Qu'est-ce qu'un bon lycée ? Mesurer les effets établissements, au-delà de la moyenne

What Makes a Good High School? Measuring School Effects beyond the Average

Pauline Givord* et Milena Suarez Castillo**

Résumé – Évaluer la capacité des établissements scolaires à faire progresser leurs élèves est un exercice complexe, car il est difficile de distinguer ce qui relève de l'effet propre de l'établissement de ce qui relève des caractéristiques des élèves qui y sont scolarisés. Cet article commence par présenter les deux principaux modèles statistiques (modèles de valeur ajoutée et modèles dits de *Student Growth Percentile*) qui sont couramment utilisés, et discute leurs apports et limites à la lumière de la littérature récente. Il propose ensuite des indicateurs qui permettent de compléter les mesures classiques de la valeur ajoutée des établissements, en évaluant notamment si les résultats obtenus par les élèves d'un lycée sont plus ou moins dispersés par rapport à ce qui serait attendu compte tenu des caractéristiques de ses élèves. Ces indicateurs sont notamment utiles pour évaluer la pertinence de l'information fournie par des indicateurs sur les effets moyens des établissements. Cette méthode est appliquée à partir des données exhaustives des notes au baccalauréat de la session 2015.

Abstract *– Assessing the ability of schools to help their students to progress is a complex exercise, as it is difficult to distinguish between the effects brought about by the school itself and those resulting from the characteristics of the students they enrol. This article starts by describing the two main statistical models currently in use (Value-Added models and Student Growth Percentile models) and discusses their advantages and limitations in the light of recent literature. It then proposes indicators to complement the traditional measures of the value-added of schools, in particular by assessing whether the results achieved by the students of a high school are more or less dispersed than would be expected given the characteristics of its students. These indicators are useful for assessing the relevance of the information provided by the indicators on average effect of the schools. This method is applied using exhaustive data on baccalaureate grades from 2015.*

Codes JEL / JEL Classification : CI20, C21, C50
Mots-clés : régression quantile, valeur ajoutée, student growth percentile
Keywords: quantile regression, school value added, student growth percentile

** Insee-LIEPP (pauline.givord@travail.gouv.fr) ; **CREST-Insee (milena.suarez-castillo@insee.fr)*
Nous remercions Fabrice Murat, Cédric Afsa, Caroline Simonis-Sueur et Fabienne Rosenwald pour avoir autorisé l'accès aux données, ainsi que, pour leur accueil à la Depp et leurs conseils, Olivier Monso, Thierry Rocher, Franck Evain et Laetitia Evrard. Nous remercions les participants du séminaire du Département des études économiques de l'Insee, et notamment Marco Paccagnella pour sa discussion stimulante sur une version précédente de cet article, ainsi que deux rapporteurs anonymes pour leur relecture précise et attentive qui a permis d'en améliorer la lecture. Les auteures restent seules responsables des erreurs ou approximations qui pourraient néanmoins subsister.

Reçu en octobre 2020, accepté en mai 2021.

Citation: Givord, P. & Suarez Castillo, M. (2021). What Makes a Good High School? Measuring School Effects beyond the Average. *Economie et Statistique / Economics and Statistics*, 528-529, 29–45. doi: 10.24187/ecostat.2021.528d.2057

Qu'est-ce qu'un bon lycée ? Sur les dernières décennies, de très nombreux travaux de recherche se sont penchés sur la mesure des effets des établissements sur la réussite des élèves, afin d'améliorer l'information disponible, accompagnant une demande et un intérêt croissants pour l'évaluation. En France, l'évaluation du système éducatif est depuis longtemps identifiée, et régulièrement réaffirmée comme importante dans l'amélioration de la qualité du service public de l'éducation, depuis les premières réflexions sur la nécessité d'une culture de l'évaluation (Thélot, 1994a ; 1994b) à la création en 2019 du Conseil d'évaluation de l'école. Pour les lycées, cela s'est traduit notamment par la publication régulière par la Direction de l'évaluation, de la prospective et de la performance (Depp, le service statistique du ministère de l'éducation nationale) des Indicateurs de Valeur Ajoutée des Lycées (les IVAL). Les IVAL fournissent un ensemble d'indicateurs sur la performance des lycées français en matière de réussite de leurs élèves au baccalauréat, mais également sur leur capacité à les accompagner jusqu'à l'examen final, en s'attachant notamment à tenir compte du profil des élèves accueillis (voir encadré 1). Dans la plupart des pays anglo-saxons, l'évaluation des établissements, essentiellement sur des critères quantitatifs, est plus ancienne, avec le développement au cours des années 80 de la culture du résultat et de l'idée de rendre responsables les établissements de la réussite de leurs élèves (dans des optiques notamment de développement du choix scolaire). L'exemple le plus emblématique de cette évolution est l'adoption en 2001 aux États-Unis de la loi fédérale *No Child Left Behind*, qui imposait aux États américains de mettre en place des tests annuels pour l'ensemble des élèves, avec des incitations fortes pour les établissements à atteindre des objectifs de réussite des élèves. Pour répondre à ces objectifs, la plupart des États ont développé des outils de mesure des établissements, voire des enseignants.

L'usage qui peut être fait de ces évaluations répond à au moins deux objectifs principaux, qui soulèvent des enjeux de mesure différents. Le premier, qui est sous-jacent notamment dans le développement de ce type de mesures dans les pays anglo-saxons, vise à fournir aux autorités publiques en charge du pilotage des établissements des instruments d'évaluation de leur efficacité, voire de leur efficience. Il pourra s'agir par exemple de mettre en regard des bons résultats d'un lycée (ou à l'inverse, des résultats décevants), avec les pratiques et les moyens mis en œuvre. Comme souligné par exemple par Raudenbush & Wilms (1995), cet objectif est particulièrement complexe, dans la mesure où les établissements scolaires n'ont pas le contrôle sur l'ensemble des dimensions qui peuvent influer sur la réussite des élèves. C'est par exemple le cas de l'influence des autres élèves sur la réussite individuelle. Ces « effets de pairs » sur la réussite sont complexes, et surtout très difficiles à mesurer (pour une synthèse récente, voir par exemple Monso *et al*., 2019). Tel qu'il est mesuré, il est donc en général impossible de distinguer dans l'effet de l'établissement sur la réussite des élèves ce qui relève de son action de ce qui relève des effets de l'interaction des élèves qu'il scolarise. En revanche, les mesures des effets lycées peuvent être utiles pour un second objectif, plus modeste, qui est de fournir aux familles une indication sur l'effet qu'elles peuvent attendre de la scolarisation dans un établissement plutôt que dans un autre, que cet effet soit lié aux pratiques de l'établissement ou à des éléments de contexte liés aux interactions avec d'autres élèves.

Même en se limitant à cet objectif d'information des familles, identifier des outils de mesure pertinents est complexe. D'abord parce que les parents peuvent avoir des critères différents pour définir ce qu'est un bon lycée. Certes, pour la plupart des parents, il s'agit d'un établissement capable d'accompagner leurs enfants jusqu'au baccalauréat, en leur assurant une scolarité sereine tout en les préparant au mieux pour l'avenir. Néanmoins, l'appréciation de la manière dont un établissement répond à ces objectifs peut varier selon les élèves. Certains adolescents pourront s'épanouir dans des établissements encourageant l'émulation et l'exigence académique quand d'autres souffriront d'une ambiance trop compétitive. Au-delà des strictes performances académiques, certains peuvent valoriser la capacité des enseignants à développer le goût d'apprendre et la confiance en soi des élèves, la qualité du climat scolaire, ou l'aide apportée aux élèves pour construire leur orientation future et pour la rendre possible.

Quelle que soit la définition que l'on se donne d'un bon lycée, identifier un établissement qui correspond aux critères retenus est encore plus complexe. Il faudrait pour cela déterminer ce qu'aurait été la scolarité d'un élève dans un autre établissement que celui où il a été scolarisé, ce qui est difficile voire impossible. En général, les parents ne disposent que de peu d'éléments pour juger d'un établissement. Les expériences passées de connaissances ou de la fratrie, le taux de réussite au baccalauréat sont des informations

certes utiles, mais qui ne renseignent qu'indirectement sur la manière dont un élève particulier s'adaptera finalement à un lycée. La réussite affichée par un lycée est avant tout le reflet des caractéristiques des élèves qui y sont scolarisés, et ne pas tenir compte des effets de sélection peut donner des images biaisées de la qualité des établissements et donc des informations peu pertinentes pour les familles. C'est pourquoi les indicateurs, comme ceux développés par la Depp, tiennent compte du niveau initial des élèves.

Les indicateurs les plus fréquemment utilisés se concentrent sur des effets moyens. Ces moyennes peuvent cependant masquer des disparités : un même effet moyen peut ainsi être mesuré aussi bien pour un lycée qui permet de faire progresser un peu tous ses élèves que pour un lycée qui ferait progresser beaucoup une minorité d'élèves. L'information apportée par l'indicateur sera plus ou moins pertinente, notamment pour des parents qui utiliseraient ces mesures pour scolariser leurs enfants dans le lycée qui permettrait une meilleure scolarité. Cette étude se propose donc d'enrichir la description de l'effet des lycées en fournissant des indicateurs qui visent à caractériser les lycées en fonction de leur propension à amplifier, ou au contraire réduire, les inégalités de performance à l'examen du baccalauréat par rapport à ce qui est attendu au regard des caractéristiques des élèves[1]. La suite de l'article commence par proposer une revue de la large bibliographie sur la mesure des effets lycées, puis détaille l'approche utilisée ici sur les lycées français des voies générales et

1. L'évaluation complète d'un établissement scolaire nécessite des informations sur ces moyens, et dépasse le propos de cet article, qui s'intéresse à la mesure de l'effet d'un établissement dans l'amélioration de la réussite scolaire de ses élèves.

Encadré 1 – **Les indicateurs de valeur ajoutée des lycées (IVAL)**

Les IVAL sont diffusés par le service statistique du ministère de l'Éducation, aujourd'hui la Direction de l'évaluation, de la prospective et de la performance, depuis 1993 (pour une présentation détaillée voir Evain, 2020).

Si la méthodologie de leur construction a évolué au cours du temps, leur objectif est de permettre des comparaisons entre des établissements en tenant compte des différences initiales des élèves qu'ils scolarisent. La « valeur ajoutée » des lycées y est mise en évidence en comparant ce qui est attendu compte tenu des caractéristiques de leurs élèves (notamment en termes de niveau scolaire et d'origine sociale), tel que prédit par un modèle, et les résultats des élèves effectivement observés.

Pour rendre compte de la difficulté d'évaluer par un seul indicateur l'action d'un établissement, plusieurs indicateurs sont proposés. Le premier s'intéresse à la probabilité, pour un élève inscrit, de réussir l'examen du baccalauréat : c'est celui qui s'apparente le plus directement aux palmarès publiés par les médias, mais en tenant compte ici de la composition initiale des établissements.

Cet indicateur sur la réussite au baccalauréat est complété par la probabilité d'obtenir ce diplôme en ayant effectué sa scolarité dans l'établissement depuis la seconde ou depuis la première, le « taux d'accès ». L'analyse des valeurs ajoutées des taux d'accès permet en creux de ne pas survaloriser les établissements dont la politique « d'écrémage » sélectionnerait les meilleurs élèves au fur et à mesure de la scolarité au lycée : ces établissements peuvent afficher de très bons résultats à l'examen final, mais au prix de l'abandon des élèves les moins prometteurs. À l'inverse, une valeur ajoutée élevée pour le taux d'accès traduit la capacité de l'établissement à accompagner ses élèves sur l'ensemble de leur scolarité[(a)].

Enfin, depuis 2017, la valeur ajoutée est également calculée pour la probabilité d'obtenir une mention à l'examen. Cela permet de mieux rendre compte des disparités de niveau entre les élèves, au-delà du seul fait d'obtenir le diplôme. En effet, le taux de réussite au baccalauréat est devenu assez peu discriminant compte tenu des niveaux très élevés observés, notamment dans les filières générales et technologiques : à la session 2019, le taux de réussite au baccalauréat était ainsi de 91 % en moyenne pour la filière générale, de 88 % pour la filière technologique, et de 82 % pour la filière professionnelle. S'intéresser à la probabilité d'obtenir une mention (soit au moins 12/20 de moyenne à l'examen) permet de distinguer plus finement les établissements entre eux.

En pratique, les valeurs ajoutées sont calculées à partir d'une modélisation logistique de la probabilité de réussite, en utilisant une modélisation à effets aléatoires pour tenir compte des effets lycées (pour des détails, voir Duclos & Murat, 2014 et Evain & Evrard, 2017). Le modèle intègre des variables individuelles des élèves : niveau scolaire, indice de position sociale[(b)], âge et sexe[(c)]. Les corrélations observées entre ces caractéristiques individuelles sont utilisées pour estimer la probabilité de réussite prédite par le modèle, ce qui agrégé au niveau du lycée permet de calculer le taux attendu de réussite. La valeur ajoutée correspond à la différence entre le taux observé et le taux attendu.

(a) Une limite cependant à l'indicateur mesurant le taux d'accès est qu'il ne permet pas de distinguer ce qui peut aussi relever des mobilités volontaires des élèves de pratiques spécifiques des établissements.
(b) L'Indice de position sociale est une mesure synthétique continue des dimensions sociales, économiques et culturelles associées à la réussite scolaire, selon la profession et la catégorie sociale (PCS) des parents (Rocher, 2016).
(c) En outre, les moyennes de ces variables sont ajoutées au modèle (voir discussion dans l'encadré 2), ce qui permet de tenir compte du fait que ces estimations des variables individuelles peuvent être biaisées dès lors qu'elles sont corrélées aux caractéristiques inobservées du lycée.

technologiques, à partir des résultats obtenus à la session 2015 du baccalauréat.

1. Mesurer l'efficacité d'un établissement ou d'un enseignant : questions méthodologiques et enjeux d'interprétation

1.1. Des effets de sélection qui rendent difficiles la mesure des effets propres des établissements ou des enseignants

L'une des difficultés majeures pour mesurer la capacité d'un établissement, ou d'un enseignant, à faire progresser ses élèves tient à l'existence d'effets de sélection importants (Felouzis, 2005). Par exemple, un lycée qui sélectionne ses élèves en fonction de leur dossier scolaire au moment de l'entrée en seconde affichera évidemment un taux de réussite au baccalauréat très élevé. Cela ne signifie pas qu'il peut être crédité d'un effort particulier pour faire progresser ses élèves. Cela ne signifie pas non plus que tout élève qui serait scolarisé dans un tel établissement, quel que soit son niveau de départ, sera assuré d'obtenir d'aussi bons résultats. En général, les établissements ne scolarisent pas les mêmes élèves, et à l'intérieur des établissements les élèves n'ont pas face à eux les mêmes enseignants. La réussite apparente de certains peut traduire simplement des différences de niveau initial entre les élèves. Ces mêmes questions se posent s'il s'agit de mesurer des « effets enseignants » (on parle aussi « d'effets maîtres »), c'est-à-dire évaluer dans quelle mesure l'action d'un enseignant a pu apporter plus ou moins aux élèves. Ces questions sont centrales dans les systèmes scolaires qui ont institutionnalisé la rémunération à la performance, comme dans certains États américains. Pour cette raison, une large littérature s'est intéressée à la question notamment de la mesure des effets enseignants (voir notamment Chetty *et al*., 2014). Si les déterminants sous-jacents des effets enseignants ou établissements sont bien évidemment différents, les deux soulèvent des questions méthodologiques identiques sur le plan statistique.

Pour comparer deux enseignants, ou deux lycées, il faudrait dans l'idéal pouvoir comparer leur capacité à faire progresser les mêmes types d'élèves. Mesurer l'effet spécifique pour un établissement demanderait en théorie de pouvoir affecter aléatoirement des élèves de profil identique dans les lycées et les classes, un exercice dont la faisabilité, pour des raisons pratiques voire éthiques, est très limitée. La plupart des modèles développés pour mesurer des effets établissements visent à réduire les biais liés aux effets de compositions différenciées des établissements ou des classes en contrôlant le niveau initial des élèves. Deux grands types de modèles ont été développés dans ce cadre : les modèles de valeur ajoutée et les modèles de *Student Growth Percentile* (« percentile de progrès des élèves »).

1.1.1. Deux modèles statistiques : les modèles de valeur ajoutée et les modèles de Student Growth Percentile

Dans leur forme la plus simple, les modèles de valeur ajoutée supposent que la variable d'intérêt (par exemple, les notes moyennes à l'examen du baccalauréat) dépend pour chaque élève de ses résultats antérieurs, d'un certain nombre de caractéristiques observables comme son niveau initial ou son milieu d'origine, et d'un effet propre à l'établissement. Ce dernier est capté en introduisant dans la modélisation une indicatrice commune à l'ensemble des élèves de l'établissement. Ce type de modèle est utilisé par exemple par la Depp pour mesurer la valeur ajoutée des lycées pour un ensemble d'indicateurs, dont notamment la probabilité de réussite au baccalauréat ou celle d'obtenir une mention, ainsi que la probabilité de réussir au baccalauréat en ayant effectué toute sa scolarité dans l'établissement (voir encadré 1).

Les modèles de *Student Growth Percentile* (SGP ensuite) ont été notamment développés aux États-Unis par l'État du Colorado (Betebenner, 2007) suivi par 18 autres États américains, tandis que les modèles de valeur ajoutée sont utilisés dans 15 États, le pionnier étant le Tennessee (voir Kurtz, 2018 pour une revue). Ces modèles ont l'intérêt, pour un usage opérationnel, d'être assez simples à interpréter. Leur principe repose sur la question suivante : comment un élève a-t-il réussi par rapport à des élèves qui avaient des résultats équivalents sur les tests antérieurs ? Les élèves sont « classés » selon leurs résultats à des tests, le rang dans ce classement étant représenté par le percentile dans la distribution des notes. Par exemple, si un élève fait mieux à un test de fin d'année que 80 % des élèves qui avaient un niveau proche du sien en début d'année, on attribuera au lycée un effet positif de 80 pour cet élève. L'efficacité de l'établissement (ou de l'enseignant) correspondra alors à la moyenne (ou la médiane) de ces effets mesurés pour l'ensemble des élèves de l'établissement (ou de l'enseignant). En pratique, ces estimations sont effectuées à partir de régressions quantiles, qui permettent de modéliser la distribution des notes à un test conditionnellement aux résultats aux tests précédents (voir encadré 2).

1.1.2. Limites statistiques des deux modèles

La mesure des effets établissements ou des effets enseignants a fait l'objet d'une intense recherche méthodologique. Cet intérêt s'explique par le fort enjeu que peuvent avoir ces indicateurs. Alors que la qualité perçue des établissements peut être un facteur important du choix d'un lycée par les parents, la publication de palmarès peut contribuer à amplifier les écarts initiaux – notamment parce que les parents les plus informés ou ayant les moyens de choisir l'établissement où scolariser leur enfant ont souvent un capital scolaire plus élevé. De manière plus radicale, ces méthodes sont parfois utilisées, comme au Royaume-Uni ou dans certains États américains, pour mesurer l'efficacité des établissements ou des enseignants avec des conséquences qui peuvent être importantes : incitations financières à la performance pour les enseignants, voire fermeture d'écoles – ou licenciement d'enseignants – dont l'efficacité est évaluée comme insuffisante[2]. Compte tenu du fort enjeu pour les acteurs concernés, disposer d'instruments valides et pertinents est crucial[3]. Cependant, les outils disponibles sont l'objet de critiques de plusieurs ordres.

En premier lieu, la plupart des contributions soulignent la difficulté de ces modèles à dépasser les limites liées notamment à l'absence de randomisation (pour une synthèse voir par exemple Everson, 2016). En particulier, la mesure des effets enseignants ou des effets établissements s'avère très sensible aux variables utilisées pour contrôler les effets de composition. Ne pas tenir compte dans les modèles de certaines des caractéristiques des élèves qui peuvent influer sur leur progression scolaire, comme leur origine sociale, réduit fortement la capacité de ces modèles à discriminer entre le fait d'avoir une pédagogie efficace et celui d'enseigner devant des élèves issus de milieux plus favorables à la réussite scolaire. Les modèles de SGP, tels qu'ils sont couramment utilisés, ne tiennent pas compte de ces dimensions et sont donc particulièrement sujets à cette critique (Guarino *et al*., 2015a). Les différentes comparaisons suggèrent que ces indicateurs ont tendance à pénaliser les enseignants face à des élèves issus de milieu social défavorisé ou avec des besoins particuliers, par rapport à des modèles de valeur ajoutée qui tiennent compte de ces dimensions (Walsh & Isenberg, 2015). Dans la mesure où toutes les informations qui seraient nécessaires ne sont pas toujours disponibles, cette question se pose également pour les modèles de valeur ajoutée. Le type de variables utilisées pour contrôler les effets de composition dans cette catégorie de modèle peut également affecter les conclusions que l'on en tire (Ehlert *et al*., 2014 ; Sass *et al*., 2014), tout comme la spécification statistique retenue (Guarino *et al*., 2015b ; Soland, 2016). Par ailleurs, comme discuté en introduction, une partie des effets de composition sur la réussite passe par les interactions entre les élèves, particulièrement complexes à mesurer (pour une mesure sur les lycées français, voir par exemple Boutchenik & Maillard, 2019), et dont l'effet propre est en général impossible à distinguer de l'effet de l'établissement.

Plus généralement, certains auteurs sont très sceptiques sur la possibilité de réduire les biais de sélection, liés notamment au fait que les caractéristiques des élèves et celles des enseignants qu'ils ont en face d'eux ne sont pas indépendantes (Rothstein, 2010 ; Sass *et al*., 2014), même si d'autres sont plus confiants sur la possibilité de s'appuyer par exemple sur la mobilité des enseignants entre établissements et entre classes pour évaluer ces effets (Chetty *et al*., 2014 ; Koedel *et al*., 2015). Par ailleurs, les effets mesurés par ces modèles peuvent être très imprécis, notamment parce qu'ils sont estimés sur des petits nombres d'observations. Une étude récente observe par exemple qu'il est possible de mettre en évidence à partir de ces modèles des pseudo-effets des enseignants sur… la taille de leurs élèves, pourtant une caractéristique qui n'est pas susceptible d'être modifiée par les pratiques pédagogiques (Bitler *et al*., 2019). Les auteurs montrent que ce résultat paradoxal s'explique par la faible taille des échantillons sur lesquels sont menées les estimations, qui conduit à attribuer à tort à l'enseignant ce qui n'est qu'un bruit statistique. Si cet effet disparaît lorsque l'on utilise des observations obtenues sur plusieurs années, cette solution n'est pas toujours retenue pour évaluer par exemple la valeur ajoutée des enseignants.

1.2. Retour à la question : peut-on mesurer ce qu'est un bon lycée ?

Au-delà de ces questions méthodologiques, l'utilisation de ce type d'instruments dans

2. Ainsi, la loi No Child Left Behind *mentionnée en introduction, qui imposait que toutes les écoles publiques devaient montrer des « progrès annuels adéquats » dans la performance de leurs élèves, telle que mesurée par des tests annuels, avec un ensemble de sanctions et d'incitations en cas d'échec à remplir les objectifs. L'échec répété à remplir ces objectifs pendant six années consécutives prévoyait un plan de restructuration complète de l'établissement, qui pouvait inclure sa fermeture, le licenciement de l'ensemble de l'équipe ou sa transformation en charter school (pour une discussion en français voir par exemple Gamoran, 2012). Cette loi a été abrogée en 2015.*

3. Pour une critique de ces pratiques, notamment étant donné les limites inhérentes à l'exercice de mesure sous-jacent, voir par exemple Jacob (2005).

l'évaluation des enseignants a aussi été critiquée pour le fait qu'elle tend à se focaliser sur ce qu'on sait le mieux mesurer (la réussite des élèves à des tests scolaires) au détriment de compétences plus fondamentales, comme la capacité des enseignants à développer chez leurs élèves la confiance en soi, l'envie d'apprendre ou l'esprit critique, dimensions qui ne se recoupent que partiellement avec les compétences cognitives. Par exemple, une étude américaine utilise l'affectation aléatoire d'élèves dans des classes dans le cadre d'une expérience randomisée pour comparer les effets enseignants sur les résultats à des tests standardisés à ceux obtenus à des questions ouvertes, ou sur l'effort et la motivation des élèves, et montre une très faible corrélation entre ces différentes dimensions (Kraft, 2019). Une autre étude montre également que les effets des enseignants sur la réussite de leurs élèves à des tests sont peu corrélés avec leurs effets sur le comportement de ces derniers (comme l'absentéisme ou le redoublement), alors même que ces dimensions sont plus prédictives de la réussite future des élèves (Jackson, 2018).

Par ailleurs, dès lors que les évaluations revêtent un fort enjeu – c'est en particulier le cas de dispositifs financiers (prime aux résultats) pour les enseignants, ou simplement de la réputation d'un établissement qui est importante pour la qualité des élèves qu'il scolarisera dans le futur – elles peuvent induire des comportements stratégiques des personnes concernées avec de possibles effets contraires à ceux escomptés (pour une contribution récente, voir Fryer, 2013 et pour une revue, Jacob, 2005). En particulier, les tentatives de manipuler les indicateurs sont fréquentes. Cela peut consister à consacrer un temps disproportionné de l'enseignement à préparer les élèves aux tests (phénomène de *teaching to the test*, voir Wall, 2000). Les modèles de SGP sont *a priori* moins susceptibles d'induire ces phénomènes de bachotage (Barlevy & Neal, 2002), parce que la mesure des effets établissements ou des effets enseignants s'appuie sur une métrique relative (la progression des élèves par rapport à ceux de même niveau initial), tandis que les modèles de valeur ajoutée demandent, pour permettre des comparaisons fiables et justes dans le temps, d'utiliser des tests standardisés dont le format et le contenu varient peu. Néanmoins, la sensibilité de ces deux modèles aux caractéristiques des élèves autres que leur niveau initial peut conduire les établissements ou les enseignants qui sont évalués à cette aune à minimiser les risques. Les établissements peuvent par exemple sélectionner les élèves les plus prometteurs, ou exclure en cours de scolarité ceux qui n'obtiennent pas des résultats suffisants. Les enseignants, lorsqu'ils peuvent choisir leur affectation, ont tendance à éviter les établissements concentrant les élèves les plus en difficulté (Walsh & Isenberg, 2015), ce qui signifie que ce sont souvent les enseignants qui n'ont pas le choix (souvent les moins qualifiés ou les moins expérimentés) qui se trouvent devant les élèves dont les besoins sont les plus grands.

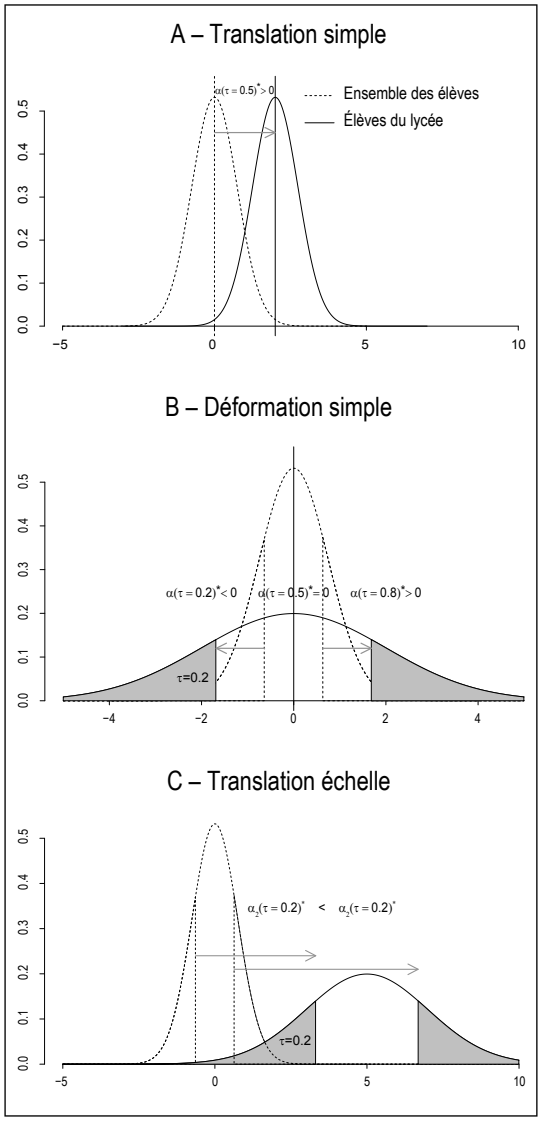
2. Mesurer la dispersion au-delà de la moyenne

Même lorsque l'on se limite aux indicateurs de performance scolaire, la qualité des établissements peut être questionnée au-delà des effets moyens qui sont mesurés classiquement. Une apparente similitude entre deux effets moyens peut masquer des réalités très différentes : un même effet moyen positif peut résulter de l'action soit d'un établissement dans lequel tous les élèves progressent, soit d'un établissement dans lequel seule une minorité d'élèves réussissent très bien tandis que d'autres au contraire ont des résultats bien plus faibles qu'attendus au regard de leurs caractéristiques.

Cette étude se propose donc d'enrichir la description qui peut être faite par les mesures classiques de l'effet des lycées en fournissant des indicateurs au-delà de la moyenne. L'objectif ici est de mesurer les effets lycée non seulement sur la moyenne de la distribution des notes, mais aussi d'évaluer dans quelle mesure un lycée tend, par rapport à des lycées identiques qui lui sont proches en termes de caractéristiques des élèves accueillis, à avoir des résultats au baccalauréat plus dispersés ou à l'inverse plus homogènes. L'intuition est illustrée par la figure I, à partir d'un exemple fictif représentant les densités théoriques des notes, telles qu'attendues en fonction des caractéristiques des élèves et en tenant compte de l'effet du lycée, dans trois cas distincts. Le premier (figure I-A) correspond à une situation où l'effet du lycée est le même pour l'ensemble des élèves : par rapport à la distribution des notes attendue, la distribution des notes observée dans ce lycée est simplement déplacée vers la droite si cet effet est positif, ou vers la gauche lorsqu'il est négatif, mais sa forme est la même. Le deuxième cas (figure I-B) représente à l'inverse une situation où le lycée a un effet très différent selon les élèves : les élèves les plus faibles y ont un niveau plus bas que celui auquel on s'attend, tandis que les meilleurs élèves obtiennent à l'inverse des notes meilleures qu'attendues. Dans ce cas fictif, l'effet est totalement symétrique et il n'y a donc

pas d'effet sur la moyenne des notes (l'effet moyen estimé sera nul), mais la dispersion des notes observée est beaucoup plus large. Enfin, le troisième cas (figure I-C) combine les deux précédents : l'effet du lycée est à la fois positif en moyenne et il a aussi tendance à augmenter la dispersion des notes.

Figure I – **Illustration des effets d'un lycée sur la dispersion et la moyenne de la distribution des notes (modèle de translation échelle)**



L'intention est ici de modéliser les effets du lycée à plusieurs niveaux de la distribution des notes dans le lycée. Pour cela, on utilise une technique statistique, celle des régressions quantiles, détaillée dans l'encadré 2. Cette modélisation permet d'aller plus loin que ce cas fictif, qui suppose que les effets sont forcément symétriques (une plus grande réussite en haut de la distribution se « paye » par une moindre réussite en bas de la distribution) : on modélisera en effet séparément le haut et le bas de la distribution des notes dans le lycée sans supposer que les effets sont symétriques. La comparaison des effets en haut et en bas de la distribution permet aussi d'estimer dans quelle mesure certains établissements peuvent amplifier ou au contraire réduire la dispersion des notes de leurs élèves par rapport à ce qui est attendu étant donné leurs caractéristiques. Il s'agit donc d'observer si certains lycées peuvent obtenir des résultats plus homogènes, ou à l'inverse plus inégalitaires, que des lycées dans lesquels les caractéristiques initiales des élèves (y compris en termes de niveau scolaire à la fin du collège) sont proches.

La méthode statistique utilisée ici s'apparente donc à une hybridation des modèles de SGP et des modèles de valeur ajoutée. Comme les premiers, elle repose sur la modélisation des effets lycée sur la distribution des notes, à partir de régressions quantiles mais, comme les seconds, en tenant compte de l'ensemble des caractéristiques observables des élèves, notamment leur niveau initial et leur origine sociale, pour essayer de réduire, autant que faire se peut, les effets de sélection.

Pour estimer les effets spécifiques à chaque lycée, au-delà des effets liés à la composition initiale, des indicatrices sont introduites dans le modèle pour chaque lycée, avec une condition de normalisation. Cette modélisation, appelée classiquement dans la littérature économétrique « à effets fixes », a l'avantage de demander très peu d'hypothèses, d'une part sur la distribution de ces effets fixes (ils peuvent être très différents d'un lycée à un autre, sans spécifier une forme particulière sur ces différences), et d'autre part sur les liens éventuels de ces effets lycées et des caractéristiques des élèves dont on souhaite mesurer l'effet. Plus précisément, il est possible d'estimer sans biais les effets des caractéristiques des élèves sur les résultats au baccalauréat, même si la répartition des élèves dans les lycées dépend à la fois de ces caractéristiques (par exemple, leur niveau scolaire) et des caractéristiques inobservées des lycées. Un risque de ce type de modèle est que les effets peuvent être mal estimés lorsque le nombre d'élèves par établissements est faible[4] : c'est

4. Ce problème est particulièrement crucial lorsqu'on modélise des variables non continues (par exemple si la variable d'intérêt est la réussite au baccalauréat plutôt que la note moyenne au baccalauréat), car la mauvaise approximation des effets fixes « contamine » l'estimation des coefficients correspondant aux caractéristiques individuelles des élèves.

la raison pour laquelle on se restreint ici aux lycées avec un nombre « suffisant » d'élèves (au moins 65 pour la filière générale et 25 pour la filière technologique, ces seuils ayant été choisis en arbitrant pour ne pas trop restreindre l'échantillon – et donc sa représentativité – tout en réduisant le risque d'obtenir des estimateurs biaisés).

Notons que les effets fixes lycées captent toutes les caractéristiques des lycées ; il n'est donc pas possible d'estimer de manière isolée l'effet de telle ou telle caractéristique (par exemple, l'ancienneté des enseignants, ou le niveau moyen des autres élèves). De plus, l'effet de ces variables est en général très difficile à estimer dès lors que des effets de sélection existent (par exemple, si les enseignants les plus expérimentés sont plus souvent affectés dans les lycées avec les élèves les plus favorisés, ou si les élèves ont tendance à se regrouper par niveau). Les modèles dits « à effets aléatoires », qui consistent à utiliser une distribution spécifique (généralement la loi normale) pour modéliser les effets propres des établissements, permettent d'estimer en même temps des coefficients pour les variables au niveau des lycées et des effets pour chaque lycée (c'est par exemple cette modélisation qui est retenue par Page *et al*., 2017). Cependant, lorsqu'il y a un lien entre les caractéristiques des élèves et les effets des lycées, les coefficients estimés sont susceptibles d'être biaisés (pour une discussion générale sur ces types de modèle dans le cadre des données utilisées ici, voir par exemple

Encadré 2 – **La régression quantile et la mesure des effets lycées**

La régression quantile est une méthode statistique de régression linéaire permettant de décrire comment une variable d'intérêt varie en fonction de covariables (pour une présentation détaillée, voir par exemple D'Haultfoeuille & Givord, 2014). Alors que la régression linéaire classique modélise comment la moyenne de la variable d'intérêt varie en fonction de variables observables, la régression quantile consiste à modéliser le quantile de cette variable conditionnel à ces observables, ces deux méthodes reposant sur une approximation linéaire. Pour le quantile $q_\tau(X)$ d'ordre τ (compris entre 0 et 1) de la distribution de la variable d'intérêt Y conditionnelle aux covariables X, on suppose donc :

$$q_\tau(X) = X\beta_\tau$$

où β_τ est le coefficient à estimer.

Il est donc possible, à partir de plusieurs régressions quantiles, d'enrichir la description de la manière dont une variable dépend d'une autre, en modélisant l'effet des covariables pour plusieurs percentiles – comme par exemple le premier décile, la médiane et le dernier décile.

En pratique, il a été montré que ce coefficient peut être obtenu sans biais par le programme linéaire :

$$\beta_\tau = argmin_\beta E\left(\rho_\tau(Y - X\beta_\tau)\right)$$

où la fonction $\rho_\tau(.)$ est une fonction test définie par $\rho_\tau(u) = (\tau - 1_{u<0})u$.

Le coefficient $\beta_{j\tau}$ pour une variable X_j correspond à la manière dont le quantile d'ordre τ de la variable d'intérêt varie en fonction d'une variation de la variable X_j. Si X_j est une variable continue, le coefficient $\beta_{j\tau}$ s'interprète comme la manière dont varie le quantile en fonction d'une variation marginale de X_j.

Parfois, une approximation linéaire peut être une forme trop simplifiée de la relation réelle entre la variable d'intérêt et les variables observables en fonction des données et, comme pour une relation linéaire, il est possible d'utiliser une forme plus complexe, soit par une transformation (par exemple, le logarithme de la variable considérée), soit par une forme polynomiale dans la variable X_j. Ici, pour approcher la relation entre les notes au baccalauréat et le niveau obtenu au brevet des collèges, il est apparu plus adapté aux données d'utiliser une forme quadratique. En pratique, cela signifie que pour interpréter l'effet de la note au brevet sur la distribution des notes au baccalauréat, il faut tenir compte des deux coefficients correspondant respectivement au niveau de la note moyenne au baccalauréat $\beta_{B\tau}$ et à son carré $\beta_{B^2\tau}$. Par exemple, pour τ = 0.80, le dernier quintile pour les élèves qui ont eu la note N au brevet (le niveau obtenu au plus par 20 % des élèves avec ce niveau initial) est inférieur de $\beta_{B\tau} + \beta_{B^2\tau} * (2N + 1)$ au quintile de la distribution des notes au baccalauréat pour les élèves ayant obtenu une note de N+1 au brevet. Ici, ces deux coefficients sont positifs pour les trois quantiles observés. Cela rend compte de l'intuition qu'une note élevée au brevet est un prédicteur de bons résultats au baccalauréat, mais signifie également que les très bons élèves au brevet peuvent encore plus marquer la différence au baccalauréat.

Il est également possible de comparer les effets d'une même variable sur les différents quantiles. Le fait qu'une variable X ait un effet plus fort sur un quantile faible que sur un quantile élevé de la distribution des notes peut s'interpréter en termes de dispersion des notes au baccalauréat. Par exemple, le fait que le coefficient de l'indicatrice qui indique qu'un élève est une fille est plus élevé pour le premier quintile que pour le dernier quintile signifie que les filles réussissent mieux, mais avec des notes moins dispersées que les garçons.

C'est ce type de comparaison qui est utilisée pour interpréter les effets lycées.

Givord & Guillerm, 2016)[5]. C'est pour cette raison qu'on préfère ici utiliser une modélisation à effets fixes.

Les régressions quantiles sont utilisées pour estimer des effets fixes des lycées sur le niveau des élèves les plus faibles (les 20 % d'élèves dont la note est inférieure à la valeur du premier quintile de la distribution des notes dans le lycée) et sur le niveau des élèves les plus forts (les 20 % d'élèves dont la note est supérieure à la valeur du dernier quintile de la distribution des notes dans le lycée), en tenant compte de la composition notamment en termes de niveau scolaire initial et de milieu social des élèves.

3. Une application à partir des résultats du baccalauréat 2015

3.1. Les données

Nous nous appuyons sur la base de données exhaustive des résultats à l'examen national du baccalauréat de la session 2015. Cette base fournit l'ensemble des notes obtenues aux différentes épreuves, mais on utilise ici la moyenne des notes des différentes matières (pondérées par leurs coefficients dans la série choisie) obtenue à la première session de l'examen[6]. Ces résultats sont complétés par le système d'information des Fichiers anonymisés pour les études et la recherche (Faere) produits et mis à disposition par la Depp. Cette base de données, constituée à des fins d'études à partir des fichiers administratifs de suivi scolaire des élèves, contient des informations individuelles telles que le sexe de l'élève, la catégorie socioprofessionnelle des parents et l'âge, ainsi que les établissements scolaires fréquentés. Elle contient également les résultats individuels à l'examen national du brevet (DNB), qui sont un indicateur du niveau scolaire de l'élève au moment de l'entrée au lycée.

Le tableau 1 permet d'illustrer les forts effets de composition dont la modélisation vise à tenir compte. Il présente des caractéristiques moyennes des établissements, estimées sur trois groupes distincts d'établissements définis en fonction des résultats obtenus en moyenne par leurs élèves au baccalauréat. Ils distinguent les 20 % des lycées dont les résultats à l'examen sont les plus faibles (soit 352 lycées généraux et 310 lycées technologiques), les 20 % des lycées dont les résultats sont les plus élevés, le troisième groupe étant constitué des lycées entre ces deux groupes extrêmes (soit 1 055 lycées généraux et 929 lycées technologiques). Ce classement porte sur les notes moyennes observées par établissement : ainsi, alors que la note moyenne sur l'ensemble des lycées généraux est de 12.2/20, elle n'est que de 10.5 pour le groupe des lycées avec les résultats les plus faibles, de 12.2 pour le groupe moyen, et de 13.8 pour les lycées du dernier groupe. Le premier constat est que, en moyenne, les lycées reproduisent en très grande partie le niveau de leurs élèves à la sortie du collège, notamment dans la filière générale. Le niveau moyen au DNB des lycéens passant le baccalauréat technologique était en moyenne plus faible, mais on retrouve également ce gradient dans l'autre sens : les lycées qui affichent les meilleurs résultats au baccalauréat sont aussi ceux qui scolarisent plus souvent les élèves qui étaient les meilleurs à la sortie du collège. Ces différences de performance peuvent être aussi reliées au niveau socio-économique des élèves, qui est l'un des déterminants les plus importants de la réussite scolaire, et dont on retrouve la hiérarchie ici. En outre, les 20 % des lycées dont les élèves ont obtenu en moyenne les résultats les plus élevés au baccalauréat scolarisent également des élèves aux profils scolaire et social plus homogènes, comme le montrent les variances de l'indice de position sociale des parents et des notes, plus faibles dans ce groupe que dans les deux autres groupes. Cela signifie notamment que ces lycées scolarisent moins souvent des élèves en difficulté, ce qui s'observe d'ailleurs dans la proportion d'élèves ayant redoublé au moins une année dans leur scolarité (dénommés redoublants dans le tableau 1 et par la suite). Dans la filière générale, les meilleurs lycées ne scolarisent que 3 % d'élèves redoublants, contre 11 % dans les lycées ayant les performances les plus faibles.

L'estimation des effets fixes par lycée permet d'évaluer les effets propres aux établissements

---

5. Il est possible d'obtenir des coefficients non biaisés de l'effet des caractéristiques individuelles des élèves, à condition d'ajouter les moyennes de ces caractéristiques, agrégées par lycée, dans le modèle (on parle de « correction de Mundlack »). Cependant, cette correction ne permet pas de corriger les biais éventuels sur les variables estimées au niveau des lycées. Ainsi, ajouter la moyenne du niveau de l'ensemble des élèves d'un lycée sur la note individuelle d'un élève permet d'estimer sans biais l'effet du niveau individuel sur la réussite, mais le coefficient obtenu pour la moyenne ne peut pas être interprété de manière causale comme l'effet du niveau de ces pairs sur le niveau d'un élève (voir Castellano et al.*, 2014).*

6. Les notes à la première session de l'examen correspondent aux notes après sessions d'harmonisation sur la notation, mais avant les épreuves de rattrapage. Ces épreuves sont proposées aux élèves ayant obtenu une note moyenne comprise entre 8 et 10, pour leur offrir la possibilité de repasser un oral pour certaines épreuves et obtenir in fine *une moyenne supérieure à 10 nécessaire pour l'obtention du diplôme. Pour cette raison, la distribution des notes après la deuxième session est très irrégulière (Givord & Suarez-Castillo, 2019), avec un point d'accumulation important juste au-dessus de 10/20 (c'est également le cas dans une moindre mesure de la distribution des notes à la première session de l'examen) et un déficit de masse entre 8 et 10. De plus, retenir la note à la deuxième session revient à comparer les résultats des élèves sur deux échelles significativement distinctes, puisque les notes portent alors sur des épreuves qui ne sont pas identiques entre les élèves.*

Tableau 1 – **Caractéristiques initiales des lycées, par groupes de performance moyenne au baccalauréat**

	Filière générale				Filière technologique			
	Total	20 % les plus faibles	Groupe médian]20,80[	20 % les plus élevés	Total	20 % les plus faibles	Groupe médian]20,80[	20 % les plus élevés
Nombres de lycées	1 759	352	1 055	352	1 549	310	929	310
Moyenne des notes au baccalauréat (1re session)	12.3	10.9	12.2	13.7	11.6	10.5	11.6	12.7
Moyenne des notes au brevet	12.3	11.2	12.3	13.3	9.7	8.9	9.8	10.5
Moyenne de l'indice de position sociale[(a)]	120.7	107.9	120.6	133.6	105.3	95.9	105.5	114.0
Variance des notes au baccalauréat (1re session)	6.4	6.6	6.6	5.6	4.4	5.4	4.3	3.8
Variance des notes au brevet	4.3	4.7	4.4	3.6	3.1	3.4	3.1	3.0
Variance de l'indice de position sociale	1 048.7	1 144.6	1 085.6	842.1	975.0	975.1	981.1	956.5
Part d'élèves redoublants (%)	6	11	5	3	18	24	17	14
Part de lycées privés (%)	26	3	19	70	20	6	16	45

[(a)] Voir encadré 1.
Note : les lycées sont groupés par filière (générale et technologique) selon la moyenne des notes de leurs élèves à la première série du baccalauréat.
Source : MENJ-Depp, Fichiers anonymisés pour les études et la recherche (Faere).

sur la réussite des élèves qu'ils scolarisent, au-delà des effets de composition. Ces estimations sont effectuées séparément pour les filières générales et technologiques. Afin de réduire la variance des estimateurs obtenus, l'échantillon est restreint aux établissements scolarisant en 2015 au moins 65 élèves dans les filières générales et 25 élèves dans les filières technologiques. Le choix de ces seuils permet de conserver 95 % des élèves dans les deux filières et résulte d'un compromis. D'un côté, il s'agit de conserver suffisamment d'élèves par lycée pour que des élèves qui pourraient avoir des profils très atypiques n'aient pas un poids trop élevé dans l'estimation des effets lycées. De l'autre, il s'agit de conserver un échantillon total d'élèves suffisamment grand pour ne pas réduire la généralisation des résultats, ce qui pourrait être le cas par exemple si les élèves scolarisés dans des lycées de grande taille ont des profils différents de ceux qui sont scolarisés dans des lycées de plus petite taille. On trouvera des détails sur la manière dont ces notes sont utilisées dans Givord & Suarez Castillo (2019). Les résultats individuels au baccalauréat sont régressés sur les caractéristiques individuelles observables des élèves : fait d'être une fille ou un garçon, origine sociale[7], fait d'avoir redoublé au cours de sa scolarité et résultats aux examens terminaux du brevet (avec une spécification quadratique), ainsi que sur un effet fixe pour tous les élèves d'un même lycée. L'effet de ces variables est estimé à trois niveaux de la distribution des notes au baccalauréat : premier et dernier quintiles et médiane. Les estimations portent sur les lycées généraux et technologiques, en séparant les deux filières. Des effets spécifiques à chaque série (trois dans la voie générale, huit dans la voie technologique) sont également ajoutés. Ils permettent de tenir compte du fait que les pratiques de notation diffèrent entre les différentes disciplines, dont le poids varie d'une série à l'autre.

3.2. La note au brevet est la variable la plus corrélée aux résultats au baccalauréat

Les corrélations entre les variables estimées et les autres variables sont conformes aux résultats obtenus classiquement (tableau 2). Comme déjà souligné par Evain & Evrard (2017) sur des données similaires, la moyenne des notes au baccalauréat apparaît très corrélée à celle des notes au brevet. Ici, les estimations suggèrent que cette dépendance s'observe à tous les niveaux de la distribution, et aussi que cette dépendance est non linéaire : le terme quadratique est positif pour les trois déciles étudiés (cf. encadré 2). Ce résultat peut s'expliquer par le fait que les très bons élèves à la fin du collège ont, très majoritairement, de très bons résultats tandis que les élèves qui ont des résultats plus faibles au brevet peuvent avoir des résultats plus variables.

Concernant l'impact du redoublement, la distribution conditionnelle des résultats au baccalauréat des élèves ayant redoublé est nettement inférieure à celle des non redoublants,

7. Telle que capturée par l'Indice de position sociale de la Depp (cf. encadré 1).

Tableau 2 – **Impact des variables explicatives sur la distribution de notes moyennes au baccalauréat (avec effets fixes lycées)**

	Q20		Q50		Q80	
	Coeff.	Écart-type	Coeff.	Écart-type	Coeff.	Écart-type
Filière générale (*N*=318 222)						
Note moyenne au brevet (niveau)	0.593***	(0.002)	0.632***	(0.002)	0.646***	(0.002)
Note moyenne au brevet (carré)	0.107***	(0.001)	0.105***	(0.001)	0.082***	(0.001)
Indice de position sociale	0.079***	(0.002)	0.079***	(0.002)	0.079***	(0.002)
Redoublant (*réf. : non redoublant*)	-0.271***	(0.008)	-0.245***	(0.007)	-0.193***	(0.008)
Fille (*réf. : garçon*)	0.08***	(0.004)	0.052***	(0.003)	0.032***	(0.004)
Série L (*réf. : ES*)	0.074***	(0.005)	0.086***	(0.005)	0.088***	(0.006)
Série S	-0.194***	(0.004)	-0.172***	(0.004)	-0.147***	(0.004)
Filière technologique[a] (*N*=122 286)						
Note moyenne au brevet (niveau)	0.358***	(0.004)	0.392***	(0.003)	0.408***	(0.004)
Note moyenne au brevet (carré)	0.018***	(0.002)	0.025***	(0.002)	0.034***	(0.002)
Indice de position sociale	0.034***	(0.004)	0.027***	(0.003)	0.027***	(0.004)
Redoublant (*réf. : non redoublant*)	-0.285***	(0.010)	-0.258***	(0.007)	-0.228***	(0.008)
Fille (*réf. : garçon*)	0.241***	(0.007)	0.211***	(0.007)	0.189***	(0.008)
ST2S (*réf. : STMG*)	-0.155***	(0.011)	-0.168***	(0.010)	-0.165***	(0.013)
STD2A	0.002	(0.036)	0.012	(0.027)	0.056**	(0.032)
STI2D	0.010	(0.014)	0.067***	(0.010)	0.168***	(0.013)
STL	0.140***	(0.020)	0.207***	(0.015)	0.261***	(0.020)
HOT	-0.360***	(0.054)	-0.397***	(0.040)	-0.456***	(0.049)

[a] Les séries de la filière technologique sont celles de 2015, soit : Sciences et technologies du management et de la gestion (STMG), de la santé et du social (ST2S), de laboratoire (STL), de l'industrie et du développement durable (STI2D), du design et des arts appliqués (STD2A), de l'hôtellerie (HOT).
Note : effets des variables explicatives sur les résultats au baccalauréat (moyenne de l'ensemble des notes) estimés par régressions quantiles pour le premier quintile (Q20), la médiane (Q50) et le dernier quintile (Q80) ; *** significatif à 1 %, ** significatif à 5 %.
Source : MENJ-Depp, fichiers Faere.

l'écart étant plus important dans le bas de la distribution. Les filles obtiennent généralement de meilleurs résultats que les garçons, et moins dispersés, comme l'illustre le fait que l'effet « fille » est plus fort dans le bas que dans le haut de la distribution. À l'inverse des autres variables explicatives étudiées ici, l'origine sociale (captée par l'indice de position sociale des parents) a un effet presque identique aux trois niveaux étudiés de la distribution des notes au baccalauréat. Par ailleurs, il s'agit aussi de la seule variable du modèle dont la corrélation avec les notes au baccalauréat est très nettement diminuée lorsque l'on introduit les effets fixes spécifiques au lycée, comme le suggère la comparaison avec des estimations qui ne comprennent pas ces effets fixes (voir Givord & Suarez Castillo, 2019). Cet effet statistique souligne les fortes différences de recrutement social entre les lycées.

Enfin, on observe de forts écarts dans la distribution des notes entre filières. Ces écarts peuvent s'expliquer par des différences de notation selon les disciplines dominantes des filières, ainsi que par des effets de composition. On observe ainsi que les élèves de la série S obtiennent en moyenne des notes au baccalauréat inférieures à celles obtenues par les élèves des deux autres séries de la filière générale, une fois tenu compte de leur niveau initial et de leurs autres caractéristiques individuelles[8].

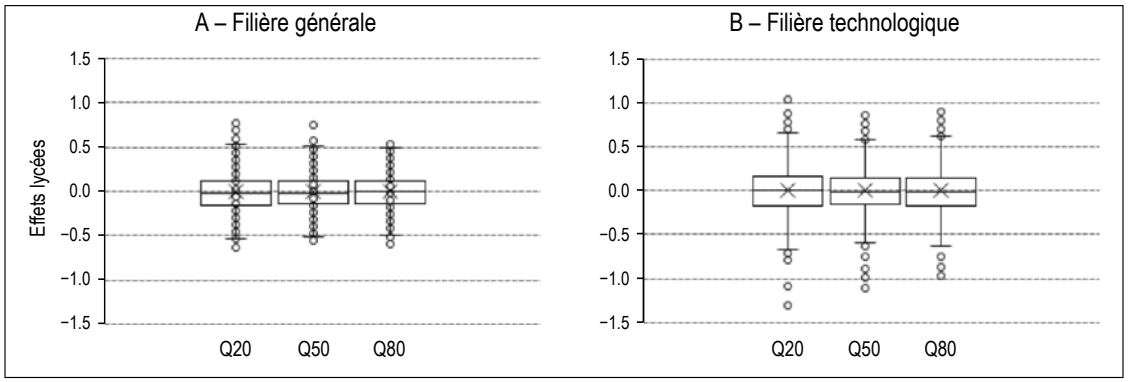
3.3. Des effets établissements très dispersés

Les effets fixes spécifiques aux lycées permettent par ailleurs de capter les effets établissements. Néanmoins, il est nécessaire de poser une contrainte d'identification – dans un modèle linéaire, il n'est pas possible d'estimer séparément la constante et les coefficients pour l'ensemble des lycées. On impose par convention que la moyenne des coefficients des lycées est nulle, ce qui signifie que pour chaque lycée, l'effet fixe estimé correspond à un écart de ce lycée par rapport à la moyenne des effets observés pour l'ensemble des lycées.

La dispersion de ces effets fixes des lycées est un peu plus élevée dans la filière technologique que dans la filière générale (figure II). Cela

8. Ce résultat suggère qu'il pourrait être utile d'interagir chaque variable individuelle par série, pour tenir compte des différences d'épreuves par séries, et introduire des attentes différenciées par séries suivant les caractéristiques des élèves. Ce choix n'a pas été fait ici, dans la mesure où cela augmente beaucoup le nombre de coefficients à estimer, alors même que le nombre d'élèves par série dans chaque lycée peut être faible et qu'il existe donc un risque de « sur-ajustement » des modèles, avec également des conséquences sur l'estimation des effets fixes des lycées. Estimer ce type de modèle serait judicieux en utilisant plusieurs années consécutives (ce qui n'a pas été possible avec les données disponibles pour cette étude).

Figure II – **Caractéristiques des distributions des effets fixes lycées**

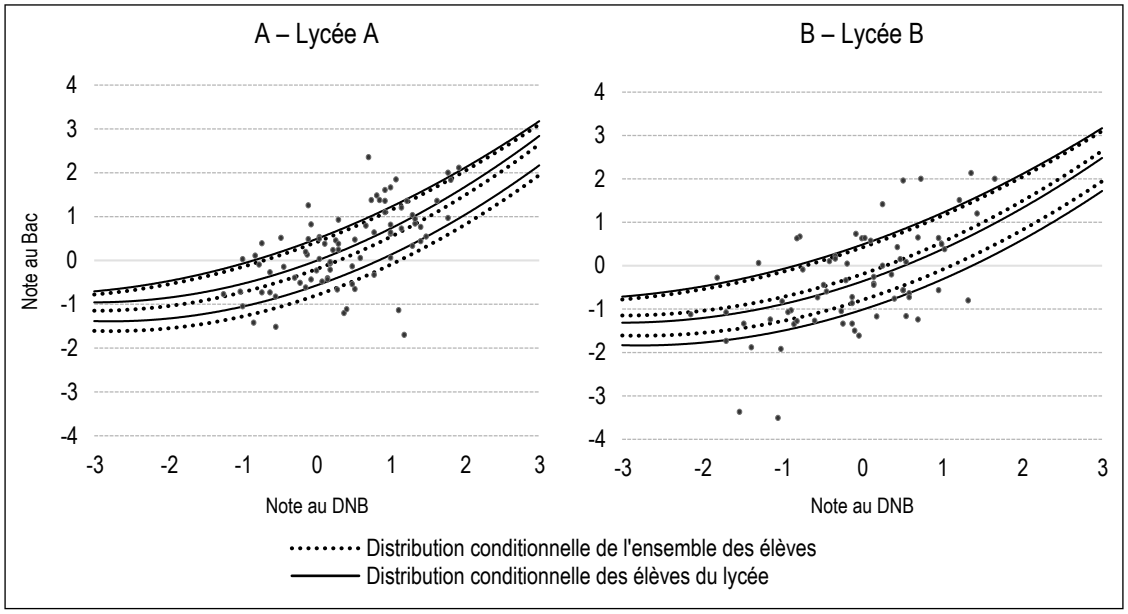


Note : effets fixes lycées obtenus par régressions quantiles pour le premier quintile (Q20), la médiane (Q50) et le dernier quintile (Q80).
Source : MENJ-Depp, fichiers Faere.

s'explique au moins en partie par le fait que pour les lycées technologiques, les effets fixes sont parfois estimés avec moins d'élèves, et donc les estimations sont moins précises. Dans les deux filières, on observe également que la dispersion est légèrement plus élevée pour les effets des lycées dans le bas de la distribution (au niveau du premier quintile) que dans le haut (au niveau du dernier quintile), avec des valeurs extrêmes nettement distinctes de la moyenne.

La figure III illustre le cas de deux lycées. Elle représente, pour chaque établissement, la relation estimée par les régressions quantiles entre les notes moyennes au baccalauréat et les notes au brevet (chaque observation correspond à un élève) pour les trois quantiles étudiés. Les courbes pleines représentent les estimations tenant compte de l'effet fixe de l'établissement – elles correspondent à une partition des élèves du lycée en fonction du niveau de la distribution auquel on s'intéresse. La courbe la plus basse correspond au premier quintile, elle est donc telle que 20 % des élèves du lycée se trouvent en dessous et 80 % au-dessus. De manière similaire, les deux autres courbes pleines constituent une partition telles que respectivement 50 % (pour la médiane) et 80 % (pour le quintile le plus élevé) des élèves du lycée se trouvent en dessous. Les courbes en pointillés sont les équivalentes

Figure III – **Notes au brevet et au baccalauréat dans deux lycées et estimations obtenues et prédites par les régressions quantiles**



Note : effets fixes lycées obtenus par régressions quantiles pour le premier quintile (Q20), la médiane (Q50) et le dernier quintile (Q80). Les courbes en pointillés correspondent aux courbes quantiles (respectivement pour Q20, Q50 et Q80) obtenues par les régressions estimées pour l'ensemble des élèves de la filière générale (par exemple 20 % des points de l'échantillon se trouvent en dessous de la courbe Q20 en traits pointillés). Les courbes en plein correspondent aux résultats de ces estimations en ajoutant les effets fixes spécifiques au lycée considéré (par exemple, 20 % des élèves du lycée A sont en dessous de la courbe Q20 en trait plein).
Source : MENJ-Depp, Fichiers Faere.

des précédentes, mais sans tenir compte des effets fixes des lycées – c'est-à-dire qu'elles correspondent aux effets attendus, d'après les corrélations observées dans l'ensemble des élèves ayant passé le baccalauréat dans cette filière.

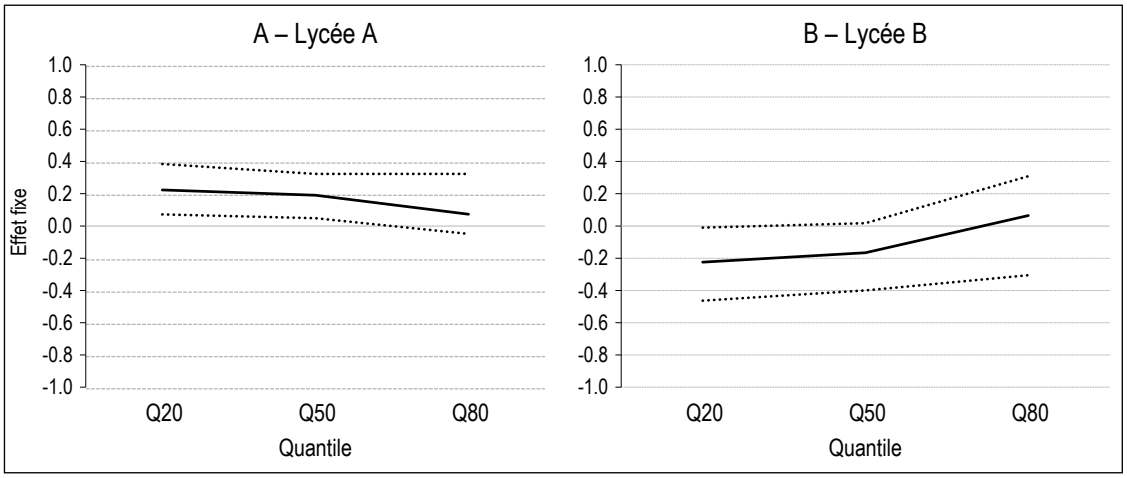
Dans les deux cas illustrés ici, les meilleurs élèves de chaque lycée étudié ne font pas moins bien que ce qui est attendu (la courbe du quintile le plus élevé se trouve légèrement au-dessus de la courbe pointillée correspondante, les différences n'étant pas significatives comme discuté plus bas). Néanmoins, les résultats des élèves de ces deux lycées sont très différents pour les autres niveaux de la distribution. Dans le lycée A, la médiane comme le quintile le plus bas sont nettement plus élevés, c'est-à-dire que ce lycée parvient à faire réussir au moins 80 % de ses élèves au-delà que ce qui est attendu ; cela signifie que ce lycée obtient des résultats supérieurs à la moyenne, sans sacrifier certains élèves. Au contraire, le lycée B parvient à faire réussir légèrement mieux qu'attendu les 20 % des meilleurs élèves mais les 20 % les plus faibles font nettement moins bien qu'attendu. À l'inverse de l'exemple précédent, ce lycée a non seulement des résultats moins bons au niveau de la médiane, mais il a également tendance à amplifier les écarts de performance, par rapport à ce qui était attendu.

Ces faits stylisés sont synthétisés dans la figure IV, qui représente les effets fixes estimés pour le premier quintile, la médiane et le dernier quintile dans les deux lycées. Pour le lycée A, tous les coefficients sont positifs, même si le coefficient correspondant au dernier quintile n'est pas significatif. Dans le lycée B, seul le coefficient correspondant au dernier quintile est positif (mais non significatif), tandis que les autres sont négatifs. Les effets sont décroissants dans le lycée A, ce qui signifie aussi que les écarts sont plus réduits qu'attendu dans ce lycée, et ils sont à l'inverse croissants dans le lycée B, ce qui signifie que les écarts y sont plus forts qu'attendu.

Cette analyse peut être faite de manière plus systématique : plus précisément, il est possible, pour chaque lycée, de comparer son effet spécifique tel qu'estimé par le modèle au niveau du dernier et du premier quintile de la distribution des notes conditionnelles. Il est alors possible de tester si la différence est significativement positive, indiquant que ce lycée a tendance à augmenter les inégalités de performance, à composition initiale donnée, ou à l'inverse, si elle est significativement négative, indiquant qu'il tend à réduire les inégalités de performance entre ses élèves. Ces tests doivent néanmoins prendre en compte le fait que l'utilisation répétée de tests statistiques (sur l'ensemble des lycées) peut conduire à accepter trop souvent des écarts significativement non nuls (pour une discussion approfondie, voir Givord & Suarez Castillo, 2019). Si, pour la majorité des lycées, on n'observe pas d'écart statistiquement différent de zéro, 8.2 % des lycées généraux, et 6 % des lycées technologiques ont tendance à augmenter significativement les écarts de performance entre leurs élèves, tandis qu'à l'inverse 8.5 % des lycées généraux, et 7.6 % des lycées technologiques ont tendance à les réduire.

La réduction de la dispersion des résultats des élèves n'est cependant pas un objectif en soi. Elle n'est pas souhaitable si elle doit aboutir

Figure IV – **Estimations des effets fixes dans deux lycées à trois niveaux de la distribution des notes**



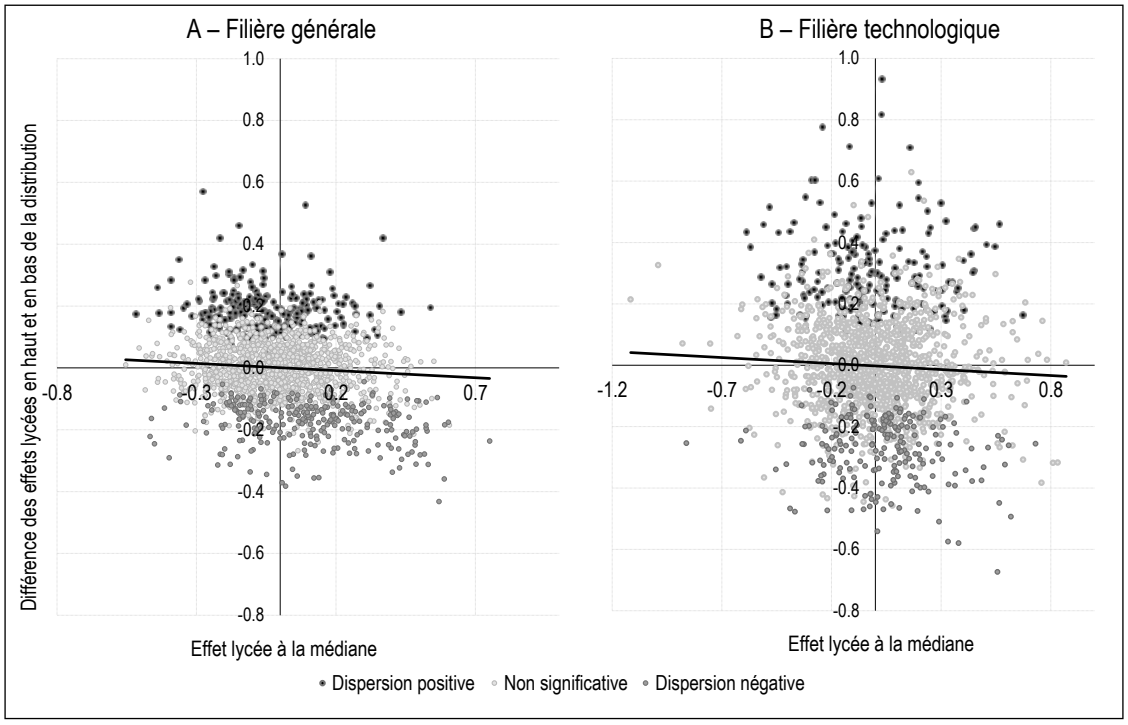
Note : coefficients des effets lycées par des régressions quantiles. Les courbes pointillées correspondent aux intervalles de confiance à 95 %.
Source : MENJ-Depp, fichiers Faere.

à un nivellement par le bas, c'est-à-dire que les résultats moins inégaux entre les élèves se font au détriment des exigences. Le cas du lycée A illustre qu'il est possible d'observer à la fois une meilleure performance et de moins grandes inégalités, Pour évaluer si ce phénomène s'observe de manière plus générale, il est possible de comparer l'effet propre de chaque lycée au niveau de la médiane (qui correspond à une approximation de sa valeur ajoutée moyenne), avec l'écart entre les effets propres mesurés respectivement au niveau du premier et du dernier quintile, qui correspond à une mesure de l'effet de ce lycée sur la dispersion des notes.

Cette relation est illustrée, séparément pour les deux filières générales et technologiques, dans la figure V. Dans cette figure, chaque point correspond à un lycée. L'abscisse correspond à l'effet lycée estimé au niveau de la médiane : une valeur positive signifie que ce lycée tend à améliorer la performance d'au moins la moitié de ces élèves, tandis qu'à l'inverse une valeur négative signifie qu'il tend à détériorer la performance pour la majorité des élèves. L'ordonnée correspond à l'écart entre les coefficients estimés pour le dernier et le premier quintile. Une valeur positive est associée à des résultats plus dispersés dans ce lycée que ce qui est attendu, ce qui signifie que ce lycée a tendance à augmenter les inégalités de performance, à composition initiale donnée, et une valeur négative signifie que le lycée tend à réduire les inégalités de performance de ces élèves.

Le premier enseignement de cette représentation est que le lycée A n'est pas un cas isolé. Sur l'ensemble des lycées observés, une augmentation de la performance est négativement corrélée avec l'augmentation des inégalités de performance. De nombreux lycées parviennent ainsi à faire réussir leurs élèves sans sacrifier les plus faibles. La pente est plus forte dans la filière générale que dans la filière technologique, cependant. Par ailleurs, cette relation entre efficacité et égalité est loin d'être déterministe. Si les lycées égalitaires, ceux qui parviennent à réduire les écarts de performance entre leurs élèves, sont plus souvent aussi performants, au sens où ils parviennent à augmenter la performance moyenne, une majorité d'élèves ont une performance moins bonne qu'attendu dans d'autres. De même, si les lycées inégalitaires (dans lesquels la dispersion des résultats entre leurs élèves est plus élevée qu'attendu) sont plus souvent moins performants que la moyenne, certains d'entre eux sont également classés parmi ceux parvenant à augmenter la performance de leurs élèves.

Figure V – **Effets lycées sur la dispersion et la médiane par filière**



Note : effets fixes obtenus par régressions quantiles aux premier et dernier quintiles et à la médiane.
Lecture : chaque point correspond à un lycée, avec en abscisse l'effet fixe estimé à la médiane, et en ordonnée la différence des effets fixes de ce lycée au dernier et au premier quintile. La droite correspond à une régression de régression linéaire.
Source : MENJ-Depp, fichiers Faere.

* *
*

L'évaluation des établissements scolaires est devenue une question centrale dans le débat public. Comme discuté dans la revue de la littérature, cette question est d'autant plus compliquée que la qualité d'un établissement est forcément multidimensionnelle, et qu'il ne peut être jugé sur un simple indicateur : c'est d'ailleurs pour cette raison que les IVAL produits par la Depp s'intéressent à plusieurs dimensions (la réussite au baccalauréat, mais également les taux de rétention compris comme la capacité des lycées à accompagner leurs élèves sur l'ensemble de leur scolarité). Cet article enrichit encore cette description, en illustrant dans quelle mesure les indicateurs focalisés sur la seule moyenne peuvent refléter aussi bien la capacité d'un établissement à faire progresser l'ensemble de ses élèves qu'une concentration sur certains.

Les résultats obtenus suggèrent que si, pour la majorité des lycées, il n'est pas possible de mettre en évidence statistiquement des effets hétérogènes (les écarts observés sont du même ordre statistique que ceux attendus), à l'inverse environ un sixième d'entre eux ont tendance soit à amplifier, soit à réduire les écarts entre les résultats de leurs élèves. Contrairement à l'opinion parfois exprimée, les lycées « inclusifs », qui parviennent à réduire les écarts de performances de l'ensemble des élèves, ne le font pas en nivelant les résultats de tous vers le bas. Au contraire, ces lycées semblent surreprésentés parmi les établissements qui parviennent à obtenir des résultats supérieurs à ce qui serait attendu au niveau de la médiane. L'interprétation de ces résultats appelle plusieurs remarques.

La première est que les effets lycées sont par nature estimés sur de petits effectifs, et sont donc imprécis. Il est alors difficile d'isoler l'exceptionnel (par exemple quelques élèves très brillants, un accident survenu dans l'établissement qui aurait perturbé la scolarité, etc.) de ce qui relève des fondamentaux du lycée (les projets d'établissements, le climat scolaire, la cohésion de l'équipe enseignante, etc.). Le risque existe de surinterpréter des écarts à la moyenne qui ne correspondraient qu'à des accidents statistiques. Pour vérifier la robustesse des résultats obtenus et réduire la volatilité des estimations, il serait intéressant de comparer les estimations obtenues pour un même lycée d'une année sur l'autre, ou d'estimer ces effets en utilisant plusieurs années quand elles sont disponibles (pour cette étude, les données pour une seule année ont pu être utilisées), comme suggéré par Bitler *et al*. (2019).

Une autre difficulté pour évaluer ces effets tient au fait qu'ils reposent sur l'hypothèse que tous les élèves qui passent le baccalauréat ont effectué l'ensemble de leur scolarité au lycée dans le même établissement. Or cette hypothèse n'est pas toujours vérifiée : certains élèves peuvent déménager en cours de scolarité, ou changer de lycée pour suivre une série qui ne serait pas proposée dans l'établissement dans lequel ils sont scolarisés en classe de seconde. Ces changements d'établissement ne sont pas seulement le fait des élèves – certains lycées peuvent choisir de ne pas accepter des élèves dont les chances de réussir l'examen sont trop faibles, par exemple en refusant l'inscription dans une série proposée, ou un redoublement. Ces comportements stratégiques des établissements peuvent fausser les indicateurs de performance liés aux résultats du baccalauréat. En excluant les élèves qui ont les résultats les plus faibles, ils peuvent conduire à surestimer la valeur ajoutée des lycées, mais également à réduire la dispersion des résultats – et donc les faire paraître plus égalitaristes qu'ils ne sont (pour une discussion plus approfondie, voir Givord & Suarez Castillo, 2019). Comme discuté plus haut, ces effets peuvent être d'autant plus importants que l'évaluation des établissements devient un enjeu pour les acteurs[9].

Traiter complètement cette question demanderait de disposer de mesures des niveaux des élèves plus fréquentes, notamment pour évaluer les progressions des élèves d'une année sur l'autre. Il faudrait aussi s'intéresser à d'autres indicateurs sur le parcours des élèves : c'est ce que permettent les indicateurs produits par la Depp en plus des indicateurs portant sur le taux de réussite au baccalauréat, qui informent sur les taux d'accès au baccalauréat depuis la classe de

9. Dans un autre registre, on peut rapprocher ce point du fait que les différentes expériences de prime à la performance pour les enseignants ne fournissent pas de résultats toujours probants en termes de progression des élèves. On trouvera dans Imberman (2015) une recension de la littérature économique sur ce sujet. Si des expériences ont montré l'efficacité des primes à la performance dans certains pays en développement, notamment en Inde (Muralidharan & Sundaraman, 2011) et en Tanzanie (Mbiti et al.*, 2019), avec des résultats plus ambigus au Kenya (Glewwe* et al.*, 2010), les différentes expérimentations menées notamment aux États-Unis fournissent des résultats qui ne permettent pas d'arriver à un consensus sur leur efficacité (Dee & Wyckoff, 2015 ; Fryer, 2013 ; Springer* et al.*, 2016). Les différentes raisons avancées pour expliquer des conséquences minimes voire négatives sur la progression des élèves sont que le montant des primes peut être trop modeste pour avoir un réel impact, le fait que les bonus financiers n'ont pas d'effet direct sur la motivation des enseignants ou encore qu'ils conduisent ces derniers à se concentrer uniquement sur les disciplines et les formats des tests standardisés utilisés pour l'évaluation. Ces résultats suggèrent que les effets de politique d'incitation aux résultats sont très sensibles aux détails de leur mise en place (Goodman & Turner, 2013), et notamment que l'évaluation des enseignants doit se faire sur plusieurs critères et ne pas reposer uniquement sur des mesures quantitatives.*

seconde, première et terminale et donc potentiellement sur ces mécanismes de sélection en cours de scolarité. Cette question rappelle, comme discuté plus haut, que l'évaluation d'un lycée ne peut se faire sur une dimension unique et qu'il est indispensable de croiser différentes dimensions. Au-delà de la performance à l'examen du baccalauréat, on pourrait par exemple s'interroger sur le climat de l'établissement et le bien-être des élèves, ou leur insertion ultérieure dans l'enseignement supérieur et sur le marché du travail.

Enfin, une dernière question centrale porte sur l'utilisation *in fine* des indicateurs sur la mesure des effets établissements. Si ces mesures peuvent servir, dans les limites détaillées dans l'introduction, comme des outils de pilotage pour différents acteurs, leur appropriation par les familles, notamment dans les situations de choix scolaire, reste à questionner. De fait, des travaux sur la ville de New-York, montrent que, dans des situations de choix scolaire, et même quand l'information sur les valeurs ajoutées des établissements est disponible, les familles ne semblent pas en tenir compte dans leur choix, mais privilégient avant tout les établissements scolarisant les meilleurs élèves (Abdulkadiroğlu *et al*., 2020). Il serait intéressant de s'interroger sur ce point dans le cas de la France, où un travail de communication important autour de la mesure des effets lycées existe depuis très longtemps. □

BIBLIOGRAPHIE

Abdulkadiroğlu, A., Pathak, P., Schellenberg, J. & Walters, C. (2020). Do Parents Value School Effectiveness? *American Economic Review*, 110 (5), 1502–1539. http://dx.doi.org/10.1257/aer.20172040

Barlevy, G. & Neal, D. (2002). Pay for Percentile. *American Economic Review*, 102(5), 1805–1831. http://dx.doi.org/10.1257/aer.102.5.1805

Betebenner, D. (2007). Estimation of Student Growth Percentiles for the Colorado Student. Technical report, National Center for the Improvement of Educational Assessment (NCIEA). https://www.researchgate.net/publication/228822935_Estimation_of_student_growth_percentiles_for_the_Colorado_Student_Assessment_Program

Bitler, M., Corcoran, S., Thusrton, D. & Penner, E. (2019). Teacher Effects on Student Achievement and Height: A Cautionary Tale. National Bureau of Economic Research, *Working paper* N° 26480. http://dx.doi.org/10.3386/w26480

Boutchenik, B. & Maillard, S. (2019). Élèves hétérogènes, pairs hétérogènes. *Éducation & Formations*, 100, 53–72. https://dx.doi.org/10.48464/halshs-02426355

Castellano, K. E., Rabe-Hesketh, S. & Skrondal, A. (2014). Composition, Context, and Endogeneity in School and Teacher Comparisons. *Journal of Educational and Behavioral Statistics*, 39(5), 333–367. http://dx.doi.org/10.3102/1076998614547576

Chetty, R., Friedman, J. & Rockoff, J. (2014). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, 104(9), 2633–2679. http://dx.doi.org/10.1257/aer.104.9.2633

D'Haultfoeuille, X. & Givord, P. (2014). La régression quantile en pratique. *Économie et Statistique*, 471, 85–111. http://dx.doi.org/10.3406/estat.2014.10484.

Dee, T. & Wyckoff, J. (2015). Incentives, Selection, and Teacher Performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, 34(2), 267–297. http://dx.doi.org/10.1002/pam.21818

Duclos, M. & Murat, F. (2014). Comment évaluer la performance des lycées. *Éducation & Formations*, 85, 72–84. https://archives-statistiques-depp.education.gouv.fr/Default/doc/SYRACUSE/10554/education-formations-n-85-novembre-2014-chap-5-comment-evaluer-la-performance-des-lycees-un-point-su

Ehlert, M., Koedel, C., Parsons, E. & Podgursky, M. J. (2014). The Sensitivity of Value-Added Estimates to Specification Adjustments: Evidence From School- and Teacher-Level Models in Missouri. *Statistics and Public Policy*, 1(1), 19–27. http://dx.doi.org/10.1080/2330443x.2013.856152

Evain, F. (2020). Indicateurs de valeur ajoutée des lycées : Du pilotage interne à la diffusion grand public. *Courrier des Statistiques*, 5, 74–94. https://www.insee.fr/fr/information/5008703?sommaire=5008710

Evain, F. & Evrard, L. (2017). Une meilleure mesure de la performance des lycées : Refonte de la méthodologie des IVAL (session 2015). *Éducation & Formations,* 94, 91–116. https://dx.doi.org/10.48464/halshs-01693896

Everson, K. (2016). Value-Added Modeling and Educational Accountability. *Review of Educational Research*, 87(1), 35–70. http://dx.doi.org/10.3102/0034654316637199

Felouzis, G. (2005). Performances et « valeur ajoutée » des lycées : le marché scolaire fait des différences. *Revue française de sociologie*, 46(1), 3–36. http://dx.doi.org/10.3917/rfs.461.0003

Fryer, R. (2013). Teacher Incentives and Student Achievement: Evidence from New York City Public Schools. *Journal of Labor Economics*, 31(2), 373–407. http://dx.doi.org/10.1086/667757.

Gamoran, A. (2012). Bilan et devenir de la loi *No Child Left Behind* aux États-Unis. *Revue française de pédagogie*, 178, 13–26. https://doi.org/10.4000/rfp.3509

Givord, P. & Guillerm, M. (2016). Les modèles multiniveaux. Insee, *Document de Travail* N° M2016/05. https://www.insee.fr/fr/statistiques/2022152

Givord, P. & Suarez Castillo, M. (2019). Excellence for all? Heterogeneity in high-schools' value-added. Insee, *Document de Travail* N° G2019/14. https://www.insee.fr/en/statistiques/4266034

Glewwe, P., Ilias, N. & Kremer, M. (2010). Teacher Incentives. *American Economic Journal: Applied Economics*, 2(3), 205–227. http://dx.doi.org/10.1257/app.2.3.205

Goodman, S. & Turner, L. (2013). The Design of Teacher Incentive Pay and Educational Outcomes: Evidence from the New York City Bonus Program. *Journal of Labor Economics*, 31(2), 409–420. http://dx.doi.org/10.1086/668676

Guarino, C., Reckase, M., Stacy, B. & Wooldridge, J. (2015a). A Comparison of Student Growth Percentile and Value-Added Models of Teacher Performance. *Statistics and Public Policy*, 2(1), 1–11. http://dx.doi.org/10.1080/2330443X.2015.1034820

Guarino, C., Reckase, M. & Wooldridge, J. (2015b). Can Value-Added Measures of Teacher Performance Be Trusted? *Education Finance and Policy*, 10(1), 117–156. http://dx.doi.org/10.1162/edfp_a_00153

Imberman, S. (2015). How effective are financial incentives for teachers? *IZA World of Labor*. http://dx.doi.org/10.15185/izawol.158

Jackson, C. (2018). What Do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes. *Journal of Political Economy*, 126(5), 2072–2107. http://dx.doi.org/10.1086/699018

Jacob, B. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, 89(5-6), 761–779. http://dx.doi.org/10.1016/j.jpubeco.2004.08.004

Koedel, C., Mihaly, K. & Rockoff, J. (2015). Value-added modeling: A review. *Economics of Education Review*, 47, 180–195. http://dx.doi.org/10.1016/j.econedurev.2015.01.006

Kraft, M. (2019). Teacher Effects on Complex Cognitive Skills and Social-Emotional Competencies. *Journal of Human Resources*, 54(1), 1–36. http://dx.doi.org/10.3368/jhr.54.1.0916.8265r3

Kurtz, M. (2018). Value-Added and Student Growth Percentile Models: What Drives Differences in Estimated Classroom Effects? *Statistics and Public Policy*, 5(1), 1–8. http://dx.doi.org/10.1080/2330443x.2018.1438938

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C. & Rajani, R. (2019). Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania. *The Quarterly Journal of Economics*, 134(3), 1627–1673. http://dx.doi.org/10.1093/qje/qjz010

Monso, O., Fougère, D., Givord, P. & Pirrus, C. (2019). Les camarades influencent-ils la réussite et le parcours des élèves ? Les effets de pairs dans l'enseignement primaire et secondaire. *Éducation & Formations*, 100, 23–52. https://www.education.gouv.fr/la-reussite-des-eleves-contextes-familiaux-sociaux-et-territoriaux-education-formations-ndeg-100-41657

Muralidharan, K. & Sundararaman, V. (2011). Teacher Performance Pay: Experimental Evidence from India. *Journal of Political Economy*, 119(1), 39–77. http://dx.doi.org/10.1086/659655

Page, G., San Martín, E., Orellana, J. & González, J. (2017). Exploring complete school effectiveness via quantile value added. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180, 315–340. https://doi.org/10.1111/rssa.12195

Raudenbush, S. W. & Wilms, J. D. (1995). The estimation of school Effects. *Journal of Educational and Behavorial Statistics*, 20(4), 307–335. https://doi.org/10.3102/10769986020004307

Rocher, T. (2016). Construction d'un indice de position sociale des élèves. *Éducation & Formations*, 90, 5–27. https://dx.doi.org/10.48464/hal-01350095

Rothstein, J. (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *Quarterly Journal of Economics*, 125(1), 175–214. http://dx.doi.org/10.1162/qjec.2010.125.1.175

Sass, T., Semykina, A. & Harris, D. (2014). Value-added models and the measurement of teacher productivity. *Economics of Education Review*, 38, 9–23. http://dx.doi.org/10.1016/j.econedurev.2013.10.003

Soland, J. (2016). Is Teacher Value Added a Matter of Scale? The Practical Consequences of Treating an Ordinal Scale as Interval for Estimation of Teacher Effects. *Applied Measurement in Education*, 30(1), 52–70. http://dx.doi.org/10.1080/08957347.2016.1247844

Springer, M., Swain, W. & Rodriguez, L. (2016). Effective Teacher Retention Bonuses. *Educational Evaluation and Policy Analysis*, 38(2), 199–221. http://dx.doi.org/10.3102/0162373715609687

Thélot, C. (1994a). Les arcanes de l'évaluation. *Courrier des statistiques,* 71-72, 3–6.

Thélot, C. (1994b). L'évaluation du système éducatif français. *Revue française de pédagogie*, 107, 5–28. https://doi.org/10.3406/rfp.1994.1261

Wall, D. (2000). The impact of high-stakes testing on teaching and learning: can this be predicted or controlled? *System*, 28(4), 499–509. http://dx.doi.org/10.1016/s0346-251x(00)00035-x

Walsh, E. & Isenberg, E. (2015). How Does Value Added Compare to Student Growth Percentiles? *Statistics and Public Policy*, 2(1), 1–13. http://dx.doi.org/10.1080/2330443x.2015.1034390