LA MOTIVATION DES ÉLÈVES À RÉPONDRE À UN TEST STANDARDISÉ

Résultats d'une étude dans le cadre de Cedre compétences langagières et littératie

Sylvie Fumel et
Saskia Keskpaik,
MENESR-DEPP, bureau de l'évaluation des élèves

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 y participant. 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. Une étude expérimentale a été effectuée en 2015, visant à comparer les performances de deux groupes d'élèves : un groupe expérimental qui participe à l'évaluation ayant au préalable reçu l'information que l'épreuve sera notée et un groupe témoin qui passe le test dans les conditions habituelles d'une épreuve standardisée non notée. L'analyse des résultats dégage une tendance pour un plus grand investissement dans une évaluation aux enjeux élevés, mais les résultats restent à confirmer à plus grande échelle.

Alors que les évaluations standardisées des élèves sont des outils essentiels du pilotage des systèmes éducatifs, ces tests standardisés peuvent être considérés comme sans enjeux par les élèves évalués. Ils sont sans conséquence directe sur leurs résultats scolaires, ce qui pose régulièrement et de façon redondante la question de la motivation et de l'implication des élèves qui participent à ce type d'évaluation, d'autant plus que, dans le système éducatif français, la notation joue un rôle prépondérant et génère de l'anxiété.

La motivation est une construction complexe et multidimensionnelle. Elle constitue une variable latente qu'on ne peut évaluer qu'indirectement à travers des variables qui semblent l'engendrer et qui semblent être affectées par elle [O'Neil, Sugrue *et alii*, 1997]. Ces variables, dont la littérature sur le sujet est abondante [*ibid*], incluent entre autres les buts motivationnels [Nicholls, 1984], l'effort [Butler et Adams, 2007] et la performance [Uguroglu et Walberg, 1979]. La note obtenue à la fin du test peut être considérée comme un facteur motivationnel important qui, à travers de l'effort ou l'implication de l'élève, peut influencer sa performance.

Plusieurs études ont été menées depuis 2011, par la direction de l'évaluation, de la prospective et de la performance (DEPP) du ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche sur la question de la motivation des élèves à répondre à des

tests standardisés [Bobineau, 2013 ; Keskpaik et Rocher, 2012, 2015]. Les limites de la présente étude ne nous permettant pas de traiter les différents aspects de la notion de motivation dans toute leur complexité. Notre objectif est d'étudier le degré d'application de l'élève évalué. Pour le mesurer, un instrument a été adapté à partir du « thermomètre d'effort » utilisé dans le Programme international pour le suivi des acquis des élèves [PISA ; Keskpaik et Rocher, 2015]. Il s'agit d'échelles chiffrées unidimensionnelles, proposées aux élèves à la fin de leur cahier d'évaluation. Cet instrument de mesure est ajouté dans presque toutes les évaluations nationales menées par la DEPP depuis 2011.

Les données quantitatives recueillies à l'aide de l'instrument de mesure de la motivation, même si elles sont très informatives, ont aussi quelques inconvénients. Tout d'abord, il s'agit de déclaration des élèves, à la fin de l'évaluation, et ce jugement sur la motivation n'est pas indépendant de la difficulté du test [Keskpaik et Rocher, 2015]. Les élèves peuvent se déclarer plus ou moins motivés à passer le test selon qu'ils ont trouvé le test facile ou plutôt difficile. Déclarer des degrés de motivation relativement peu élevés peut constituer, pour certains élèves, une façon de « légitimer » leur réussite ou leur échec. Un élève qui se sent en difficulté pendant le test peut se déclarer peu motivé à la fin de la passation afin d'anticiper et d'expliquer, en partie, son faible score attendu.

Afin d'éviter les désavantages de données déclaratives sur la motivation *a posteriori*, nous avons souhaité étudier la motivation dans des conditions plus expérimentales. Ainsi, une étude a été organisée par la DEPP en 2014 dans le contexte de l'étude Cedre mathématiques ↘ **Encadré 1**. Les résultats, relativement modestes en ce qui concerne la généralisation statistique à cause du faible effectif de l'échantillon, mettent cependant en évidence une plus grande motivation des élèves face à une épreuve notée qu'une épreuve sans enjeux. Afin de vérifier si les tendances observées dans cette étude se confirment à plus grande échelle et dans d'autres domaines disciplinaires, cette expérimentation a été reconduite au printemps 2015.

MISE EN PLACE DE L'ÉTUDE, ÉCHANTILLON ET MÉTHODOLOGIE

La DEPP a mis en place un dispositif national d'évaluation, le Cycle des évaluations disciplinaires réalisées sur échantillons (Cedre), pour accompagner et suivre les changements du système éducatif. Ces évaluations suivent, au plus près, les évolutions du système scolaire et informent les décideurs des effets de ces changements. Elles s'inscrivent dans la durée et permettent de dresser un état de l'ensemble des compétences attendues par rapport aux programmes et d'analyser les facteurs d'efficacité des contextes dans lesquels se passent les enseignements. Ces évaluations concernent des échantillons représentatifs d'élèves de collèges publics et privés sous contrat. Un protocole d'évaluation Cedre compétences langagières et littératie motivation vise à comparer les performances de deux groupes d'élèves : un groupe expérimental qui reçoit au préalable l'information que l'épreuve sera notée et comptera dans la moyenne trimestrielle de français et un groupe témoin qui passe le test dans les conditions habituelles des évaluations standardisées réalisées par la DEPP.

Sous l'autorité pédagogique de l'inspection générale de Lettres et en coordination avec deux inspecteurs académiques-inspecteurs pédagogiques régionaux (IA-IPR) des académies de Grenoble et de Toulouse, la DEPP a mis en œuvre cette expérimentation. Afin de préparer la passation, les établissements ont reçu un courrier pour prendre connaissance du protocole.

Encadré 1

EXPÉRIMENTATION DE MOTIVATION DANS LE CADRE DE CEDRE MATHÉMATIQUES 2014

Une étude expérimentale a été effectuée en 2014, visant à comparer les performances de deux groupes d'élèves : un groupe expérimental qui participe à l'évaluation ayant au préalable reçu l'information que l'épreuve sera notée et un groupe témoin qui passe le test dans des conditions habituelles d'une épreuve standardisée non notée. 14 classes de troisième dans 9 établissements publics ont participé à cette étude, divisées en deux groupes à profil similaire en termes d'indice social et de taux d'élèves en retard scolaire. Le choix des classes est facilité par le réseau personnel et s'est fait à partir du volontariat des professeurs de mathématiques de ces classes.

Afin de préparer la passation, le groupe d'enseignants s'est réuni à la DEPP pour prendre connaissance du protocole. Les enseignants ont également reçu des consignes sur quelques points à observer lors de la passation de l'épreuve (comportement des élèves, utilisation du temps, etc.). L'épreuve s'est déroulée au cours du mois d'avril lors d'une séquence de travail de 45 minutes. Il s'agit d'une partie des épreuves Cedre mathématiques qui contient 33 items dont 4 ouverts. La correction des cahiers s'est faite lors d'une seconde journée à la DEPP. Chaque enseignant a corrigé les cahiers d'une classe autre que la sienne (mais similaire en termes du profil socio-économique) avec des consignes identiques à celles données lors de l'évaluation Cedre mathématiques.

Les observations effectuées par les professeurs montrent que la mise au travail des élèves n'a pas posé de difficultés. Les deux groupes ont été motivés par cette expérimentation. En moyenne, le groupe expérimental a utilisé les 45 minutes prévues pour l'épreuve alors que le groupe témoin a été nettement plus rapide (25 minutes en moyenne)[1].

Les élèves du groupe expérimental semblent ainsi prendre le test plus au sérieux, en parcourant le cahier plusieurs fois et relisant leurs productions.

Le taux de participation est plus élevé dans le groupe témoin : 89 % absents pour l'ensemble des élèves, 85 % dans groupe expérimental, 92 % dans groupe témoin. Les absents sont plus souvent les garçons (58 % contre 53 % parmi les élèves participants), issus des milieux socio-économiques moins favorisés (indice moyen – 0,14 contre 0,03) et en difficulté scolaire (24 % d'élèves en retard parmi les absents contre 12 % parmi les participants, note au diplôme national du brevet moyenne 16,5 sur 40 contre 18,9). Le score est plus élevé et moins dispersé dans le groupe expérimental. Après avoir contrôlé les caractéristiques individuelles et celles des classes (sexe, retard, indice de position sociale individuel et de classe, note au diplôme national du brevet), on observe un effet positif mais non significatif de l'expérimentation sur la performance des élèves. Ainsi, selon cette étude, les élèves semblent plus investis dans une évaluation aux enjeux élevés, mais on ne peut pas en conclure qu'ils réussissent significativement mieux lorsque l'évaluation « compte ». La taille insuffisante de l'échantillon rend difficile la généralisation des résultats. En outre, la sélection basée sur le volontariat suppose que les professeurs participants sont souvent plus impliqués dans le sujet d'étude et adoptent une approche particulière de la notation, ce qui peut être source de potentiels biais dans les résultats (certains des enseignants volontaires pratiquent uniquement une évaluation par compétences dans leur classe).

1. Temps mesuré lorsque les deux-tiers de la classe ont fini l'épreuve.

Les enseignants de français étaient destinataires d'un questionnaire d'observation lors de la passation de l'épreuve sur le comportement des élèves et l'utilisation du temps. Tous les élèves ont été évalués entre le 18 et le 29 mai 2015 lors d'une séquence de travail de quarante-cinq minutes, placée sur une heure de français de l'emploi du temps des élèves. Les élèves étaient en situation de devoir sur table avec leur professeur de français, ce qui n'est pas le cadre d'organisation des évaluations nationales (Cedre) ou internationales (PISA).

Le cahier de l'élève comprend deux épreuves issues de l'évaluation Cedre compétences langagières et littératie (CLL) 2015 : la première épreuve, 25 minutes de production écrite sur l'argumentation, et la seconde, 20 minutes de compréhension de l'écrit. Dans cette évaluation, quatre grands domaines de compétences sont évalués en compréhension de l'écrit (prélever une information, traiter et intégrer des informations, réfléchir et évaluer, expliquer et raisonner), et trois autres en production écrite (élaborer des contenus adaptés à la situation de communication, assurer l'organisation et la cohérence du texte et maîtriser les outils de la langue).

Suivant l'appartenance au groupe témoin ou au groupe expérimental, les élèves ont reçu les consignes suivantes avant de commencer la passation :

Consignes pour le groupe témoin

« Au cours des 45 prochaines minutes, vous allez répondre à un test.
Vous êtes sollicités pour participer à une évaluation qui permettra de connaître les performances des élèves de troisième. Ce travail ne donnera pas lieu à une note pour votre carnet scolaire et les résultats resteront anonymes.
Mais il est très important que vous fassiez de votre mieux.
Ce cahier est constitué d'une séquence d'écriture et d'une séquence de compréhension de l'écrit. »

Consignes pour le groupe expérimental

« Au cours des 45 prochaines minutes, vous allez répondre à un test.
Vous êtes sollicités pour participer à une évaluation qui permettra de connaître les performances des élèves de troisième. Ce travail donnera lieu à une note pour votre carnet scolaire et les résultats seront pris en compte pour le DNB (diplôme national du brevet).
Vous avez devant vous un CAHIER.
Ce cahier est constitué d'une séquence d'écriture et d'une séquence de compréhension de l'écrit. »

La correction des cahiers de l'académie de Grenoble s'est faite à Grenoble, lors d'une journée organisée par la DEPP et l'IA-IPR de Grenoble. Les correcteurs de Grenoble ont ensuite corrigé à domicile les cahiers de l'académie de Toulouse. Chaque enseignant a corrigé les cahiers d'une ou plusieurs autres classes que la sienne avec des consignes de codage identiques à celles de l'évaluation Cedre CLL. Il n'a pas été possible d'organiser la correction des cahiers de l'académie de Toulouse par les enseignants de cette académie pour des contraintes financières.

En tout, 42 classes ont été sélectionnées pour participer à cette expérimentation, soit au total environ 1 040 élèves. Les classes ont été divisées en deux groupes, expérimental et témoin, par un tirage aléatoire mais assurant l'équilibre dans la « composition » des deux groupes (niveau socio-économique et part des élèves redoublants). 40 classes parmi ces 42 ont participé

à l'étude. Deux classes parmi ces 40 ont passé l'évaluation en tant que membre du groupe témoin alors qu'elles étaient sélectionnées pour le groupe expérimental. Afin de conserver l'équilibre des groupes en termes des caractéristiques de contrôle (indice social moyen et pourcentage d'élèves en retard), elles ont été écartées de la plupart de nos analyses. L'échantillon participant retenu est ainsi composé de 38 classes dans 35 établissements.

Différentes méthodes d'analyse statistique ont été employées pour exploiter les données de cette étude expérimentale : statistiques descriptives, modèles de régression, analyse multiniveau et méthode d'appariement ↘ **Encadré 2**. Un score global a été calculé pour chaque élève en comptant toutes ses réponses correctes. Ce score a ensuite été standardisé sur la moyenne du groupe témoin.

Encadré 2

MÉTHODES D'ANALYSE

La méthode de régression quantile consiste à diviser la distribution de la performance des élèves en sous-groupes et à étudier l'effet de l'expérimentation sur ces groupes. Au lieu de procurer une estimation moyenne des paramètres pour l'ensemble de la distribution de scores, le modèle de régression quantile permet ainsi de calculer ces estimations séparément dans chaque subdivision de la distribution du score. Dans notre cas, la distribution du score est divisée en 10 intervalles ou parties selon les quantiles, chaque intervalle contenant le nombre égal d'élèves.

L'analyse multiniveau prend en compte la structure hiérarchique des données, en l'occurrence organisée sur deux niveaux : élève et établissement (la classe). L'emploi des modèles multiniveaux permet d'évaluer la variabilité des résultats à l'intérieur des établissements (entre les élèves) ainsi qu'entre les établissements. Dans l'analyse multiniveau, on peut distinguer entre les effets fixes du modèle, qui sont les coefficients de régression, et les effets aléatoires du modèle, c'est-à-dire les estimations de variance. Le coefficient de corrélation intraclasse montre la part de la variance totale du score qui peut être expliquée par les différences entre les établissements[2].

La méthode d'appariement (*matching*) analyse l'effet moyen de l'expérimentation. Cette méthode permet de réduire les différences entre les groupes qu'on compare et de limiter le biais dû au fait que le groupe témoin et le groupe expérimental ne sont pas complètement identiques en termes de composition. Cette méthode consiste à comparer, par paire, les scores des élèves de chacun de groupes qui sont similaires au regard des caractéristiques observables dont on dispose et dont on sait qu'ils influencent la performance (sexe, retard, indice de position sociale). Cette méthode permet de comparer ce qui est comparable – c'est-à-dire, comparer la performance de deux élèves ayant des profils similaires et dont la seule différence est l'appartenance au groupe expérimental ou au groupe témoin. Nous avons effectué l'appariement avec remise : la performance d'un élève du groupe témoin peut être comparée à celle de plusieurs élèves dans le groupe expérimental. L'analyse se fait en deux étapes : nous calculons d'abord le score de propension en utilisant un modèle logistique pour ensuite effectuer l'appariement sur ce score.
En contrôlant les caractéristiques des élèves (sexe, retard scolaire, indice de position sociale), nous calculons la différence moyenne entre les deux groupes.

2. Voir l'encadré sur les modèles multiniveaux dans Keskpaik et Rocher [2015]. Pour en savoir plus sur l'analyse multiniveau, le lecteur est invité à consulter l'ouvrage de Goldstein [2003].

NOTATION ET PERFORMANCE

Quelques statistiques descriptives

L'épreuve proposée aux élèves peut être considérée comme plutôt facile puisqu'un élève sur cinq a réussi 18 items ou plus sur un total de 20 items. La **figure 1** montre la répartition des scores bruts des élèves (*i.e.* le nombre de réponses correctes) et met en évidence l'asymétrie « positive » de cette distribution.

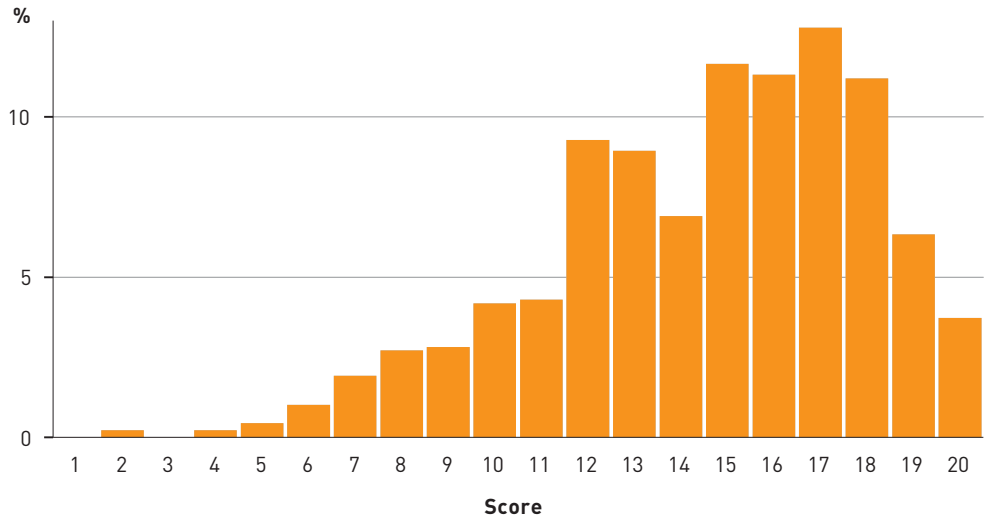
Avant de commencer à analyser en détail les effets possibles de l'expérimentation, la comparabilité des deux groupes a été étudiée. Ainsi, l'équilibre de ces groupes selon la répartition filles-garçons, le pourcentage des élèves redoublants, et le niveau socio-économique des élèves, a été analysé pour voir si l'équilibre visé lors du tirage de l'échantillon était maintenu. Le **tableau 1** présente les résultats de cette analyse pour les 42 classes échantillonnées (échantillon initial) et les 38 classes qui ont répondu et sont retenues dans les analyses. Les scores moyens, le nombre moyen de réponses manquantes et le nombre moyen des réponses manquantes aux questions ouvertes sont également présentés pour ces populations d'élèves.

Le **tableau 1** met en évidence que le groupe expérimental est dès le départ légèrement plus favorisé et cette tendance s'accentue dans l'échantillon participant. On note une moindre part des élèves en retard parmi les participants que dans l'échantillon initial tiré. Les garçons sont également relativement moins présents parmi les participants, surtout dans le groupe témoin. Les élèves en retard et les garçons ont été ainsi plus souvent absents lors de la passation des tests de cette étude. En comparant l'indice de position sociale des élèves [Rocher, 2016] et la répartition filles-garçons selon les groupes, on note que le léger déséquilibre initial s'amplifie. Les garçons et les élèves plus défavorisés socialement sont davantage absents dans le groupe témoin que dans le groupe expérimental. Ces élèves – plus souvent en difficulté en lecture [Daussin, Keskpaik, Rocher, 2011] – semblent ainsi plus enclins à participer à un test à « forts enjeux » qu'à une épreuve qui n'a pas de conséquence directe sur leurs résultats scolaires.

Les élèves du groupe expérimental ont obtenu un score moyen plus élevé que les élèves du groupe témoin ↘ **Tableau 1**. De plus, on observe un taux de non-réponse et plus spécifiquement un taux de non-réponse aux questions ouvertes moins élevé. On peut supposer qu'être noté amène les élèves à donner davantage de réponses. Cependant, lorsque nous étudions la non-réponse en prenant en compte les différences dans la composition des groupes, notamment du fait que le groupe expérimental est légèrement plus favorisé, nous observons que cette différence entre les deux groupes n'est pas statistiquement significative.

Différentes études internationales et nationales (PISA, Cedre, etc.) ont montré que les filles sont en moyenne plus performantes en lecture que les garçons. Notre expérimentation confirme ces résultats généraux. Néanmoins, la différence de scores entre les filles et les garçons est moins importante dans le groupe expérimental que dans le groupe témoin ↘ **Tableau 2**. Lorsqu'ils sont notés, les garçons feraient-ils relativement plus d'efforts ? On observe que la différence de scores entre les élèves « à l'heure » et ceux qui sont en retard par rapport à leur parcours scolaire tend à se renforcer dans le groupe expérimental. L'augmentation de l'enjeu du test semble ainsi impacter davantage les élèves « à l'heure » que les élèves en retard (les plus en difficulté). Nous allons, par la suite, employer des méthodes statistiques plus complexes qui permettent de tester ces hypothèses.

↘ Figure 1 Distribution du score



Éducation & formations n° 93 © DEPP

Lecture : 113 élèves, soit 12,8 % ont eu 17 réponses correctes sur 20.
Note : les données présentées sur cette figure portent sur l'ensemble de 884 élèves participants (y compris les élèves des deux classes ayant changé de groupe, puis qui sont écartés des analyses suivantes).

↘ Tableau 1 Statistiques descriptives

	Effectif	% en retard	% garçons	Indice de position sociale		Score brut		Score standardisé (sur groupe témoin)		Nombre de non-réponses		Nombre de non-réponses aux questions ouvertes	
				Moyen	Écart-type	Moyen	Écart-type	Moyen	Écart-type	Moyen	Écart-type	Moyen	Écart-type
Échantillon : ensemble	1 042	17,0	50,1	108	34								
Échantillon : groupe expérimental	530	17,7	52,5	110	36								
Échantillon : groupe témoin	512	16,2	47,7	107	32								
Répondants : ensemble	844	13,0	49,2	111	34	14,6	3,4	0,09	0,97	1,14	1,94	0,55	0,88
Répondants : groupe expérimental	415	13,0	53,3	115	35	14,9	3,3	0,19	0,93	0,94	1,77	0,47	0,79
Répondants : groupe témoin	429	13,1	45,2	107	33	14,3	3,5	0,00	1,00	1,34	2,06	0,62	0,95

Éducation & formations n° 93 © DEPP

Lecture : 17 % d'élèves sont en retard dans l'échantillon tiré contre 13 % parmi les répondants.
Parmi les répondants, les élèves du groupe expérimental ont eu, en moyenne, 14,9 et ceux du groupe témoin 14,3 réponses correctes (sur 20).
Note : indice de position sociale défini par Rocher [2016].

↘ Tableau 2 Score moyen selon le sexe et le retard scolaire

	Garçon	Fille	« À l'heure »	En retard
Ensemble	14,3	14,9	15,0	11,9
Groupe expérimental	14,9	15,0	15,4	12,0
Groupe témoin	13,7	14,7	14,6	11,9

Éducation & formations n° 93 © DEPP

Lecture : les garçons du groupe expérimental ont obtenu un score moyen de 14,9 (sur 20).

Analyses « toutes choses égales par ailleurs »

Afin de pouvoir séparer l'éventuel impact de l'expérimentation sur la performance des « effets de structure », nous avons employé différentes méthodes de régression[3] (**encadré 2** p. 109). Ces méthodes permettent d'isoler le lien entre la performance et une variable d'intérêt donnée (l'expérimentation de notation dans notre cas) et d'étudier ce lien « toutes choses égales par ailleurs », *i.e.* en maintenant constantes toutes les autres relations ajoutées dans le modèle[4]. À titre d'exemple, nous avons vu que les élèves issus des milieux socio-économiques défavorisés tendent à être davantage absents dans le groupe expérimental pour la passation de l'épreuve. On sait que ces élèves obtiennent des scores moins élevés [Daussin, Keskpaik, Rocher, 2011]. Ainsi, le score moyen plus élevé observé dans le groupe expérimental peut potentiellement être dû au fait que les élèves de niveau socio-économique défavorisé, avec des performances moyennes moins élevées, sont relativement moins représentés. Les méthodes « toutes choses égales par ailleurs » permettent d'éviter ce type de biais dus à la composition des populations. Dans notre exemple, nous pourrons vérifier à l'aide de ces méthodes si à un niveau socio-économique donné, le fait d'appartenir au groupe expérimental entraîne des scores plus élevés.

La **méthode de régression quantile** permet de découper la distribution du score en 10 groupes et d'étudier la relation entre l'enjeu du test et la performance selon différents niveaux de performance (**encadré 2** p. 109). Autrement dit, l'application de cette méthode permet d'analyser si le fait d'appartenir au groupe expérimental joue un rôle plus important parmi les élèves faibles, parmi les élèves forts, ou parmi ceux qui se situent au milieu de la distribution du score. On construit un modèle en régressant le score sur la variable de l'expérimentation, en contrôlant le sexe, le retard scolaire et l'indice de position sociale.

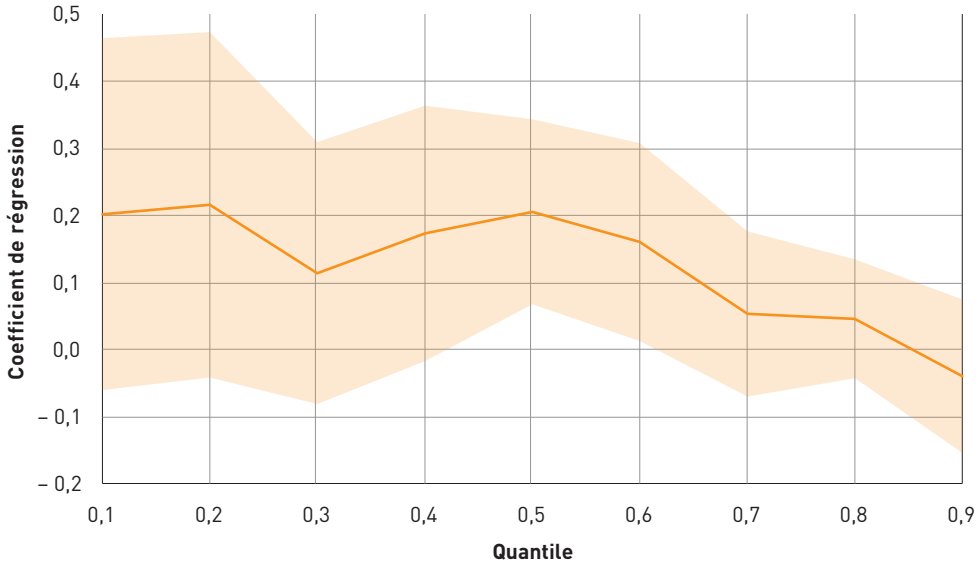
Les résultats de cette analyse montrent que l'effet de l'enjeu du test semble plus prononcé parmi les élèves moyens ↘ **Figure 2**. Cet effet est significatif dans les groupes de performance 4, 5 et 6, dans lesquels son ampleur s'étend de 0,16 à 0,20 écart-type de score. Les élèves moyens seraient ainsi plus sensibles aux enjeux du test, s'investissant davantage lorsque celui-ci a des conséquences directes sur leurs résultats scolaires.

Nous avons ensuite employé la **méthode d'appariement** (*matching*) qui permet de « corriger » le déséquilibre entre les deux groupes, témoin et expérimental, en termes de répartition filles-garçons, de pourcentage d'élèves en retard et de distribution de l'indice de position sociale, puis de comparer ainsi ce qui est comparable (**encadré 2** p. 109). Les résultats mettent

3. Les analyses qui suivent sont effectuées sur l'échantillon de 38 classes, soit 844 élèves répondants. Les deux classes qui ont changé de groupe ont été écartées, ainsi que 16 élèves ayant trop de valeurs manquantes. 415 de ces élèves (49 %) sont dans le groupe expérimental et 429 sont dans le groupe témoin.

4. « Toutes choses » incluent uniquement les variables de contrôle ajoutées dans le modèle.

↘ Figure 2 Coefficient de régression selon les quantiles de score



Éducation & formations n° 93 © DEPP

Lecture : parmi les élèves dont le score se situe dans le groupe 4 (*i.e.* la distribution d'élèves entre les quantiles 0,3 et 0,4), ceux appartenant au groupe expérimental obtiennent un score plus élevé de 0,17 écart-type par rapport à ceux faisant partie du groupe témoin.

en évidence un effet de 0,17 écart-type de score et cet effet est significatif au seuil de 0,01. En tenant compte des différences dans la composition des groupes, les élèves du groupe expérimental ont ainsi tendance à obtenir des scores légèrement supérieurs à ceux du groupe témoin.

Enfin, nous avons affiné nos analyses en construisant des **modèles multiniveaux**. Cette méthode permet de prendre en compte la structure « imbriquée » des données (les élèves qui sont regroupés dans des classes) et de distinguer l'influence des variables du niveau classe de celle des caractéristiques individuelles (du niveau élève) (**encadré 2** p. 109). On construit d'abord un modèle sans variables explicatives (un modèle vide) et on calcule le coefficient de corrélation intraclasse qui montre la part de la variance totale du score qui peut être expliquée par les différences entre les établissements. Ce coefficient vaut 0,13, ce qui veut dire que 13 % de la variabilité du score sont dus aux différences entre les classes. La part majeure de la variance du score est ainsi expliquée par des différences au sein de classes, c'est-à-dire entre les élèves.

Le groupe, témoin ou expérimental, a ensuite été ajouté au modèle en tant que variable explicative pour observer si cet ajout amène à une réduction de la variance du score entre classes **↘ Tableau 3** p. 115, modèle 2. Autrement dit, le fait d'appartenir au groupe expérimental ou témoin explique-t-il une partie de la variation du score d'un établissement à l'autre ? L'ajout de cette variable dans le modèle ne conduit pas à une réduction de la variabilité des scores aux niveaux classe et élève (la variance inter-classes passe de 0,14 à 0,13 et la variance inter-élèves passe de 0,86 à 0,87). Les indices de sélection de modèle (AIC, BIC) ne changent pas, ce qui indique que le pouvoir explicatif du modèle reste le même. Le modèle 2 ne rend ainsi pas mieux compte de la structure des données que le modèle 1. On observe un effet de 0,162 écart-type de score de cette variable, mais cet effet n'est pas statistiquement significatif.

Le modèle 3 inclut des caractéristiques sociodémographiques et scolaires des élèves telles que le sexe, le retard scolaire et l'indice de position sociale. On observe que les indices AIC et le BIC diminuent, ce qui indique une augmentation du pouvoir explicatif du modèle. Enfin, le modèle 4 intègre également quelques variables de classe – le pourcentage des élèves en retard et l'indice social moyen de la classe. On observe que les indices AIC et le BIC continuent à diminuer. Par rapport au modèle vide, le modèle complet (modèle 4) amène à une réduction de la variance inter-classes de 85 % (de 0,14 à 0,02) et de la variance inter-élèves de 10 % (de 0,86 à 0,77).

« Toutes choses égales par ailleurs », un garçon obtient un score de 0,141 écart-type moins élevé qu'une fille et un élève en retard un score de 0,614 écart-type moins élevé qu'un élève « à l'heure ». Une unité supplémentaire sur l'échelle de l'indice de position sociale amène à une augmentation de score de 0,006 écart-type. Le score varie également en fonction des caractéristiques d'établissements, le pourcentage d'élèves en retard et l'indice social moyen de la classe jouant un rôle sur le score. Les élèves provenant des classes où il y a relativement plus d'élèves en retard obtiennent des scores moins élevés, toutes choses égales par ailleurs. L'augmentation d'une unité de cet indicateur est associée à une perte de score de 0,011 écart-type. Plus l'indice de position sociale moyen de la classe est élevé, meilleure est la performance. Un gain de score de 0,006 écart-type est associé à l'augmentation d'une unité sur l'indice.

Les résultats de ces différents modèles statistiques mettent en évidence un rôle relativement modeste, mais non négligeable, de l'enjeu du test sur la performance. Notons aussi que l'ordre de grandeur de cet effet de l'expérimentation est comparable à celui observé lors de l'étude expérimentale en mathématiques (de 0,15 à 0,20 écart-type de score selon la méthode utilisée).

Qu'en est-il de la motivation des élèves ? Les élèves du groupe expérimental qui ont reçu les consignes expliquant que le test serait noté et compterait pour leurs résultats scolaires seraient-ils plus motivés ? Nous allons ensuite analyser le degré de motivation déclaré par les élèves ayant participé à l'étude.

MOTIVATION DIFFÉRENCIÉE SELON LE GROUPE ?

Motivation initiale

Après avoir reçu les consignes spécifiques à leur groupe et avant de commencer à répondre aux items cognitifs, les élèves étaient interrogés sur leur degré de motivation à passer le test. Une question a été intégrée au début de chaque cahier d'évaluation afin de mesurer la motivation initiale à l'aide d'une échelle à quatre positions. Les résultats de cet instrument sont présentés dans le **tableau 4**.

Les élèves du groupe expérimental se placent plus souvent que les élèves du groupe témoin, sur les positions extrêmes de l'échelle. Ils se déclarent plus souvent à la fois « très motivé(e)s » (14,2 % d'entre eux contre 10,2 % dans le groupe témoin) et « pas du tout motivé(e)s » (8,7 % contre 5,6 %). On ne peut ainsi pas conclure que l'expérimentation de la notation a un effet net et clair sur la motivation des élèves à répondre au test, telle que déclarée au début de l'évaluation.

↘ Tableau 3 Modèles multiniveaux

Paramètres	Modèle 1	Modèle 2	Modèle 3	Modèle 4
Effets fixes				
Constante	– 0,023 (0,068)	– 0,1 (0,092)	– 0,715*** (0,13)	– 1,135** (0,381)
Groupe expérimental		0,162 (0,133)	0,137 (0,098)	0,114 (0,081)
Garçon			– 0,131** (0,062)	– 0,141** (0,061)
Retard scolaire			– 0,648*** (0,094)	– 0,614*** (0,095)
Indice de position sociale			0,007*** (0,001)	0,006*** (0,001)
Pourcentage de retard scolaire, moyenne par classe				– 0,011** (0,005)
Indice de position sociale, moyenne par classe				0,006* (0,003)
Effets aléatoires				
Niveau 2 (classe) : variance des constantes	0,135*** (0,04)	0,127*** (0,039)	0,055** (0,022)	0,024** (0,014)
Niveau 1 : variance inter-élèves	0,865*** (0,043)	0,865*** (0,043)	0,767*** (0,038)	0,766**** (0,038)
AIC	2335,3	2335,9	2221,8	2208,3
BIC	2340,2	2342,4	2233,3	2223,1

Éducation & formations n° 93 © DEPP

Significativité : * au seuil critique de 10 % ; ** au seuil critique de 5 % ; *** au seuil critique de 1 %.
Lecture : toutes les autres variables comprises dans le modèle 4 tenues constantes, un garçon obtient un score de 0,141 écart-type moins élevé qu'une fille.
Note : les indicateurs de position sociale sont exprimés en points et le pourcentage de retard par classe en points de pourcentage. Les autres variables explicatives n'ont pas d'unité de mesure, mais une catégorie de référence (les élèves du groupe expérimental par rapport aux élèves du groupe témoin, les garçons par rapport aux filles, etc.). Les erreurs standards associées aux effets sont présentées entre parenthèses.

↘ Tableau 4 Répartition et score moyen selon la motivation initiale

Réponse des élèves à la question *« Quel est votre degré de motivation pour faire cette évaluation ? »*	Groupe expérimental			Groupe témoin		
	En %	Score moyen	Score moyen standardisé sur le groupe témoin	En %	Score moyen	Score moyen standardisé sur le groupe témoin
Je suis très motivé(e)	14,2	15,7	0,40	10,2	14,9	0,18
Je suis motivé(e)	49,3	15,0	0,20	54,5	14,7	0,11
Je suis peu motivé(e)	27,9	14,5	0,07	29,7	13,8	– 0,14
Je ne suis pas du tout motivé(e)	8,7	15,4	0,33	5,6	13,4	– 0,24

Éducation & formations n° 93 © DEPP

Lecture : 14,2 % des élèves du groupe expérimental se sont déclarés très motivés au début du test. Ces élèves ont obtenu un score moyen de 15,7 (sur 20).

En observant le score moyen par rapport au degré de motivation initiale, on note que, de manière générale, le score est d'autant moins élevé que la motivation initiale déclarée par élève est faible. Néanmoins, les élèves du groupe expérimental qui se déclarent « *pas du tout motivés* », ont obtenu un score comparable à celui des élèves de ce groupe ayant déclaré qu'ils sont très motivés.

Difficulté perçue du test et implication

Outre la question concernant la motivation initiale, quelques questions portant sur le test (sur la perception de la difficulté du test et sur la motivation au test) étaient posées aux élèves à la fin de l'évaluation[5]. Lorsqu'on compare les réponses à ces questions dans les deux groupes, on observe que les élèves du groupe expérimental se déclarent, en moyenne, plus impliqués ↘ **Tableau 5**.

Les élèves du groupe expérimental ont trouvé le test légèrement plus difficile que ceux du groupe témoin. Les résultats d'une régression linéaire sur la difficulté et l'implication (standardisés sur le groupe témoin) montrent que la difficulté perçue est liée à l'enjeu du test ↘ **Tableau 6**. Ceci reste vrai après avoir contrôlé les caractéristiques sociodémographiques et scolaires des élèves. Les élèves du groupe expérimental jugent ainsi le test plus difficile que ceux du groupe témoin, en maintenant constant le sexe, l'indice de position sociale et le retard scolaire. En revanche, l'implication n'a pas de lien significatif avec l'expérimentation lorsqu'on contrôle les variables de contexte.

GROUPE TÉMOIN *VERSUS* ÉCHANTILLON CEDRE COMPÉTENCES LANGAGIÈRES ET LITTÉRATIE

Afin de vérifier si les conditions de passation ont eu un effet sur les résultats, la méthode d'appariement a été utilisée pour comparer le groupe témoin de l'étude aux élèves ayant participé à l'évaluation Cedre CLL. Si les premiers ont passé l'épreuve dans les conditions habituelles du travail en classe pendant un cours de français, ces derniers ont participé à une évaluation standardisée dans des conditions particulières (épreuve plus longue que les contrôles habituels en classe, avec un superviseur autre que leur professeur de français, etc.).

En comparant les élèves du groupe témoin et ceux de l'évaluation Cedre CLL à propos du taux de non-réponse et du niveau d'implication (tel que déclaré à la fin du test), nous observons des différences significatives. D'une part, les élèves du groupe témoin laissent plus de questions sans réponse (différence de 0,32 écart-type de l'indicateur standardisé sur l'échantillon Cedre CLL, significatif au seuil de 0,01) ; d'autre part, ils se placent plus haut sur l'échelle d'implication (un effet de 0,15 écart-type de l'indicateur standardisé sur l'échantillon Cedre CLL, significatif au seuil de 0,10). Les élèves du groupe témoin de l'expérimentation se déclarent ainsi plus impliqués que ceux ayant passé l'évaluation Cedre CLL « habituelle » tout en accusant un taux de non-réponse plus élevé.

En revanche, aucune différence n'a été constatée dans les performances de ces deux groupes d'élèves.

5. « *Sur une échelle de difficulté allant de 1 à 10, comment avez-vous trouvé les exercices de cette évaluation ?* », « *Comment vous êtes-vous appliqué(e) pour faire cette évaluation ? Indiquez votre niveau d'application sur une échelle allant de 1 à 10* ».

↘ Tableau 5 Difficulté perçue du test par les élèves et implication déclarée

		Moyenne	Écart-type
Groupe expérimental	Difficulté perçue	5,0	1,9
	Implication	7,1	2,0
Groupe témoin	Difficulté perçue	4,7	1,8
	Implication	6,9	1,9

Éducation & formations n° 93 © DEPP

Lecture : en moyenne, les élèves du groupe expérimental ont jugé le test de difficulté 5 (sur 10).

↘ Tableau 6 Difficulté perçue du test et implication selon l'appartenance au groupe expérimental ou témoin et selon caractéristiques d'élèves

Variable	Coefficient de régression	
	Difficulté perçue	Implication
Constante	0,112	– 0,321**
Groupe expérimental	0,193**	0,129
Groupe témoin	réf.	réf.
Fille	0,110	0,329***
Garçon	réf.	réf.
En retard	0,113	– 0,186
« À l'heure »	réf.	réf.
Indice de position sociale	– 0,002	0,002*

Éducation & formations n° 93 © DEPP

Significativité : * au seuil critique de 10 % ; ** au seuil critique de 5 % ; *** au seuil critique de 1 %.
Lecture : les autres variables tenues constantes, un élève du groupe expérimental juge le test de 0,193 écart-type plus difficile qu'un élève du groupe témoin.

ÉLÉMENTS QUALITATIFS DE CONTEXTE

Les enseignants de français responsables de passation ont complété un questionnaire d'observation de passation, noté leurs commentaires sur l'épreuve et relevé les remarques de leurs élèves à la fin de la passation. Il apparaît que les deux groupes ont travaillé avec sérieux, qu'ils ont posé de nombreuses questions avant de commencer sur la façon d'écrire sur le cahier, de raturer ou pas, parce que l'aspect imprimé du cahier les « *intimidait* ».

Pour la première partie du cahier, la production écrite, les enseignants des deux groupes relèvent la même gestion du temps par les élèves : ils se mettent rapidement au travail. Le sujet proposé est un bon déclencheur d'écriture déjà testé auparavant. La moitié des élèves n'utilise que 15 minutes sur les 25 prévues. Les professeurs observent que les élèves écrivent au fil de la plume, sans révision du texte et, même s'il leur reste du temps, très peu relisent.

Les conditions de mise en œuvre et de passation spécifiques à cette évaluation ont favorisé, à chaque étape, l'implication et la motivation des adultes (inspectrices, chefs d'établissements, professeurs de français responsables de passation et correcteurs) et des élèves des deux groupes. En effet, à tous les échelons de la mise en œuvre, les liens entre la DEPP et les

intervenants entre eux ont été beaucoup plus fréquents, directs et suivis, par téléphone, courriel et réunions que lors de la mise en œuvre des évaluations Cedre avec échantillon national. Autres différences importantes à noter, le responsable de la passation était le professeur de français de la classe et l'épreuve était passée pendant une heure de cours de français de l'emploi du temps des élèves, soit les conditions d'un « devoir sur table », ce qui ne correspond pas aux conditions de passation d'une évaluation standardisée.

Il est à noter qu'à la fin de la passation certains enseignants ont souhaité prolonger le travail en engageant une discussion avec leurs élèves sur la question de la notation, « *une réflexion autour de la note et de son usage* » dit l'un d'eux.

CONCLUSION

Cette étude confirme les tendances et les ordres de grandeur de l'effet de notation observé lors de l'expérimentation Cedre mathématiques motivation, mais celui-ci n'est pas toujours significatif. On constate une différence de performance entre groupe expérimental et témoin d'environ 15 % et 20 % d'écart-type de score selon les méthodes, ce qui correspond à 7 à 8 points sur l'échelle de Cedre.

Les résultats restent à confirmer puisque les deux études Cedre motivation 2014 et 2015 rencontrent les mêmes limites d'ordre méthodologique et de taille insuffisante de l'échantillon.

Cette étude incite à la prudence dans les analyses et l'interprétation des tendances à la hausse ou à la baisse des résultats dans les comparaisons temporelles des évaluations Cedre. Il est important de tenir compte de la motivation des élèves et de contrôler si la mesure de la motivation reste stable ou varie afin de relativiser les constats.

L'évolution programmée de Cedre, avec le passage d'une évaluation papier à une évaluation numérique, permet de penser que le facteur motivation ne sera pas sans conséquence sur la performance des élèves, en distinguant les effets sur les filles et sur les garçons. Il sera important de le mesurer.

Afin de sensibiliser les élèves aux enjeux de ce type d'enquête, une information plus complète sur l'importance à accorder aux résultats des évaluations nationales (Cedre) et internationales (PISA) et à leur retentissement, leur permettrait d'en prendre toute la mesure. Au lieu de vivre ces évaluations comme une charge de travail supplémentaire, voire « une punition injuste » (pourquoi moi ?), les élèves de l'échantillon se sentiraient ainsi valorisés d'être sélectionnés dans un échantillon représentatif des élèves français en de fin de troisième.

↘ BIBLIOGRAPHIE

Bobineau M., 2013, *Évaluations dites « à faibles enjeux » : quelle perception et implication de la part des élèves ? Étude qualitative à partir de Cedre sciences 2013,* Rapport de stage, MEN-DEPP.

Butler J., Adams R. J., 2007, "The Impact of Differential Investment of Student Effort on the Outcomes of International Studies", *Journal of Applied Measurement,* vol. 8, n° 3, p. 279-304.

Daussin J.-M., Keskpaik S., Rocher T., 2011, « L'évolution du nombre d'élèves en difficulté face à l'écrit depuis une dizaine d'années », *France, portrait social,* Insee, p. 137-152.

Goldstein H., 2003, *Multilevel Statistical Models*, Third Edition, London, Arnold.

Keskpaik S., Rocher T., 2012, « Les évaluations à faibles enjeux : quel rôle joue la motivation ? Une expérience à partir de PISA », Communication dans le cadre du 24e colloque de l'Admée-Europe, Luxembourg.

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 », *Éducation & formations,* n° 86-87, MENESR-DEPP, p. 119-139.

Nicholls J. G., 1984, "Achievement motivation: Conceptions of Ability, Subjective Experience, Task Choice, and Performance", *Psychological Review,* n° 91, p. 328-346.

O'Neil H. F., Sugrue B., Abedi J., Baker E. L., Golan S., 1997, *Final Report of Experimental Studies on Motivation and NAEP Test Performance,* CSE Technical Report 427, Los Angeles, University of California.

Rocher. T., 2016, « Construction d'un indice de position sociale des élèves », *Éducation & formations,* n° 90, MENESR-DEPP, p. 5-27.

Uguroglu M. E., Walberg H. J., 1979, "Motivation and achievement: A Quantitative Synthesis", *American Educational Research Journal,* n° 16, p. 375-389.