**

Le tableau 14 présente la répartition de la population ciblée selon le secteur d'enseignement dans le premier degré.

Table 14: Répartition dans la base de sondage - Langues vivantes étrangères École

	Anglais		Allemand	
Strate	École	Eleves	École	Eleves
Public hors EP	19 327	503 272	807	18 766
EP	2 921	104 538	59	2 596
Privé	3 969	112 597	36	1 950
Total	**26 217**	**720 407**	**902**	**23 312**

Le tableau 15 présente la répartition de la population ciblée selon le secteur d'enseignement dans le second degré.

Table 15: Répartition dans la base de sondage - Langues vivantes étrangères collège

	Anglais		Allemand		Espagnol	
Strate	Établissements	Élèves	Établissements	Élèves	Établissements	Élèves
Public hors EP	3 798	419 706	2 806	72 498	3 155	279 517
EP	940	102 042	632	15 388	774	67 213
Privé	1 471	141 368	1 013	23 899	1 254	93 109
Total	**6 209**	**663 116**	**4 451**	**111 785**	**5 183**	**439 839**

Échantillon
Le tableau 16 présente la répartition de l'échantillon selon le secteur d'enseignement dans le premier degré. Au total, 298 écoles différentes ont été sélectionnés.

Table 16: Répartition dans l'échantillon - Langues vivantes étrangères école

	Anglais		Allemand	
Strate	Écoles	Élèves	Écoles	Élèves
Public hors EP	111	2 792	130	3 288
EP	17	647	12	543
Privé	20	631	8	325
Total	**148**	**4 070**	**150**	**4 156**

Le tableau 17 présente la répartition de l'échantillon selon le secteur d'enseignement dans le second degré. Au total, 1097 établissements différents ont été sélectionnés.

Table 17: Répartition dans l'échantillon - Langues vivantes étrangères collège

	Anglais		Allemand		Espagnol	
Strate	Établissements	Élèves	Établissements	Élèves	Établissements	Élèves
Public hors EP	144	3 757	302	3 819	225	5 079
EP	44	963	80	858	59	1 210
Privé	50	1 302	113	1 305	80	1 789
Total	**238**	**6 022**	**495**	**5 982**	**364**	**8 078**

2.3 État des lieux de la non-réponse

2.3.1 Non-réponse totale

Parmi la non-réponse totale, nous distinguons la non-réponse des établissements de la non-réponse des élèves des établissements participants. Les chiffres suivants ont été observés pour 2022.

Entre 97 % et 99 % des établissements des échantillons ont répondus aux évaluations à l'école (tableau 18). Áu collège ils sont entre 95 % et 98 % (tableau 19).

Table 18: Non-réponse des écoles - Langues vivantes étrangères école

Strate	Anglais			Allemand		
	Écoles attendus	Écoles répondants	% Écoles répondants	Écoles attendus	Écoles répondants	% Écoles répondants
Public hors EP	111	110	99	130	129	99
EP	17	17	100	12	11	92
Privé	20	20	100	8	6	75
Total	**148**	**147**	**99**	**150**	**146**	**97**

Table 19: Non-réponse des établissements - Langues vivantes étrangères collège

Strate	Anglais			Allemand			Espagnol		
	Étab attendus	Étab répondants	% Étab répondants	Étab attendus	Étab répondants	% Étab répondants	Étab attendus	Étab répondants	% Étab répondants
Public hors EP	144	140	97	302	297	98	225	215	96
EP	44	41	93	80	78	98	59	54	92
Privé	50	47	94	113	110	97	80	76	95
Total	**238**	**228**	**96**	**485**	**364**	**98**	**364**	**345**	**95**

Entre 96 % et 97 % des effectifs attendus ont participé aux évaluations du premier degré (tableau 20).

Table 20: Non-réponse des élèves - Langues vivantes étrangères école

Strate	Anglais			Allemand		
	Élèves attendus	Élèves répondants	% Élèves répondants	Élèves attendus	Élèves répondants	% Élèves répondants
Public hors EP	2 792	2 725	98	3 288	3 217	98
EP	647	595	92	543	483	89
Privé	631	627	99	325	272	84
Total	**4 070**	**3 947**	**97**	**4 156**	**3 972**	**96**

Entre 89 % et 95 % des effectifs attendus ont participé aux évaluations du second degré (tableau 21).

Table 21: Non-réponse des élèves - Langues vivantes étrangères collège

Strate	Anglais			Allemand			Espagnol		
	Élèves attendus	Élèves répondants	% Élèves répondants	Élèves attendus	Élèves répondants	% Élèves répondants	Élèves attendus	Élèves répondants	% Élèves répondants
Public hors EP	3 757	3 563	95	3 819	3 642	95	5 079	4 526	89
EP	963	901	94	858	787	92	1 210	1 011	84
Privé	1 302	1 187	91	1 305	1 270	97	1 789	1 675	94
Total	**6 022**	**5 651**	**94**	**5 982**	**5 699**	**95**	**8 078**	**7 212**	**89**

2.3.2 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 lors de sa passation.
- La non-réponse terminale : un élève s'est arrêté avant la fin de sa passation 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 "NA").

Table 22: Non-réponse des élèves - Langues vivantes étrangères

	Anglais		Allemand		Espagnol	
	Non-réponse papier	Non-réponse numérique	Non-réponse papier	Non-réponse numérique	Non-réponse papier	Non-réponse numérique
École	5%	7%	8%	10%	NA	NA
Collège	4%	5%	6%	3%	3%	2%

La non-réponse terminale a également été traitée par séquence.

- À l'école en anglais 1 % des élèves sont en non-réponse terminale sur les séquences papier et numérique

- À l'école en allemand 2 % des élèves sont en non-réponse terminale sur les séquences papier et numérique
- Au collège en anglais 4 % des élèves sont en non-réponse terminale sur la séquence papier et 5 % sur la séquence numérique
- Au collège en allemand 1 % des élèves sont en non-réponse terminale sur la séquence papier et 4 % sur la séquence numérique
- Au collège en espagnol 2 % des élèves sont en non-réponse terminale sur la séquence papier et 3 % sur la séquence numérique

2.4 Redressement

La correction de la non-réponse totale est effectuée en deux étapes. D'abord une post-stratification, selon les strates de sondage :

- 1. Public hors EP
- 2. REP
- 3. Privé sous contrat

Ensuite un calage sur marge est effectué en utilisant les deux variables de calage suivantes:

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

Table 23: Comparaison entre les marges de l'échantillon et les marges dans la population - Anglais école

	Modalité ou variable	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population	Pourcentage post-calage
Retard	1	38 328	56 540	5,2	7,4	7,4
	2	695 405	705 012	94,8	92,6	92,6
Sexe	1	380 368	388 745	51,8	51,0	51,0
	2	353 365	372 807	48,2	49,0	49,0
Strate	1	525 276	525 276	69,0	69,0	69,0
	2	119 570	119 570	15,7	15,7	15,7
	3	116 706	116 706	15,3	15,3	15,3

Table 24: Comparaison entre les marges de l'échantillon et les marges dans la population - Allemand école

	Modalité ou variable	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population	Pourcentage post-calage
Retard	1	948	1 802	4,0	7,4	7,4
	2	22 662	22 468	96,0	92,6	92,6
Sexe	1	12 104	12 389	51,3	51,0	51,0
	2	11 505	11 881	48,7	49,0	49,0
Strate	1	19 303	19 303	79,5	79,5	79,5
	2	2 900	2 900	11,9	11,9	11,9
	3	2 067	2 067	8,5	8,5	8,5

Table 25: Comparaison entre les marges de l'échantillon et les marges dans la population - Anglais collège

	Modalité ou variable	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population	Pourcentage post-calage
Retard	1	59 689	71 261	8,1	9,6	9,6
	2	674 657	672 003	91,9	90,4	90,4
Sexe	1	368 170	374 407	50,1	50,4	50,4
	2	366 176	368 857	49,9	49,6	49,6
Strate	1	463 588	463 588	62,4	62,4	62,4
	2	117 016	117 016	15,7	15,7	15,7
	3	162 660	162 660	21,9	21,9	21,9

Table 26: Comparaison entre les marges de l'échantillon et les marges dans la population - Allemand collège

	Modalité ou variable	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population	Pourcentage post-calage
Retard	1	7 363	8 234	5,8	6,4	6,4
	2	120 290	121 402	94,2	93,6	93,6
Sexe	1	67 287	67 331	52,7	52,7	52,7
	2	60 365	62 305	47,3	47,3	47,3
Strate	1	82 446	82 446	63,6	63,6	63,6
	2	18 117	18 117	14,0	14,0	14,0
	3	29 073	29 073	22,4	22,4	22,4

Table 27: Comparaison entre les marges de l'échantillon et les marges dans la population - Espagnol collège

	Modalité ou variable	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population	Pourcentage post-calage
Retard	1	46 560	54 397	8,3	9,6	9,6
	2	514 366	512 577	91,7	90,4	90,4
Sexe	1	275 899	282 424	49,2	49,8	49,8
	2	285 027	284 550	50,8	50,2	50,2
Strate	1	352 053	352 053	62,1	62,1	62,1
	2	88 790	88 790	15,7	15,7	15,7
	3	126 131	126 131	22,2	22,2	22,2

2.5 Précision

L'erreur standard de la moyenne (SEM) est calculée comme l'écart-type divisé par la racine carrée de la taille de l'échantillon :

$$SEM = \frac{\sigma}{\sqrt{n}} \tag{3}$$

où σ est l'écart type de la population et n est la taille de l'échantillon (tableau 29).

Table 28: Scores moyens et erreurs standard associées - Langues vivantes étrangères à l'école

	Année	Score CO	Écart-type CO	Score CE	Écart-type CE
Anglais école	2004	250	2,17	250	1,73
	2010	261	2,16	272	1,73
	2016	257	1,92	277	1,86
	2022	263	1,76	280	1,59
Allemand école	2004	250	3,39	250	3,49
	2010	254	1,82	262	1,75
	2016	255	2,16	263	2,41
	2022	250	1,74	254	2,18

Table 29: Scores moyens et erreurs standard associées - Langues vivantes étrangères à l'école

	Année	Score CO	Écart-type CO	Score CE	Écart-type CE
Anglais collège	2004	250	1,97	250	1,64
	2010	240	1,82	252	1,99
	2016	256	2,12	278	2,4
	2022	261	1,50	279	1,68
Allemand collège	2004	250	2,8	250	2,28
	2010	239	1,49	242	1,46
	2016	240	1,59	257	1,6
	2022	228	1,10	244	1,43
Espagnol collège	2010	250	1,72	250	1,49
	2016	247	1,35	256	1,27
	2022	249	0,89	255	0,96

Pour savoir par exemple si l'évolution entre 2016 et 2022 est significative , il faut calculer la valeur suivante :

$$\frac{\left|\hat{\bar{Y}}_{2022} - \hat{\bar{Y}}_{2016}\right|}{\sqrt{se^2_{\hat{\bar{Y}}_{2022}} + se^2_{\hat{\bar{Y}}_{2016}}}} \tag{4}$$

Par exemple en espagnol collège compréhension de l'écrit entre 2016 et 2022, on obtient une valeur de 0.7 (inférieure à 1.96). Cela signifie que l'évolution du score moyen n'est pas statistiquement significative.

3 Analyse des items

3.1 Méthodologie

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

3.1.1 Approche classique

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

Réussite et score

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

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

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

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

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

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

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

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

d'informations, cf. par exemple Laveault et Gregoire, 2002).

Fidélité

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

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

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

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

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

Indices de discrimination

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

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

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

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

En effet, on peut montrer que

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

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

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

3.1.2 Analyse factorielle des items

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

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

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

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

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

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

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

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

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

3.2 Codage des réponses aux items

3.2.1 Valeurs manquantes

Trois types de valeurs manquantes sont distinguées :

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

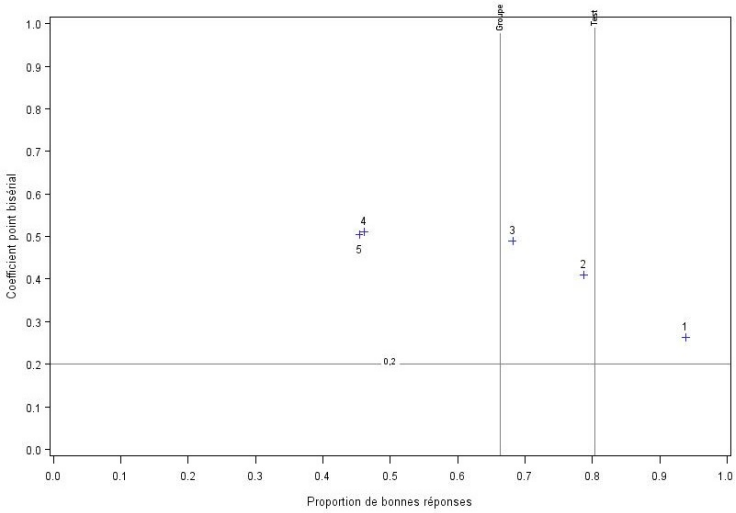
- Absence de réponse : l'élève a vu l'item mais n'y a pas répondu. L'absence de réponse est alors considérée comme une erreur de la part de l'élève.
- Non-réponse terminale : l'élève s'est arrêté au cours de l'épreuve, potentiellement en raison d'un manque de temps. Des choix sont effectués pour déterminer le traitement de ces valeurs. Nous considérons que si un élève a vu moins de 50 % du test alors nous ne disposons pas d'assez d'informations sur les compétences de cet élève, il est donc enlevé de l'analyse. Sinon elles sont traitées comme des échecs.

3.2.2 Regroupement des items

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

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

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



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

4 Modélisation

4.1 Méthodologie

4.1.1 Modèle de réponse à l'item

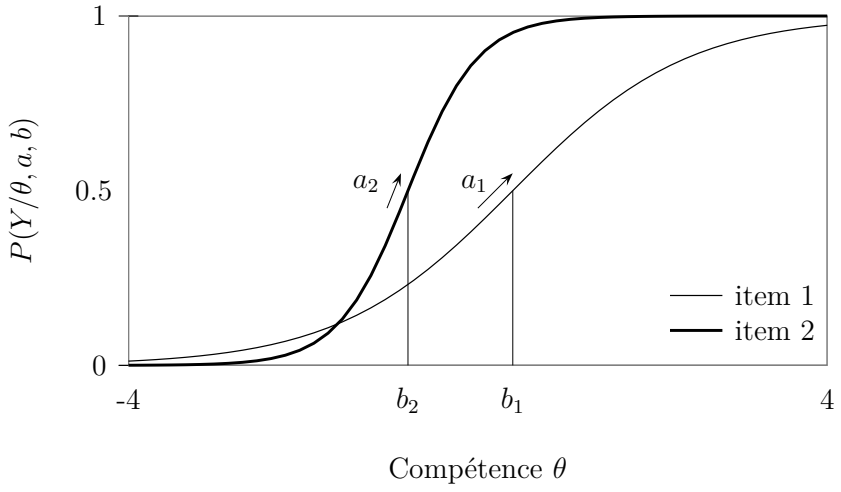
Le modèle de mesure utilisé est un modèle de réponse à l'item à deux paramètres avec une fonction de lien logistique (MRI 2PL):

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

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

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

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



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

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

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

4.1.2 Procédures d'estimation

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

Estimation des paramètres des items

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

Notons tout d'abord que les modèles présentés ne sont pas identifiables. En effet, les transformations $\theta_i^* = A\theta_i + B$, $b_j^* = Ab_j + B$ et $a_j^* = a_j/A$ avec A et B deux constantes $(A > 0)$, conduisent aux mêmes valeurs des probabilités. Dans CEDRE, nous levons l'indétermination en standardisant la distribution des θ pour les données du premier cycle (en l'occurrence, moyenne de 250 et écart-type de 50 pour l'année 2016).

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

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

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

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

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

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

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

avec g la densité de θ.

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

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

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

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

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

Estimation des niveaux de compétence

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

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

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

avec

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

et

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

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

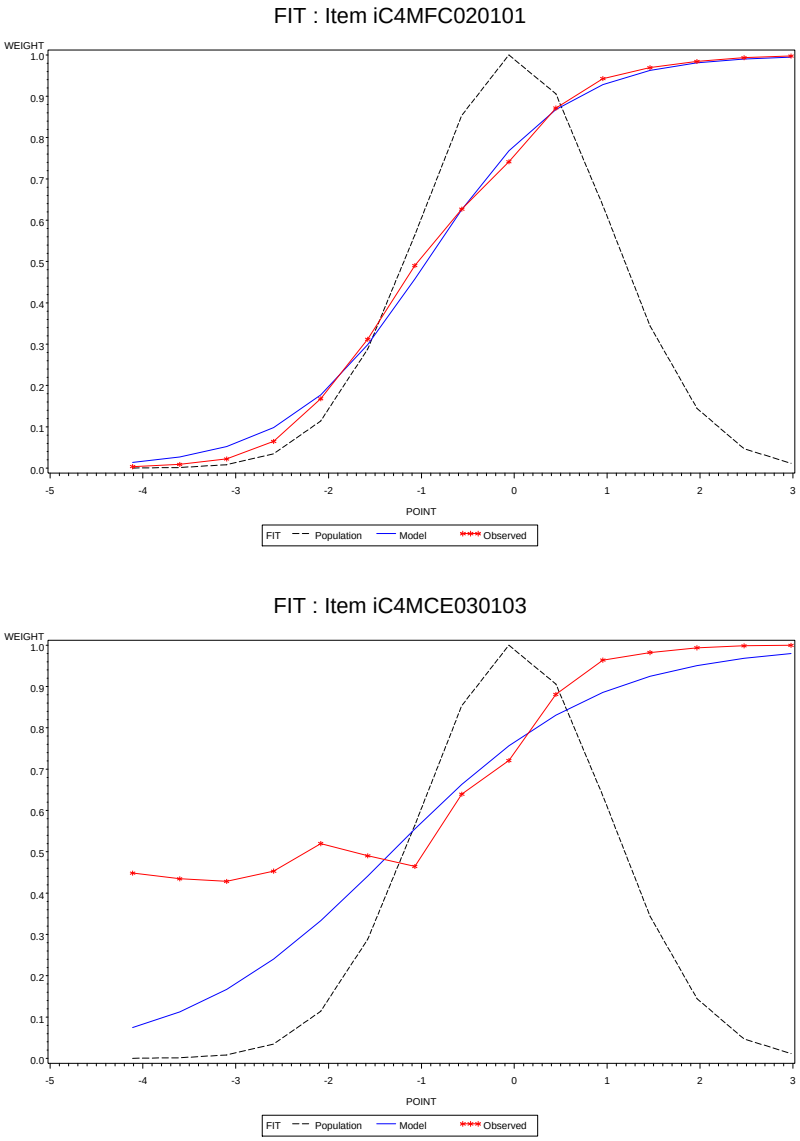
$$\frac{\partial lnL}{\partial \theta_i} + \frac{J}{2I} = 0 \tag{16}$$

4.1.3 Indice d'ajustement (FIT)

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

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

Figure 3: Exemples d'ajustements (FIT)



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

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

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

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

4.1.4 Fonctionnement Différentiel d'Item (FDI)

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

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

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

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

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

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

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

4.1.5 L'information du test

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

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

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

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

4.2 Résultats

4.2.1 Pouvoir discriminant des items

- En anglais à l'école 14 items, 5 items de 2016 et 9 de 2022, ont été éliminés pour cause de rbis-point trop faible en compréhension de l'oral. En compréhension de l'écrit ils sont 12, 8 items de 2016 et 4 de 2022.
- En allemand à l'école 5 items tous de 2022 ont été éliminés pour cause de rbis-point trop faible en compréhension de l'oral. En compréhension de l'écrit ils sont 20, 15 items de 2016 et 5 de 2022.
- En anglais au collège 4 items, 2 items de 2016 et 2 de 2022, ont été éliminés pour cause de rbis-point trop faible en compréhension de l'oral. En compréhension de l'écrit ce sont 3 items de 2016 et 1 de 2022 qui ont été éliminés.
- En allemand au collège 19 items, 13 de 2016 et 6 de 2022 ont été éliminés pour cause de rbis-point trop faible en compréhension de l'oral. En compréhension de l'écrit ils sont 7, 5 items de 2016 et 2 de 2022.
- En espagnol au collège 8 items, 7 de 2016 et 1 de 2022, ont été éliminés pour cause de rbis-point trop faible en compréhension de l'oral. En compréhension de l'écrit aucun item n'est enlevé.

4.2.2 Bilan de l'analyse des items

Table 30: Évolution du nombre d'items dans l'évaluation

		Anglais école	Allemand école	Anglais collège	Allemand collège	Espagnol collège
Compréhension de l'écrit	Initial	145	136	116	102	192
	dont ancrage papier	133	116	112	100	192
	dont nouveaux numériques	1	17	0	1	4
	mauvais Rbis papier	8	15	3	5	0
	mauvais Rbis numérique	4	5	1	2	0
	DIFF	1	17	0	1	4
Compréhension de l'oral	Initial	150	102	158	128	138
	dont ancrage papier	133	116	112	100	192
	dont nouveaux numériques	1	17	0	1	4
	mauvais Rbis papier	5	0	2	13	7
	mauvais Rbis numérique	9	5	2	6	1
	DIFF	1	17	0	1	4

4.3 Calcul des scores

Comme indiqué précédemment, une analyse conjointe des données des 2 années a permis d'estimer les paramètres des items, puis les niveaux de compétences θ des élèves. Afin de lever l'indétermination du modèle, la moyenne des θ a été fixé à 250 et leur écart-type à 50, pour l'échantillon de 2015. Le tableau 31 présente les résultats obtenus.

Table 31: Niveaux de compétences (moyennes des scores et écarts-types) Langues vivantes étrangères École et Collège

	2016		2022	
	Score	écart-type	Score	écart-type
Ecole Allemand CE	263	48,9	254	47,6
Ecole Allemand CO	256	38,9	250	36,1
Ecole Anglais CE	277	51	281	50,6
Ecole Anglais CO	257,5	45,7	263,3	47,6
Collège Allemand CE	257	51	245	48
Collège Allemand CO	240	40	228	33
Collège Anglais CE	278	74	279	68
Collège Anglais CO	256	55	261	55
Collège Espagnol CE	256	40	255	40
Collège Espagnol CO	247	38	249	34

5 Construction de l'échelle

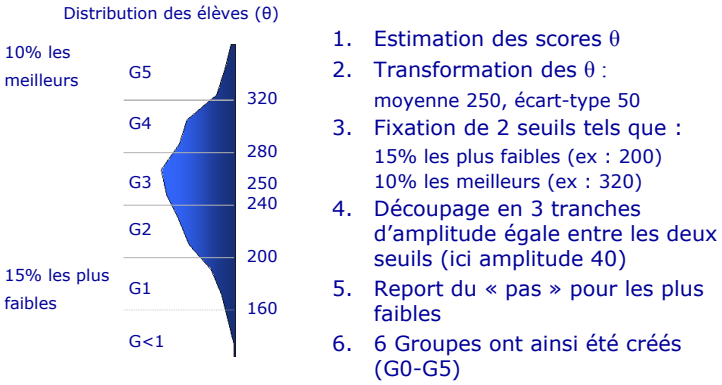
5.1 Méthode

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 Langues vivantes étrangères 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 2004. 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



6 Variables contextuelles et non cognitives

6.1 Variables sociodémographiques et indice de position sociale

Un certain nombre de variables sociodémographiques permettent d'enrichir l'analyse des résultats. Le score moyen des élèves est ainsi analysé en fonction du genre, du retard scolaire et quand les effectifs le permettent en fonction du secteur d'enseignement. Le lecteur est invité à consulter les Notes d'Informations pour plus de détails (Yann Eteve, Marguerite Garnero, Damien Laflaquière, Louis Philbert, & Hugo Rogie, 2022; Yann Eteve, Marguerite Garnero, Florence Lozachmeur, Louis Philbert, & Hugo Rogie, 2022; Stéphane Boucé, Yann Eteve, Marguerite Garnero, Louis Philbert, & Hugo Rogie, 2022).

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

Il n'a été possible d'établir des comparaisons qu'en termes de niveau social des écoles et établissements, et non au niveau individuel. En effet, la PCS des parents n'est pas disponibe au niveau individuel pour les écoles. Pour les établissement en 2022, la PCS des parents est disponible pour chaque élève, mais elle ne l'était pas dans le cycle antérieur. Pour chaque écoles et établissement des échantillons de 2016 et 2022, la moyenne de l'indice de position sociale a été calculée et la population a ensuite été découpée en quatre groupes selon les quartiles (tableau 6.1).

Indice moyen de l'école	Année	Score moyen	Ecart type
Groupe 1 (25% les plus défavorisés)	2016	259	**43**
	2022	**272**	50,5
Groupe 2	2016	276	49
	2022	277	49,8
Groupe 3	2016	277	51
	2022	278	47,1
Groupe 4 (25% les plus favorisés)	2016	294	53
	2022	291	51,1

Note de lecture : en 2022, le score moyen en compréhension de l'écrit en anglais des élèves appartenant au premier quart des écoles les plus favorisées (Groupe 1) augmente de 13 points par rapport à 2016. Les évolutions significatives sont indiquées en gras.

6.2 Élaboration des questionnaires de contexte

Afin d'enrichir l'analyse des résultats, un questionnaire de contexte a été élaboré. A l'école, les élèves ont été invités à remplir ce questionnaire à l'issue de l'évaluation au format papier. Les 20 questions qui le constituaient figuraient à la fin du cahier concernant la compréhension de l'écrit. Au collège, le questionnaire élève a été ajouté à la fin de l'évaluation numérique. Il se composait au total de 27 questions. Que le questionnaire ait été destiné à des écoliers ou des collégiens, il interrogeait les élèves sur des dimensions telles que leur scolarité, leur appétence pour la langue étrangère évaluée, leur pratique de cette langue en classe et en dehors de la classe.

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.

References

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

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

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

Rocher, T. (1999). *Psychométrie et théorie des sondages* (Unpublished master's thesis). 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* (Unpublished doctoral dissertation). 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.

Stéphane Boucé, Yann Eteve, Marguerite Garnero, Louis Philbert, & Hugo Rogie. (2022). Espagnol en fin de collège : une stabilité des résultats à l'oral comme à l'écrit. *Note d'information*, *24.39*.

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

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

Yann Eteve, Marguerite Garnero, Damien Laflaquière, Louis Philbert, & Hugo Rogie. (2022). Anglais en fin d'école et de collège : une hausse du niveau des élèves en compréhension de l'oral en 2022. *Note d'information*, *24.37*.

Yann Eteve, Marguerite Garnero, Florence Lozachmeur, Louis Philbert, & Hugo Rogie. (2022). Allemand en fin d'école et de collège : une baisse

significative du niveau des élèves en 2022. *Note d'information*, *24.38*.

Liste des tableaux

Liste des figures