conseil d'analyse économique

Focus

N° 027-2018
Octobre 2018

Impact de l'évaluation par compétence

Yann Algan, Jean Constantin, Samuel Delpeuch,
Élise Huillery et Corinne Prost(1)

1. Introduction

En sciences de l'éducation, en psychologie sociale et en sciences cognitives, une distinction est faite entre l'évaluation sommative et l'évaluation formative. L'évaluation sommative utilise une métrique commune – le plus souvent des notes – afin d'indiquer le niveau de performances des élèves. La métrique la plus classique est la note chiffrée, mais elle peut aussi reposer sur des lettres, des couleurs, ou toute autre échelle de valeurs comme des smileys. L'évaluation sommative a pour fonction l'attestation des apprentissages. *A contrario*, l'évaluation formative a pour fonction de favoriser la progression des apprentissages. Elle est réalisée en cours d'activité et vise à faire état des progrès des élèves ainsi que des difficultés rencontrées. L'évaluation formative repose sur deux piliers : premièrement le diagnostic, qui identifie les progrès effectués et les points à améliorer, et deuxièmement la remédiation, qui porte sur les moyens de dépasser les difficultés. L'évaluation formative repose sur des échanges entre élève et enseignant dans une double logique de développer la confiance en soi et le droit à l'erreur comme facteur de progrès, ces deux éléments constituant la base de la motivation de l'élève. Parce qu'elle se concentre sur l'apprenant plus que sur le groupe, la comparaison entre élèves est absente de l'évaluation formative, qui se focalise davantage sur les différentes performances de l'élève lui-même au cours du temps. Ainsi, l'évaluation formative valorise la correction, la progression et l'auto-évaluation, et diminue drastiquement la compétition entre élèves.

En théorie rien n'oppose ces deux formes d'évaluation qui servent deux objectifs distincts. Elles peuvent donc être utilisées de façon combinée et complémentaire, par exemple de façon continue pour l'évaluation formative afin d'encourager les élèves dans leur progression personnelle au quotidien, et de façon ponctuelle pour l'évaluation sommative lorsque l'enseignant veut faire le point sur le niveau d'acquisition de ses élèves. En pratique, comme l'évaluation sommative peut desservir les objectifs

(1) Respectivement : Doyen de l'École d'affaires publiques de Science Po, membre du CAE ; Assistant de recherche au CAE ; Chargé d'études au CAE ; Professeur à l'Université Paris Dauphine, membre du CAE ; Chercheuse affiliée au CREST, membre du CAE.

recherchés par l'évaluation formative, un dosage est nécessaire. L'évaluation sommative peut d'ailleurs être utilisée par l'enseignant sans pour autant que les résultats soient nécessairement communiqués aux élèves et aux familles, comme c'est le cas en Finlande jusqu'à l'âge de 11 ans.

En France, l'évaluation sommative domine fortement et ce depuis toujours. Tandis que d'autres pays comme le Royaume-Uni, la Finlande ou le Canada ont amorcé depuis les années 1980-1990 un rééquilibrage fort en faveur de l'évaluation formative, les enseignants français sont peu formés à ces pratiques et utilisent de façon quotidienne l'évaluation sommative, quelle qu'en soit la métrique. Un processus de réflexion sur l'évaluation des élèves a pourtant débuté dès 2000 dans une circulaire de Claude Allègre qui soulignait et regrettait la fonction « répressive » de la notation. En 2013 des directives claires ont été données par la Direction générale de l'enseignement scolaire (DGESCO) pour faire évoluer la façon de noter les élèves sous l'impulsion de Vincent Peillon qui appelait de ses vœux « une transformation réelle des pratiques en classe » avec la conscience que l'évaluation est une activité centrale avec des répercussions importantes sur la motivation et la progression des élèves. Enfin, une conférence nationale sur l'évaluation des élèves a vu le jour en décembre 2014 ; le rapport du jury citoyen constitué à cette occasion recommandait davantage de place à l'évaluation formative et une diminution de l'importance de l'évaluation sommative. Pourtant, si les notes chiffrées ont fortement battu en brèche au cycle 2 (CP-CE1-CE2) au profit des points de couleur (vert-jaune-rouge), des lettres (A-B-C-D), et des appréciations (TB-B-AB-à revoir)[(2)], il ne s'agit aucunement d'un changement d'approche : les enseignants continuent d'évaluer le niveau de performance des élèves. Le terme « sans note » est parfois utilisé avec abus pour désigner le recours à l'évaluation formative. En réalité ce n'est pas l'absence ou la présence de notes qui distinguent les deux formes d'évaluation, mais l'objectif recherché (attester d'un niveau ou encourager la progression), et par voie de conséquence la façon dont l'enseignant dialogue avec l'élève au cours du processus d'évaluation, l'encourage, et élabore avec lui des stratégies pour s'améliorer.

En France, depuis la fin des années 2000, un nombre croissant de collèges ont abandonné les notes chiffrées pour adopter l'« évaluation par compétence ». Ces collèges sont communément désignés par les services de l'Éducation nationale comme « collèges sans note ». Bien que très variables d'un collège à l'autre, les nouvelles pratiques d'évaluation par compétence ont pour point commun l'abandon de la note chiffrée et de toute autre forme d'évaluation synthétique d'un ensemble d'acquisitions diverses. À la place, les enseignants évaluent une par une un ensemble de compétences spécifiques, en utilisant généralement pour chacune de ces compétences trois catégories : « acquis », « en cours d'acquisition » et « non acquis ». Ainsi, au lieu d'obtenir une note synthétique masquant les points forts et les points faibles, les élèves reçoivent un bulletin listant l'ensemble des compétences et leur niveau d'acquisition pour chacune d'elles. Par contraste avec la notation classique, l'évaluation par compétence permet à l'élève d'identifier ce qu'il sait déjà faire et ce qu'il ne sait pas encore faire. De plus, l'évaluation par compétence limite la compétition entre élèves là où l'évaluation classique chiffrée classe les élèves entre eux et stimule la compétition. Dans certains cas, ce changement de système d'évaluation s'accompagne d'un changement plus radical qui s'apparente clairement à l'évaluation formative. Le projet des classes coopératives du collège Lucien Vadez à Calais, mis en place en 2014, en est un exemple frappant : ces classes proposent une pédagogie centrée sur l'évaluation formative, l'entraide au sein de la classe, et sur des temps de communication privilégiés. Les changements observés dans ces classes coopératives par rapport aux classes classiques sont d'ailleurs spectaculaire[(3)]. Cependant, si on en reste à l'évaluation par compétence elle-même, nous devons reconnaître qu'elle ne franchit pas encore le pas d'une réelle évaluation formative : premièrement elle ne dévie pas fondamentalement de l'objectif d'attester d'un niveau d'acquisition, fût-il détaillé compétence par compétence au lieu d'être synthétique ; deuxièmement, elle ne s'accompagne pas nécessairement d'un échange personnalisé entre l'élève et

(2) Le rapport de l'IGEN sur l'évaluation des élèves paru en 2013 fait état de 17% des écoles utilisant la note chiffrée au cycle 2, suivi d'un saut à 71 % en CM1-CM2.

(3) Alors que les classes coopératives affichaient les mêmes performances académiques et des origines sociales similaires en 6e, les classes coopératives affichaient en moyenne 10 demi-journées d'absence de moins que les classes classiques, les retards étaient deux fois plus bas, les punitions et les exclusions étaient aussi plus rares, et les résultats académiques en forte hausse avec 100 % de réussite au brevet contre 83,3 % dans les classes classiques, *cf*. Caron (2018).

l'enseignant sur ses progrès, ses points forts, ses points faibles, les pistes de progression. Or, ce dialogue est la condition nécessaire pour que l'évaluation puisse renforcer la confiance en soi et motiver l'élève à dépasser ses difficultés. L'évaluation par compétence constitue donc une forme d'évaluation à mi-chemin entre l'évaluation sommative classique et l'évaluation formative.

Le but de cette étude est d'estimer l'impact de l'évaluation par compétence sur les apprentissages scolaires des collégiens bénéficiant de cette nouvelle forme d'évaluation en France. Concrètement, il s'agit de savoir si le fait de passer d'une évaluation classique à une évaluation par compétence a un impact sur le niveau d'acquisition à la fin de la 3e, que nous mesurons par les résultats à l'examen du Diplôme national du brevet (DNB). Nous utilisons une méthode par appariement combinée à une méthode de différence-en-différence afin de mesurer les effets de l'évaluation par compétence sur le niveau d'acquisition des élèves. Deux traitements sont étudiés : le premier consiste en l'évaluation par compétence en 6e seulement (« traitement de faible intensité »), tandis que le second consiste en l'évaluation par compétence durant tout le collège(4) (« traitement de forte intensité »).

2. Revue de littérature

Selon Brookhart (2004), l'évaluation dans la salle de classe répond à une logique formative et sommative qu'il convient de séparer. Au cours d'une leçon, l'évaluation et la notation doivent maximiser leurs valeurs informationnelles pour l'élève. Une évaluation formative encourage l'élève dans son apprentissage et est toujours accompagnée d'un retour de l'enseignant précis et individuel sur les acquis, les faiblesses et les points à travailler de chaque élève. La note chiffrée n'est pas obligatoirement supprimée, mais elle est accompagnée de commentaires constructifs. Une évaluation sommative, avec une note classique, peut être réalisée à la fin de la période consacrée à une notion pour témoigner des compétences acquises par l'élève de la façon la plus claire possible.

Butler et Nissan (1986) montrent l'effet positif de l'évaluation formative au cours d'une expérience randomisée sur 261 élèves âgés de 12 à 13 ans. Les élèves recevant des retours précis et détaillés affichaient des résultats plus élevés à des tests quantitatifs sur des épreuves faisant appel à la créativité que ceux recevant des notes et ceux ne recevant aucun retour. L'évaluation formative avait aussi réduit l'anxiété, et augmenté l'intérêt des élèves pour l'exercice. Hudesman *et al*. (2013) évaluent l'impact de l'évaluation formative et de cours méta-cognitifs, c'est-à-dire « apprendre à apprendre », sur 15 000 étudiants en mathématiques d'une université américaine. Après avoir reçu des conseils personnalisés pour améliorer leurs résultats et leurs méthodes d'apprentissage, les étudiants traités obtenaient des scores significativement plus élevés que des groupes contrôles sélectionnés aléatoirement ou identiques en performances pré-traitement. Hattie (2008) réalise une grande méta-analyse de la littérature scientifique dédiée aux facteurs de la réussite scolaire. Il en déduit que la présence de retours pertinents de l'enseignant sur les travaux de l'élève constitue le facteur le plus puissant de réussite scolaire, avec une amélioration moyenne de 73 % d'un écart-type. En comparaison, des interventions plus populaires, mais aussi plus coûteuses, telles que la réduction de la taille des classes ou les devoirs, mènent à des améliorations de 21 et 29 % d'un écart-type respectivement.

L'évaluation formative semble aussi bénéficier aux élèves les plus performants comme le suggèrent Crouzevialle *et al*. (2015). Parmi 110 étudiants à l'université, les individus à forte capacité de mémoire de travail obtenaient de moins bons résultats dans une situation marquée par la performance et la comparaison sociale. Un tel environnement, typiquement induit par les notes chiffrées, provoque de l'anxiété chez les bons élèves pour maintenir leur statut. Dans leur méta-analyse, Fuchs et Fuchs (1986) montrent que l'évaluation formative avait aussi eu un impact positif et significatif sur les performances scolaires d'enfants handicapés.

(4) Comme expliqué plus loin, la réapparition des notes classiques est cependant fréquente en 3e du fait de la prise en compte du contrôle continu dans le score global du diplôme national du Brevet jusqu'en 2016.

Les mécanismes cognitifs et psychologiques liant l'évaluation formative à de meilleures performances sont nombreux. Hayek *et al*. (2015) et (2017) montrent que la présence de notes chiffrées réduit la capacité d'élèves de l'université à travailler en équipe ainsi que les comportements coopératifs, menant à de plus faibles performances lors de travaux de groupes. Crouzevialle et Butera (2013) prouvent que chez des étudiants à l'université, la présence de notes est vécue comme une menace. L'anxiété et le désir de surpasser les autres induisent des pensées parasites qui accaparent les capacités cognitives verbales des étudiants entrainant des performances plus faibles sur des tâches complexes.

La nature de la motivation des élèves semble aussi jouer un rôle déterminant. Pulfrey *et al*. (2013) démontrent que pour 89 collégiens, la présence de note réduit l'intérêt pour le contenu de l'exercice et la motivation intrinsèque, pourtant bénéfique à l'apprentissage sur le long terme. À l'inverse, les élèves n'étaient plus motivés que par la note (motivation extrinsèque). Selon Butler (1988), lors d'un exercice noté, l'ego devient la source principale de motivation. Les bons élèves gardent un haut niveau d'engagement car ils sont en mesure de réaffirmer leurs capacités, à l'inverse l'engagement des moins bons élèves s'effondre à l'introduction d'une note chiffrée. L'effet négatif des notes sur l'engagement est observé pour les deux groupes lors d'activités créatives.

L'auto-handicap est aussi un mécanisme possible. Jones et Berglas (1978) le définissent comme une stratégie de protection de l'estime de soi lorsque la valeur personnelle ou les compétences d'un individu doivent être démontrées. Pour éviter un possible échec, l'individu s'inflige des obstacles et des difficultés externalisant ainsi la raison de l'échec, le rendant encore plus probable. Ainsi, Zuckerman *et al.* (1998) montrent que les étudiants recourant le plus aux stratégies d'auto-handicap obtenaient de moins bons résultats à résultats académiques antérieurs égaux.

3. L'évaluation par compétence

L'évaluation par compétence consiste en l'abandon de la note chiffrée et de toute autre forme d'évaluation synthétique agrégeant les acquisitions de notions diverses. À la place, les enseignants évaluent une par une un ensemble de compétences spécifiques, qui ne sont pas seulement des connaissances mais aussi des capacités d'apprentissage (raisonnements, mobilisation des ressources), et des attitudes (travail collaboratif, maîtrise de soi, persévérance). Pour chacune de ces compétences, l'évaluation consiste généralement en trois niveaux d'acquisition : « acquis », « en cours d'acquisition », ou « non acquis » (mais ces niveaux peuvent bien entendu varier dans leur nombre et leur dénomination). Ainsi, au lieu d'obtenir une note synthétique masquant les points forts et les points faibles et favorisant la compétition, les élèves reçoivent un bulletin listant l'ensemble des compétences et leur niveau d'acquisition pour chacune d'elles. Les compétences à acquérir sont lisibles pour tous, élèves, parents et enseignants, et les élèves ne peuvent pas directement se comparer les uns aux autres.

Il est important de noter que lorsqu'elle est pratiquée, l'évaluation par compétence est lancée à l'initiative des équipes d'enseignants ou de l'administration des collèges, dans une démarche volontaire et décentralisée. Parce qu'elle ne suit pas un cadre unifié et commun, l'évaluation par compétence peut prendre des formes variées dans les différents collèges qui l'ont adoptée. L'évaluation par compétence peut être menée pour une seule matière ou pour toutes les matières, et pour une classe seulement, toutes les classes d'un seul niveau, ou toutes les classes dans tous les niveaux. L'évaluation par compétence peut être directement appliquée à plusieurs niveaux mais la plupart du temps, l'application commence pour une seule cohorte et est étendue à d'autres niveaux au fur et à mesure que la cohorte initiale progresse.

Dans cette étude, nous considérons uniquement l'évaluation par compétence appliquée à l'ensemble d'un niveau et à l'ensemble des matières. De plus, deux intensités de traitement sont étudiées :

- léger : les élèves ont été exposés à l'évaluation par compétence uniquement en classe de 6^{e} ;

- intense : les élèves ont été exposés à l'évaluation par compétence trois ou quatre années lors de leur collège, soit en 6e, 5e, 4e et éventuellement en 3e[(5)].

En se basant sur une liste auto-déclarative de collèges recueillie par le ministère de l'Éducation nationale auprès des rectorats, une liste de 593 collèges ayant utilisé ou utilisant l'évaluation par compétence est établie, issus de 20 académies. Nous avons enrichi les informations transmises par les rectorats en les recoupant avec les informations contenues par les sites internet des collèges pour limiter les erreurs de mesure. Nous avons ensuite procédé à certaines restrictions sur cet ensemble d'établissements afin de rendre possible l'étude d'impact de l'évaluation par compétence et son interprétation. Nous avons éliminé :

- 97 établissements qui ont appliqué l'évaluation par compétence suivant un schéma atypique différent des deux traitements décrits ci-dessus[(6)] ;
- 27 établissements qui n'appliquaient pas l'évaluation par compétence sur l'ensemble des classes d'un même niveau ;
- 29 établissements qui n'appliquaient pas cette pratique à l'ensemble des matières et/ou continuaient à utiliser l'évaluation sommative en parallèle ;
- 328 établissements dans lesquels l'utilisation de l'évaluation par compétence a commencé récemment et les élèves exposés à l'évaluation par compétence n'avaient pas encore passé le DNB en juin 2017 (dernières observations du DNB disponibles au moment de notre étude). Le fait que plus de la moitié des collèges soient passés à l'évaluation par compétence récemment suggère que cette tendance se développe et que des études d'impact ultérieures bénéficieront de plus nombreuses observations.

Finalement, 89 établissements sont gardés dans l'échantillon : 58 d'entre eux correspondent au schéma « traitement léger » et 31 au schéma « traitement intense ».

4. Stratégies d'identification

Appliquée de manière décentralisée et sous l'impulsion des collèges, l'évaluation par compétence n'a pas été pensée dans la perspective d'une expérimentation nationale intégrant un référentiel commun de pratiques et un protocole d'évaluation d'impact par la méthode préconisée de l'assignation aléatoire. Par conséquent, l'évaluation par compétence prend des formes naturellement différentes d'un collège à l'autre, et cette étude rend compte de l'impact moyen sans qu'il soit possible de déterminer l'impact local de telle ou telle application. Notre étude apporte toutefois une information pertinente : elle mesure l'impact qu'aurait l'évaluation par compétence sur les apprentissages des élèves quand celle-ci se développe de manière décentralisée et volontaire.

(5) La réforme du DNB, effective à partir de la session 2017, remplace le contrôle continu classique fondé sur l'ensemble des notes sur 20 obtenues en 3e par des points attribués en fin d'années scolaires selon les « degrés de maîtrise des compétences du socle ». L'attribution de ces points peut s'appuyer en partie sur les notes obtenues dans l'année mais elle peut aussi s'appuyer sur l'évaluation par compétence, donc le nouveau DNB ne nécessite plus un contrôle continu avec des notes chiffrées. Avant l'année scolaire 2016-2017, les collèges étaient en revanche limités dans la mise en place de l'évaluation par compétence en 3e car le contrôle continu était nécessaire au calcul du score final au DNB. Les professeurs ont parfois contourné le problème et ont attribué des notes chiffrées indicatives basées sur les compétences acquises par les élèves, mais on ne peut exclure que les notes aient dû parfois être réintroduites en 3e. À partir de 2017, au contraire, l'évaluation par compétence a pu se poursuivre pleinement pendant l'année de 3e.

(6) Tous les cas de figure étant possibles, nous nous sommes limités aux deux traitements décrits ci-dessus et avons éliminé les situations atypiques dans lesquelles les élèves ont été exposés à l'évaluation par compétence une ou plusieurs années pendant le collège de façon incohérente. Par exemple, les expositions tardives ne commençant pas en 6e ont été exclues. De même, les expositions discontinues avec une alternance d'années avec évaluation par compétence et d'années avec évaluation classique ont aussi été exclues. Ces exclusions sont motivées par le fait que ces situations ne correspondent pas à des choix pédagogiques délibérés, mais aux aléas de calendrier dans la mise en œuvre ou le retrait de l'évaluation par compétence.

De plus, une attention particulière a été portée aux biais de sélection qui peuvent différencier les élèves issus des collèges sans notes des élèves issus des collèges classiques. En effet, au-delà de l'hétérogénéité des mesures, la participation volontaire des collèges à l'évaluation par compétence est de nature à générer des biais de sélection. Cette étude combine deux stratégies d'identification pour contrôler ces biais, que nous détaillons par la suite. Mais avant de détailler ces stratégies, une première étape consiste à mesurer les caractéristiques des collèges qui font le choix d'utiliser l'évaluation par compétence, telles que le score moyen au brevet, la composition sociale des collèges ou encore leur taille.

4.1. Caractéristiques observables des collèges sans note

À l'aide de régressions logistiques au niveau des collèges, on calcule la probabilité d'appartenir à l'un des deux groupes de traitement en fonction de ses caractéristiques observables dans les bases de données administratives de la Direction de l'évaluation de la performance et de la prospective (DEPP)[7] sur la période allant de 2007 à l'année précédant la mise en place de l'évaluation par compétence. Le tableau 1 présente les résultats de ces régressions.

Premièrement, on observe que la composition sociale des collèges n'influence pas significativement la probabilité d'utilisation de l'évaluation par compétence. En revanche, le niveau scolaire des élèves, tel que mesuré par les résultats au brevet, semble avoir un fort pouvoir prédictif. En effet, une plus grande proportion de mention TB entre 2007 et l'année précédant le traitement est associée avec une moins grande probabilité d'être traité, et ce pour les deux traitements léger et intense. Inversement, des proportions élevées de mention B et de candidats non-admis augmentent la probabilité d'appartenir aux groupes de traitement de façon significative. Une augmentation d'un percentile dans la distribution nationale de la proportion de non-admis est associée avec une augmentation de la probabilité de participer aux traitements léger et intense de 4,6 et 4,1 points de pourcentage respectivement. Globalement, un niveau scolaire plus faible au départ augmente donc la propension de l'établissement à passer à l'évaluation par compétence.

Les régressions montrent aussi un effet académie important, ce qui suggère que la force d'impulsion du rectorat joue un rôle important dans le déploiement de ce nouveau mode d'évaluation. Les collèges des académies de Poitiers ou encore de Dijon montrent une plus forte probabilité d'être traités, traduisant potentiellement des environnements plus innovants au sein de ces académies.

(7) Convention DEPP de mise à disposition des données n° 2018-08.

Tableau 1. Probabilités de traitement en fonction des caractéristiques observables des collèges entre 2007 et l'année précédant le traitement

	Traitement léger		Traitement intense	
Pourcentage d'élèves de classe sociale				
Défavorisée	– 0,314153	(– 0,52)	– 0,9161582	(– 1,14)
Intermédiaire	– 0,3001342	(– 0,57)	– 0,9581709	(– 1,40)
Favorisée	– 0,7906693	(– 1,46)	– 0,3291814	(– 0,55)
Proportion dans la distribution nationale				
• mention TB	– 0,0324066	(– 3,20)	– 0,048503	(– 3,48)
• mention B	0,0545909	(4,20)	0,0335073	(2,23)
• mention AB	– 0,0049937	(– 0,46)	0,0151573	(1,07)
• admis sans mention	0,0009021	(0,10)	– 0,0031319	(– 0,26)
• non-admis	0,0455948	(2,98)	0,0414805	(2,27)
Taille des effectifs (distrib. nationale)	0,006286	(1,12)	0,0005158	(0,07)
Académie de Bordeaux	0		1,824286	(1,51)
Académie de Caen	0		3,070148	(2,60)
Académie de Clermont-Ferrand	1,4079	(1,77)	1,716014	(1,17)
Académie de Dijon	1,926279	(2,78)	1,862818	(1,26)
Académie de Grenoble	0,7526436	(1,07)	2,216607	(1,92)
Académie de Limoges	0		3,76444	(3,09)
Académie de Lyon	0		0,8269478	(0,57)
Académie de Montpellier	0,0361393	(0,04)	0	
Académie de Nancy-Metz	1,130367	(1,50)	0	
Académie de Nantes	0,7374807	(0,98)	0	
Académie de Nice	– 0,3215787	(– 0,29)	1,113156	(0,77)
Académie de Orléans-Tours	– 0,6125954	(– 0,54)	0	
Académie de Poitiers	1,762578	(2,52)	3,443506	(2,98)
Académie de Rouen	0,6739094	(0,78)	0	
Académie de Strasbourg	2,900667	(4,97)	2,833025	(2,22)
Académie de Toulouse	– 0,2852592	(– 0,25)	1,684657	(1,31)

En outre, des facteurs déterminants tels que l'attitude des équipes pédagogiques, leur niveau de formation, leur dynamisme, ou encore leur âge et leur expérience, peuvent être liées au choix du collège d'opter pour l'évaluation par compétence mais ne sont pas disponibles dans nos données. Typiquement, les collèges dont les équipes pédagogiques (enseignants et administration) sont les plus dynamiques, motivées et innovantes, sont plus susceptibles de prendre part au traitement que les collèges au personnel plus traditionnel et sans volonté de changement. Les résultats observés au DNB peuvent alors refléter cet effet « équipe enseignante innovante » qui peut favoriser les apprentissages indépendamment du mode d'évaluation dans le groupe de traitement (collèges pratiquant l'évaluation par compétence) par rapport au groupe de contrôle (collèges pratiquant l'évaluation classique).

Pour pallier ces différents biais de sélection, nous avons choisi de combiner d'une part un appariement statistique de chaque collège traité avec un collège contrôle « jumeau » pour former deux groupes très similaires avant la mise en place de l'évaluation par compétence, et d'autre part une stratégie de différence-en-différences permettant d'isoler l'impact de l'évaluation par compétence de l'écart résiduel entre les groupes traitement et contrôle et de l'évolution tendancielle naturelle des résultats au DNB au cours du temps. Nous utilisons deux groupes de contrôle différents que nous présentons ci-après successivement ; le deuxième groupe de contrôle, notamment, permet de mieux traiter la sélection sur inobservables (dynamisme de l'équipe pédagogique).

4.2. Stratégie 1 : groupe de contrôle externe

La première étape de l'étude consiste en la construction d'un groupe de contrôle parmi l'ensemble des collèges n'ayant jamais utilisé l'évaluation par compétence, appelé « groupe de contrôle externe ». Concrètement, pour chaque établissement traité, on se place dans les années antérieures (à partir de 2007) à la mise en place de l'évaluation par compétence et on calcule des variables cibles incluant l'origine sociale des élèves (proportions d'élèves favorisés et défavorisés), les résultats au brevet (taux de réussites et taux d'obtention des différentes mentions) ainsi que la situation géographique de l'établissement (académie). Ces variables cibles sont également calculées pour tous les autres établissements de France n'ayant jamais eu recours à l'évaluation par compétence. Des paires de collèges « jumeaux » sont établies suivant deux critères :

- ils doivent appartenir à la même académie ;
- ils minimisent la somme des distances entre les variables cibles : proportion d'élèves favorisés, proportion d'élèves défavorisés, proportion de non-admis, proportion d'admis sans mention, proportion de mention AB, proportion de mention B, et proportion de mention TB.

Deux collèges jumeaux présentent ainsi des caractéristiques très similaires sur l'ensemble de la période qui précède l'introduction de l'évaluation par compétence. Les graphiques ci-dessous montrent l'évolution des proportions d'élèves non-admis (NA), et ayant obtenu une mention très bien (TB) pour les collèges traités et leurs jumeaux contrôles pour les sessions du brevet précédant le début du traitement. On voit que les résultats des deux groupes non seulement suivent une trajectoire très similaire avant le traitement, mais sont aussi très proches en niveau. Les résultats des régressions « placebo »[(8)] montrent qu'il n'y avait pas de différences significatives avant la mise en place de l'évaluation par compétence entre l'évolution au cours du temps des résultats au DNB dans les collèges traitement et cette même évolution dans les collèges contrôle. Ainsi on peut faire l'hypothèse que le groupe contrôle constitue un bon contrefactuel et que l'évolution de ses résultats au brevet après traitement représenterait l'évolution des résultats du groupe traité si le traitement n'avait pas lieu. Par conséquent, les différences de résultats au brevet observées entre le groupe contrôle et le groupe traité après le début du traitement pourront être imputées entièrement au programme d'évaluation par compétences : c'est l'impact de ce mode d'évaluation par rapport à la notation classique.

Avant de comparer les résultats des collèges au sein de chaque paire, les élèves ayant changé d'établissement entre la 6^{e} et la 3^{e} sont retirés de la base de données. Cette procédure, qui n'impacte que marginalement les scores des collèges, permet de s'assurer qu'aucun élève du collège de contrôle n'a été concerné par l'évaluation par compétence et inversement, que les élèves du collège traité ont pleinement bénéficié du traitement étudié.

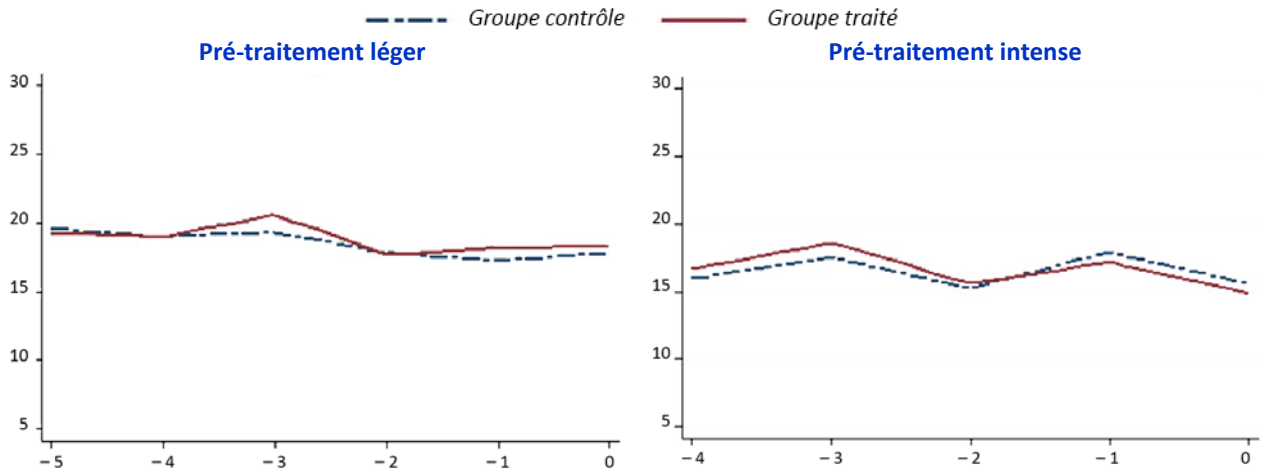
L'expression « différence en différence » traduit la formule mathématique correspondant à cet exercice. Au sein de la paire de collège $i - j$, où le collège i applique l'évaluation par compétence à partir du moment t et que le collège j ne le pratique à aucun moment, l'effet du traitement peut être retrouvé en faisant la différence des écarts (voir équation 1).

(1) $$Effet = \left(Résultat_{it} - Résultat_{jt}\right) - \left(Résultat_{it-1} - Résultat_{jt-1}\right)$$

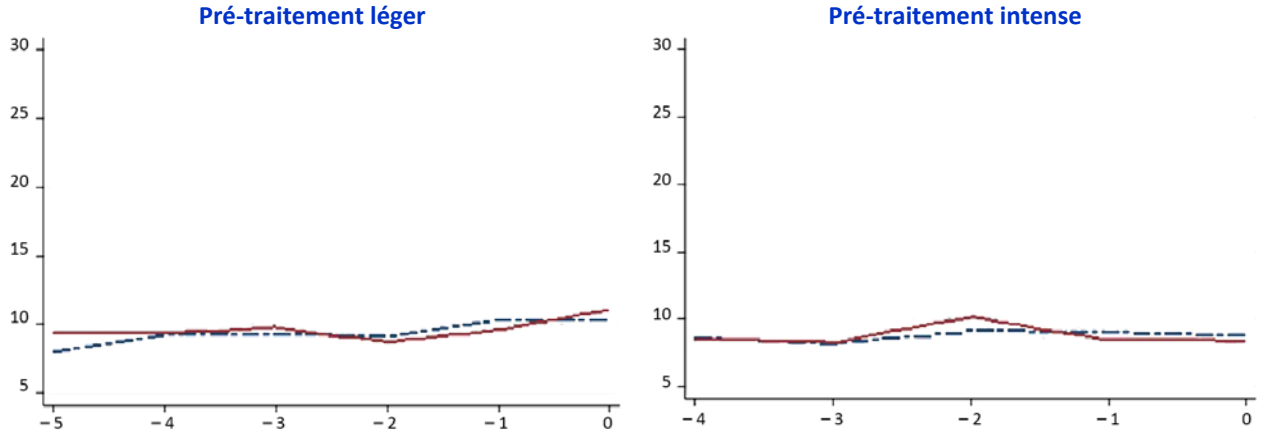
(8) Les régressions placebo sont les mêmes que celles présentées dans la partie 5.2 en prenant pour date de mise en place du traitement des dates fictives antérieures de 1 ou 2 ans à la date réelle de mise en place du traitement. L'hypothèse d'évolution identique avant le traitement est validée si ces régressions donnent des coefficients non significatifs.

1. Stratégie 1 : pré-traitements léger et intense

a. Proportion de mentions NA (non admis) selon le nombre d'années avant traitement (en %)



b. Proportion de mentions TB (très bien) selon le nombre d'années avant traitement (en %)



Source : Auteurs d'après bases 'Apprenants et Océan DNB' (2007-2016), DEPP, ministère de l'Éducation nationale.

4.3. Stratégie 2 : groupe de contrôle interne

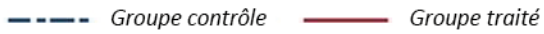
À la différence de la première stratégie, les collèges utilisés comme collèges de contrôle dans la formation des paires sont tirés de la base de données des 598 collèges pratiquant l'évaluation par compétence mais choisis de sorte que l'application de cette méthode pédagogique ne soit pas encore effective sur la cohorte étudiée. Ce groupe de contrôle est ainsi qualifié « d'interne » car il est composé de collèges tardivement traités. Parmi ces collèges « tardifs », on assigne à chaque collège traité un jumeau statistique en termes de notes et de composition sociale d'une façon similaire à la stratégie 1. La seule différence avec les critères de la stratégie 1 est que l'identification du jumeau statistique n'est pas limitée à l'académie des collèges traités. En effet, l'application de ce critère aurait réduit trop fortement l'ensemble des établissements contrôles potentiels et n'auraient pas permis une bonne qualité d'appariement.

Si l'éventail de choix pour trouver un collège très similaire au collège étudié se réduit du fait de la plus petite taille de cet échantillon, cette stratégie permet de mieux neutraliser l'effet « équipe pédagogique innovante » inobservable par nature. En effet, puisqu'au sein d'une paire les deux collèges ont décidé (à des moments différents) d'appliquer l'évaluation par compétence, le degré de motivation et d'innovation des équipes est probablement plus proche entre les deux groupes.

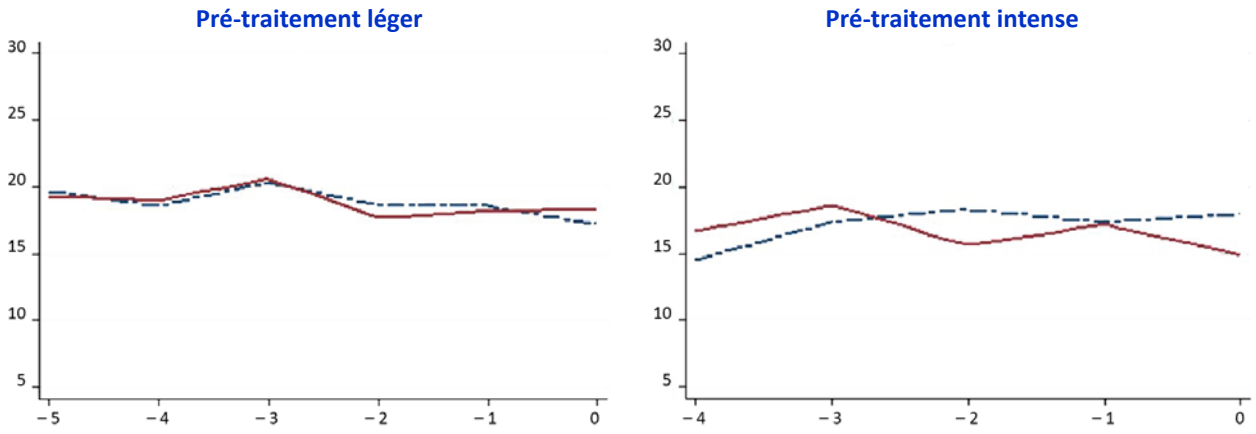
Les graphiques ci-dessous montrent que les paires de collèges sont très similaires avant le début du traitement pour le traitement léger. Les proportions de non-admis (NA) et de mention très bien (TB) pour les collèges traités et contrôle suivent des trajectoires très proches, le groupe contrôle constitue donc un contrefactuel efficace. Les régressions « placebo » montrent à nouveau que les évolutions des deux groupes de collèges sont bien identiques avant la mise en place du traitement. La comparaison de leurs résultats au brevet après le début du traitement permettra donc d'identifier l'effet pur de l'évaluation par compétence.

En revanche pour le traitement intense, les tendances pré-traitement ne sont pas parallèles et les régressions « placebo » obligent à rejeter l'hypothèse selon laquelle les évolutions étaient identiques dans les deux groupes avant la mise en place de l'évaluation par compétence. Aussi, nous ne pourrons pas exploiter le groupe de contrôle interne pour évaluer l'impact du traitement intense. Nous pouvons exploiter le groupe de contrôle interne seulement pour mesurer l'impact du traitement léger.

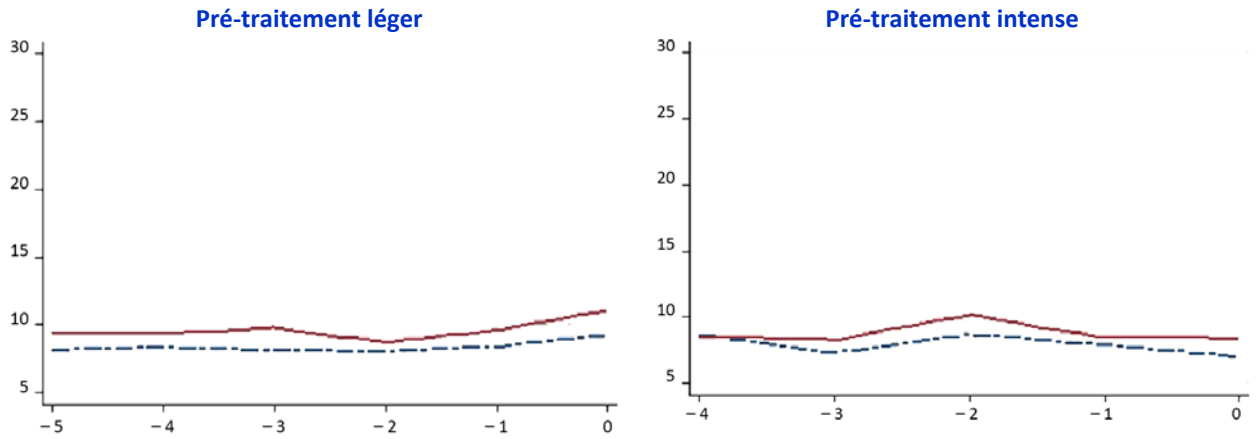
2. Stratégie 2 : pré-traitements léger et intense



a. Proportion de mentions NA (non admis) selon nombre d'années avant traitement (en %)



b. Proportion de mentions TB (très bien) selon nombre d'années après traitement (en %)



Source : Auteurs d'après bases 'Apprenants et Océan DNB' (2007-2016), DEPP, ministère de l'Éducation nationale.

5. Résultats

5.1. Taux d'admis et de non-admis par groupe avant et après le traitement

Le tableau 2 présente les différences entre les groupes contrôles et témoins avant et après traitement. L'avantage de cette présentation est de donner une image claire et transparente des écarts entre les deux groupes et des écarts avant et après le traitement, permettant ainsi de visualiser simplement l'impact du traitement. Les termes d'erreur sont obtenus par la méthode du *bootstrap*. Pour simplifier la présentation des résultats, nous reportons dans ce tableau les effets pour seulement deux variables d'intérêt : ne pas être admis au DNB et être admis au DNB.

Pour la stratégie 1, le taux d'échec au brevet a baissé plus fortement dans les collèges concernés par le traitement léger que dans les collèges « jumeaux » du groupe de contrôle. Au contraire, la proportion d'échec au brevet a augmenté plus fortement pour les élèves traités par 3 ou 4 années d'évaluation par compétence que dans les collèges contrôle. Ces effets ne sont toutefois pas statistiquement significatifs, c'est-à-dire qu'il n'est pas possible de différencier ces petites variations d'un simple bruit statistique.

En utilisant le groupe de contrôle interne (stratégie 2), les évolutions entre les collèges ayant mis en place l'évaluation par compétence en 6e et les collèges contrôle sont encore plus proches. Ces résultats indiquent une absence d'impact de l'évaluation par compétence en 6e sur le taux d'admission et de non-admission au DNB, ce qui suggère que les apprentissages académiques des élèves n'ont été affectés ni positivement ni négativement par l'évaluation par compétence.

Tableau 2. Différences entre les groupes contrôles et témoins avant et après traitement

	Pré-traitement		Post-traitement		Différence post-pré	
Stratégie 1. Traitement léger						
• Non admis						
– Traitement léger	0,169		0,116		– 0,0521(***)	(– 7,26)
– Contrôle	0,167		0,126		– 0,0401(***)	(-6,02)
– Différence T – C	0,00200	(0,50)	– 0,0100	(– 1,11)	– 0,0120	(-1,38)
• Admis						
– Traitement léger	0,795		0,852		0,0572(***)	(6,85)
– Contrôle	0,799		0,843		0,0433(***)	(5,84)
– Différence T – C	– 0,00364	(– 0,80)	0,00925	(0,99)	0,0129	(1,34)
Stratégie 1. Traitement intense						
• Non admis						
– Traitement intense	0,148		0,131		– 0,0168	(– 1,65)
– Contrôle	0,143		0,116		– 0,0276(***)	(– 2,50)
– Différence T – C	0,00431	(0,77)	0,00431	(0,77)	0,0107	(0,97)
• Admis						
– Traitement intense	0,813		0,840		0,0270(***)	(2,66)
– Contrôle	0,821		0,849		0,0286(***)	(3,20)
– Différence T – C	– 0,00796	(– 1,33)	– 0,00959	(– 0,92)	– 0,00163	(– 0,14)
Stratégie 2. Traitement léger						
• Non admis						
– Traitement 1	0,169		0,116		– 0,0521(***)	(– 7,26)
– Contrôle	0,171		0,117		– 0,0541(***)	(– 6,87)
– Différence T – C	– 0,00247	(– 0,57)	– 0,000508	(– 0,05)	0,00196	(0,20)
• Admis						
– T	0,795		0,852		0,0572(***)	(6,85)
– C	0,789		0,851		0,0620(***)	(7,51)
– T – C	0,00633	(1,33)	0,00153	(0,13)	– 0,00480	(– 0,44)

Lecture : Les estimations des erreurs-type sont obtenues en utilisant la méthode du *bootstrap*. La statistique t de Student est présentée entre parenthèses.

Notes : (***) Significatif au seuil de 1 % ; (**) Significatif au seuil de 5 % ; (*) Significatif au seuil de 10 %.

Source : Auteurs d'après bases 'Apprenants' et 'Océan DNB' (2007-2016), DEPP, ministère de l'Éducation nationale.

5.2. Spécification de différence-en-différences avec effets fixes

Pour gagner en précision, nous évaluons maintenant l'impact des deux traitements en utilisant une spécification permettant d'exploiter plus finement les variations intra-collège avant et après la mise en place de l'évaluation par compétence. L'estimation est réalisée selon la méthode des moindres carrés ordinaires en utilisant une spécification qui suit le modèle de différence-en-différences avec effets fixes établissements *i* et années *j* :

$$(2)\ Résultat\ DNB = \alpha + \gamma, Traitement, Post + \sum_i \beta_i, Etablissement_i + \sum_j \beta_j, Années_j + \varepsilon_{ij}$$

La variable « résultat DNB » est une *dummy* indiquant si l'élève a été non admis / admis sans mention / admis avec mention AB / admis avec mention B / admis avec mention TB. Nous utilisons également une autre variable dépendante qui complète le résultat au DNB : une variable qui indique si l'élève a redoublé au cours de sa scolarité au collège. Le coefficient du terme d'interaction γ mesure la différence entre l'écart entre les élèves d'un même collège avant et après traitement dans les collèges traités et ce même écart dans les collèges contrôle. Ce paramètre donne donc l'impact de l'évaluation par compétence de manière très fine. Les termes d'erreur sont clusterisés au niveau du collège pour tenir compte des facteurs communs partagés par les élèves issus du même établissement, qui invalident l'hypothèse d'indépendance.

Cette régression est appliquée aux deux traitements en utilisant la stratégie 1 (groupe de contrôle externe) (les résultats utilisant la stratégie 2, soit le groupe de contrôle interne, pour le traitement léger donne les mêmes résultats). Nous présentons les résultats pour l'ensemble des élèves, ainsi que par sous-groupes selon la catégorie socio-professionnelle des parents : la catégorie favorisée comprend les cadres, les professeurs, les ingénieurs et les professions libérales, la catégorie intermédiaire correspond aux techniciens, artisans, commerçants et agriculteurs, enfin la catégorie défavorisée rassemble les ouvriers et les chômeurs. En effet, il est important de savoir non seulement l'effet moyen de l'évaluation par compétence sur les apprentissages, mais aussi les effets spécifiques selon l'origine sociale des élèves, ceux-ci pouvant être en principe différents entre eux. Les résultats sont présentés dans le tableau 3.

Aucun des coefficients de différence-en-différences n'est significativement différent de 0, et ce quelle que soit l'intensité du traitement et quelques soit l'origine sociale des élèves. Bien que les élèves traités aient reçu une à plusieurs années d'évaluation par compétence, leurs résultats au brevet ne sont statistiquement pas différents de ceux de leurs pairs scolarisés dans un environnement classique avec évaluation chiffrée classique. Les effets des traitements léger et intense sur les performances des élèves traités sont donc absents ou trop faibles pour être différenciés de 0.

De légères tendances sont toutefois observées. Le traitement léger semble être associé à une baisse très marginale de l'échec au brevet de l'ordre de 1 point de pourcentage (soit 6%) sur l'ensemble de l'échantillon. Cet effet provient majoritairement des élèves issus de classes sociales défavorisées pour qui l'échec a baissé de 2 points de pourcentage comparés aux élèves défavorisés non traités, cette baisse n'étant toutefois pas assez précisément estimée (t = 1,42). Les élèves défavorisés recevant le traitement léger semblent également redoubler 10 % moins souvent durant le collège que leurs pairs non traités (baisse de – 1,27 point de pourcentage marginalement significative, t = 1,52). La pratique du redoublement pourrait donc avoir légèrement diminuée dans les collèges pratiquant l'évaluation par compétence, ce qui semblerait cohérent avec une approche moins conservatrice de la notation et des décisions de passage. Il faut cependant noter que cette légère baisse du taux de redoublement entraîne potentiellement une sous-estimation des effets de l'évaluation par compétence sur les résultats au DNB. En effet, les élèves les plus fragiles se présentent au DNB un an plus tôt dans les collèges traitement par rapport aux collèges contrôle. Comme nous n'observons que la première cohorte exposée au traitement dans nos données, les élèves les plus fragiles sont plus nombreux dans le groupe traitement que dans le groupe contrôle, ce qui entraîne un biais à la baisse des écarts entre les deux groupes dans la période

post-traitement[9]. Concernant le traitement intense, il paraît entrainer une très faible baisse des mentions élevées menant à une redistribution sur les mentions intermédiaires principalement pour les catégories socio-professionnelles favorisées et intermédiaires. La validité de ces observations est cependant limitée en raison de leur absence de significativité.

Tableau 3. Différences entre les groupes contrôles et témoins avant et après traitement (différence en différences)

	Non-admis	Admis sans mention	Mention assez bien	Mention bien	Mention très bien	Redoublement au collège
Stratégie 1. Traitement léger						
• Population totale	– 0,0100	0,00643	0,0119	– 0,00768	– 0,000577	– 0,00685
(N = 111 005)	(– 1,10)	(0,78)	(1,30)	(– 0,98)	(– 0,05)	(– 1,16)
Moyenne	0,15968	0,25761	0,24693	0,20297	0,13280	0,09501
• Population favorisée	– 0,00550	0,00112	0,00845	– 0,00906	0,00499	– 0,00170
(N = 23 213)	(– 0,63)	(0,10)	(0,49)	(– 0,56)	(0,17)	(– 0,30)
Moyenne	0,04411	0,13846	0,24038	0,30069	0,27635	0,04295
• Population intermédiaire	– 0,00400	0,00465	0,00392	– 0,00214	– 0,00242	– 0,00295
(N = 40 906)	(– 0,39)	(0,45)	(0,33)	(– 0,20)	(– 0,21)	(– 0,34)
Moyenne	0,12670	0,25568	0,26945	0,21811	0,12976	0,08487
• Population défavorisée	– 0,0184	0,00615	0,0192	– 0,00824	0,00133	– 0,0127
(N = 46 886)	(– 1,45)	(0,56)	(1,62)	(– 0,81)	(0,15)	(– 1,52)
Moyenne	0,24540	0,31828	0,23054	0,14139	0,06439	0,12971
Stratégie 1. Traitement intense						
• Population totale	0,00863	0,00140	0,0147	– 0,0148	– 0,00987	– 0,00890
(N = 50 491)	(0,79)	(0,10)	(1,16)	(– 1,43)	(– 0,61)	(– 1,18)
Moyenne	0,14107	0,26013	0,26486	0,20911	0,12483	0,09933
• Population favorisée	0,000648	0,00160	0,0362*	– 0,0289	– 0,00959	0,00513
(N = 10 491)	(0,07)	(0,07)	(1,73)	(– 1,31)	(– 0,31)	(0,55)
Moyenne	0,04185	0,15432	0,25088	0,30331	0,24964	0,05186
• Population intermédiaire	0,0133	– 0,00103	0,0171	– 0,00991	– 0,0195	– 0,00368
(N = 21 114)	(0,97)	(– 0,06)	(1,06)	(– 0,60)	(– 1,14)	(– 0,35)
Moyenne	0,11689	0,26603	0,28000	0,21805	0,11902	0,09234
• Population défavorisée	0,00987	0,00993	– 0,000220	– 0,0127	– 0,00688	– 0,0211
(N = 18 886)	(0,53)	(0,59)	(– 0,01)	(– 0,90)	(– 0,45)	(– 1,50)
Moyenne	0,22323	0,31229	0,25569	0,14678	0,06200	0,13346

Lecture : Les estimations sont issues de régressions moindres carrés ordinaires. Le coefficient reporté dans ce tableau est celui de la variable d'interaction entre l'indicatrice « traitement » et l'indicatrice « année postérieure à la mise en place de l'évaluation par compétence ». La régression inclut également des effets fixes « collège » et des effets fixes « année ». Les termes d'erreurs sont clusterisés au niveau des collèges pour tenir compte des facteurs communs partagés par les élèves issus du même établissement. Les statistiques t de Student sont entre parenthèses.

Notes : (***) Significatif au seuil de 1 % ; (**) Significatif au seuil de 5 %, * Significatif au seuil de 10 %.

Source : Calculs des auteurs d'après bases 'Apprenants' et 'Océan DNB' (2007-2016), DEPP, ministère de l'Éducation nationale.

(9) Toutefois, même en supposant que le biais est maximal (donc que 1,3 % d'élèves défavorisés qui n'ont pas redoublé dans les collèges traitement ont tous été non admis), la baisse du taux de non admis chez les élèves défavorisés ne serait pas très importante : elle serait de 1,8 + 1,3 = 3,1 points de pourcentage au lieu de 1,8 point, soit une baisse de 13 %. Une telle baisse constitue la borne supérieure de l'effet de l'évaluation par compétence en 6e sur le taux d'échec au DNB.

6. Interprétation et conclusion

On peut donc conclure que l'évaluation par compétence n'a pas d'effets statistiquement détectables sur les apprentissages tels que mesurés par les épreuves obligatoires du DNB en fin de collège. Les élèves, qu'ils aient été exposés à l'évaluation par compétence seulement en 6e ou pendant tout ou presque leur scolarité au collège, ont obtenu les mêmes résultats au DNB que les élèves exposés au mode d'évaluation classique par notes chiffrées. Cette absence d'effets sur les résultats au DNB se retrouve chez tous les élèves quelle que soit leur origine sociale. Si quelque chose se passe, il s'agirait d'une légère baisse du taux de non-admis chez les élèves défavorisés, mais ceci reste davantage une piste à approfondir qu'une certitude. Un tel résultat pourrait être confirmé ou infirmé en prenant des échantillons plus importants, ce qui sera possible dans les années futures grâce à la montée en puissance de l'évaluation par compétence.

L'absence d'effet de l'évaluation par compétence pourrait refléter un manque de transformation profonde des pratiques pédagogiques autour de l'évaluation des élèves. L'une des limites de cette étude est le statut auto-déclaratif de la pratique de « l'évaluation par compétence ». Les établissements concernés n'évoluent pas dans un cadre défini incluant un référentiel commun et une formation spécifique à destination des enseignants. Cette évaluation d'impact s'est donc attachée à quantifier l'impact de pratiques relativement hétérogènes malgré un affichage uniforme. Il est intéressant de constater qu'en moyenne les pratiques mise en œuvre ne produisent pas d'effets sur les apprentissages en fin de collège, ce qui n'exclut pas bien évidemment que dans tel ou tel cas spécifique – et minoritaire – des effets soient apparus.

De plus, l'évaluation formative est une pratique pédagogique riche qui nécessite une transformation radicale qui va au-delà d'un changement de métrique, comme le passage de l'échelle de notation sur vingt à des points de couleurs, des mentions « acquis », etc. Comme nous l'avons rappelé en introduction de ce Focus, l'évaluation formative implique un changement radical dans l'objectif même de l'évaluation, en passant d'une évaluation des apprentissages à une évaluation pour les apprentissages, c'est-à-dire dont le but premier est de motiver les élèves et de les encourager à progresser. Une recherche complémentaire devrait donc être menée pour évaluer si les établissements qui utilisent l'évaluation par compétence ont réellement changé leurs pratiques d'évaluation dans le sens de l'évaluation formative, ou s'ils sont restés finalement plus proche de l'approche sommative bien qu'utilisant une grille détaillée compétence par compétence. L'évaluation formative requiert une transformation structurelle de la façon d'enseigner qui n'a peut-être pas eu lieu : mettre en avant les points forts et les points faibles de chacun, renforcer la confiance en soi et le sentiment d'auto-efficacité, accorder le droit à l'erreur, apporter des conseils individualisés pour progresser. Une méthodologie claire, ainsi que des formations seraient nécessaires pour observer une véritable application de l'évaluation formative. Les collèges engagés dans l'innovation pédagogique devraient pouvoir bénéficier d'un cadre permettant de systématiser l'évaluation scientifique de leurs pratiques et de les disséminer dans d'autres établissements.

L'arrivée de nouvelles cohortes traitées en troisième ainsi que l'institutionnalisation progressive de l'évaluation formative devraient permettre de mener de nouvelles évaluations d'impact et de dépasser les limites évoquées ci-dessus.

Références bibliographiques

Brookhart S.M. (2004) : « Assessment Theory for College Classrooms", *New Directions for Teaching and Learning*, n° 100, pp. 5-14.

Butler R. (1988) : « Enhancing and Undermining Intrinsic Motivation: The Effects of Task-Involving and Ego-Involving Evaluation of Interest and Performance », *British Journal of Educational Psychology*, vol. 58, n° 1, pp. 1-14.

Butler R. et M. Nisan (1986) : « Effects of No Feedback, Task-Related Comments, and Grades on Intrinsic Motivation and Performance », *Journal of Educational Psychology*, n° 78, pp. 210-216.

Caron C. (2018) : *La cohorte coopérative*, Rapport sur le projet de classes coopératives au Collège Lucien Vadez à Calais. Année scolaire 2017-2018, Bibliothèque des expérimentations pédagogiques, 'Expérithèque Lille'.

Crouzevialle M. et F. Butera (2013) : « Performance-Approach Goals Deplete Working Memory and Impair Cognitive Performance », *Journal of Experimental Psychology: General*, vol. 142, n° 3, pp. 666-678.

Crouzevialle M., A. Smeding et F. Butera (2015) : « Striving for Excellence Sometimes Hinders High Achievers: Performance-Approach Goals Deplete Arithmetical Performance » in *Students with High Working Memory Capacity*, Sutherland (ed.) PLoS ONE.

Duflo E. (2001) : « Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment », *American Economic Review*, vol. 91, n° 4, pp. 795-813.

Fuchs L.S. et D. Fuchs (1986) : « Effects of Systematic Formative Evaluation: A Meta-Analysis », *Exceptional Children*, vol. 53, n° 3, pp. 199-208.

Hattie J. (2009) : *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*, Routledge.

Hayek A.S., C. Toma, D. Oberlé et F. Butera (2015) : « Grading Hampers Cooperative Information Sharing in Group Problem Solving », *Social Psychology*, vol. 46, n° 3, pp. 121-131.

Hayek A.S., C. Toma, S. Guidotti, D. Oberlé et F. Butera (2017) : « Grades Degrade Group Coordination: Deteriorated Interactions and Performance in a Cooperative Motor Task », *European Journal of Psychology of Education*, vol. 32, n° 1, pp. 97-112.

Hudesman J., S. Crosby, B. Flugman, S. Issac, H. Everson et D.B. Clay (2013) : « Using Formative Assessment and Metacognition to Improve Student Achievement », *Journal of Developmental Education*, vol. 37, n° 1, pp. 2-13.

Jones E.E. et S. Berglas (1978) : « Control of Attributions About the Self Through Self-Handicapping Strategies: The Appeal of Alcohol and the Role of Under Achievement », *Personality and Social Psychology Bulletin*, n° 4, pp. 200-206.

Pulfrey C., C. Darnon et F. Butera (2013) : « Autonomy and Task Performance: Explaining the Impact of Grades on Intrinsic Motivation », *Journal of Educational Psychology*, vol. 105, n° 1, pp. 39-57.

Zuckerman N., S.C. Kieffer et C.R. Knee (1998) : « Consequences of Self-Handicapping: Effects on Coping, Academic Performance, and Adjustment », *Journal of Personality and Social Psychology*, vol. 74, n° 6, pp. 1619-1628.