Évaluer l'action éducative des lycées à travers les indicateurs de valeur ajoutée des lycées : quand le " bruit " s'immisce dans l'administration de la preuve

Fernando Núñez-Regueiro, Pascal Bressoux

Évaluer l'action éducative des lycées à travers les indicateurs de valeur ajoutée des lycées : quand le « bruit » s'immisce dans l'administration de la preuve

Fernando NÚÑEZ-REGUEIRO
Pascal BRESSOUX

Résumé. Chaque année, le ministère de l'Éducation nationale publie des indicateurs de valeur ajoutée des lycées (IVAL) visant à mesurer l'efficacité des actions éducatives. Conçus comme outils de pilotage fiables, ces IVAL jouent un rôle décisionnel important dans le travail des équipes éducatives. Pourtant, les IVAL sont limités par la non-prise en compte d'une erreur d'échantillonnage qui confond l'action des lycées avec du « bruit » statistique (*i.e.*, les effets du hasard). Retraçant l'historique de ces indicateurs – depuis leurs origines dans la recherche sur les effets-établissement, jusqu'à leur diffusion en France –, cette étude montre en quoi cette non-prise en compte de l'erreur d'échantillonnage pose un problème de mesure qui contrarie une évaluation fiable de l'action éducative. Pour pallier cette difficulté, des stratégies d'estimation alternatives sont proposées. La pertinence de cette critique est illustrée à travers l'étude de 112 lycées de l'académie de Grenoble et de leur action éducative sur la réussite au baccalauréat.

Mots-clés. IVAL – VALEUR AJOUTEE – ÉVALUATION – LYCEE – DIPLOME – ANALYSE MULTINIVEAU

Les indicateurs de performance des établissements scolaires constituent des objets d'analyse controversés de la sociologie de l'éducation. Ils ont en effet pour objectif d'établir le rôle joué par un établissement dans l'explication de phénomènes éducatifs fondamentaux pour l'insertion sociale et professionnelle (e.g., acquisitions scolaires, accès à un diplôme), compte tenu des caractéristiques de ses élèves et de ses formations (Duclos et Murat, 2014 ; Perry, 2016 ; Raudenbush et Willms, 1995). D'une part, ces indicateurs ont un intérêt pour le pilotage en offrant la possibilité, pour les acteurs éducatifs, d'évaluer l'efficacité globale de leurs interventions au-delà des caractéristiques *a priori* favorables ou défavorables de leurs élèves, suivant la logique d'un raisonnement « toutes choses étant égales par ailleurs » sur ces caractéristiques. D'autre part, ces indicateurs sont difficiles à construire et à interpréter, ce qui a favorisé de nombreuses critiques à leur égard, notamment dans les pays où les établissements ont été évalués, récompensés ou pénalisés en fonction de leur niveau de performance sur ces indicateurs (e.g., en Angleterre, aux États-Unis, en Thaïlande ; Felouzis, 2005 ; Goldstein et Spiegelhalter, 1996 ; Raudenbush et Willms, 1995). Du fait de l'importance de l'enjeu (*i.e.*, assigner une valeur à l'action éducative de l'établissement) et de sa difficulté de réalisation (*i.e.*, obtenir une valeur fiable et interprétable), la place accordée à ces indicateurs de performance ne va pas de soi.

Fortes de ce genre de constats, plusieurs études ont été menées pour juger de la pertinence de ces indicateurs et parvenir à des préconisations pour leur utilisation (Goldhaber *et al.*, 2015 ; Perry, 2016 ; Raudenbush et Willms, 1995). Il ressort de ces études, tout d'abord, que les indicateurs de performance peuvent être valides sous certaines stratégies de modélisation multiniveau qui intègrent les caractéristiques individuelles et

contextuelles des établissements, suivant la tradition de recherche sur les effets-établissement (Bressoux 1994, 2010 ; Raudenbush et Bryk 2002 ; Snijders et Bosker 1999). Ensuite, ces études mettent en avant la nécessité d'intégrer une multiplicité d'indicateurs afin d'offrir une appréciation plus globale de l'action éducative, et d'associer les acteurs dans l'appropriation et l'interprétation de l'information fournie par ces indicateurs (Goldhaber *et al.*, 2015). Ces études préconisent donc une stratégie d'évaluation multidimensionnelle et participative des indicateurs de performance au service du pilotage éducatif. Cette approche correspond pour beaucoup à celle couramment privilégiée en France. Ainsi, à partir de modèles statistiques multiniveaux, le ministère de l'Éducation nationale publie chaque année trois indicateurs de valeur ajoutée des lycées (IVAL) concernant le taux de réussite, le taux de mentions et le taux d'accès au baccalauréat (c'est-à-dire, la probabilité d'obtenir le baccalauréat au sein du même lycée, à l'entrée en classe de 2de). Ces IVAL sont conçus comme « des outils qui les aident [*i.e.*, les acteurs éducatifs des lycées] à améliorer l'efficacité de leurs actions » (MENJS-DEPP, 2020, p. 3).

Dans le prolongement de ces interrogations, le présent travail vise à renforcer la pertinence des indicateurs de performance en explicitant et en dépassant une limite méthodologique des IVAL ministériels, pouvant avoir des effets non désirables pour l'évaluation des lycées. Cette limite tient au fait que, par construction, les IVAL ministériels ignorent une erreur d'échantillonnage qui confond, dans la valeur ajoutée des lycées, l'effet de leur propre action avec les effets du hasard. Pourtant, les IVAL font aujourd'hui partie intégrante du pilotage éducatif des établissements et de la volonté de soutenir les pratiques éducatives efficaces, en se centrant sur la preuve fournie par les IVAL. Il convient alors d'« évaluer l'évaluation » (Felouzis, 2005) en estimant la place qu'occupe le bruit statistique dans la preuve administrée à travers les IVAL. Illustrant l'importance de ce bruit et la précarité des évaluations qui en découlent pour de nombreux lycées, cet article propose des méthodes d'estimation alternatives permettant de contrôler les effets du hasard afin d'obtenir une information plus fiable sur l'action éducative des lycées.

Genèse, objectifs et critiques des IVAL en France

Les indicateurs de valeur ajoutée ont été introduits en France afin d'aiguiller l'action éducative des établissements (MEN-DEP, 1994 ; MENJS-DEPP, 2020), en fournissant aux acteurs de terrain un outil d'évaluation fiable et accessible permettant d'implémenter une culture d'action fondée sur la preuve (*evidence-based practice*). Mais de quelle « preuve » est-il question, exactement ? Loin de représenter un objet d'étude évident, la question de la preuve est ici intimement liée à celles de l'existence et de la mesure de l'« effet établissement », qui ont longuement nourri les débats sociologiques sur les inégalités d'apprentissages en contexte scolaire. Si l'existence de cet effet est aujourd'hui bien établie, le problème de sa mesure dans les indicateurs de performance demeure épineux et a pris, nous le verrons, une forme particulière dans le contexte français.

À l'origine, un questionnement sur les effets-établissement

Les établissements ont-ils un effet spécifique sur la réussite scolaire des élèves ? Au cours des années 1960, les premières études affirmèrent la quasi-inexistence des effets-établissements (« *Schools Make No Difference* »), en montrant que les moyens financiers allouées aux écoles (*i.e.*, dotations matérielles et humaines) n'impactaient pas ou très faiblement les acquisitions cognitives des élèves. Au contraire, le milieu social d'origine des élèves et le contexte social qui en découlait jouaient des rôles plus déterminants (voir, ici, l'emblématique rapport Coleman ; Coleman *et al.*, 1966). Si des inégalités de moyens existaient bel et bien d'une école à une autre, ces inégalités ne permettaient pas d'expliquer les différences de réussite entre élèves.

Par opposition aux premiers travaux, les tenants de « l'efficacité de l'école » (*school effectiveness*) contestèrent, dans les années 1970 et 1980, l'assimilation des effets-établissement avec celle des moyens financiers, argumentant que l'action éducative se caractérise avant tout par des processus organisationnels concernant, par exemple, la gestion des cours, l'affectation des élèves aux classes, le climat disciplinaire, ou encore les approches éducatives (Brookover *et al.*, 1979 ; Rutter *et al.*, 1979). Plus qu'une unité de production déterminée par ses intrants institutionnels (*i.e.*, les moyens financiers), l'école serait une entité sociale dont l'efficacité tiendrait à la cohérence de son fonctionnement interne (*i.e.*, les processus organisationnels ; Bressoux, 1994). En harmonisant ces processus, les écoles pouvaient donc faire une différence (« *Schools Can Make a Difference* » ; Brookover *et al.*, 1979). En parallèle à cette critique, le développement de méthodes de décomposition de la variance sur plusieurs niveaux d'analyse (*i.e.*, élèves et écoles) permirent de montrer que les effets-établissement existaient bel et bien et qu'ils s'avéraient faibles à modérés sur les acquisitions scolaires (*i.e.*, entre 8 % et 15 % de la variance à l'étranger, mais moins de 5 % de la variance en France ; Duru-Bellat *et al.*, 2004), et modérés à forts sur le décrochage scolaire (e.g., 25 % de la variance aux États-Unis, 10 % à 16 % de la variance en France ; Núñez-Regueiro, 2018 ; Núñez-Regueiro *et al.*, 2022; Khouaja et Moullet, 2016 ; Li, 2007).

À partir des années 1990, le débat se focalisa davantage sur la nature des effets-établissement afin de savoir dans quelle mesure l'action éducative était indépendante ou non du public accueilli. Des études ont par exemple montré que les enseignants s'adressant à des élèves issus de milieux défavorisés plutôt que favorisés passent davantage de temps à gérer des problèmes disciplinaires en classe, font moins confiance en la capacité à apprendre et à réussir des élèves, et ont des objectifs d'apprentissage moins ambitieux, ce qui se solde par des apprentissages plus modestes et par un risque plus élevé de décrochage scolaire (Duru-Bellat *et al.*, 2004 ; Rumberger et Palardy, 2005 ; Rumberger et Thomas, 2000). Pour certains auteurs, de telles relations entre processus éducatifs et public accueilli obéreraient en principe la possibilité d'identifier l'action éducative indépendamment du contexte social (Thrupp, 1995). Pour d'autres, il s'agissait d'une question empirique ouverte, requérant de modéliser plus finement ces relations (Duru-Bellat, 2002). À cet égard, on remarquera que, si les effets de composition sociale transitaient pour partie ou entièrement par ces processus éducatifs dans lesdites études, une large part de l'effet-établissement demeurait inexpliquée (e.g., entre 60 % et 85 % pour les apprentissages, entre 45 % et 55 % pour le décrochage ; Rumberger et Palardy, 2005 ; Rumberger et Thomas, 2000). Ce genre

de constat suggère que l'effet de l'action éducative d'un établissement est, au moins en partie, indépendant des profils d'élèves accueillis et des processus qui en découlent.

Mesurer l'action éducative grâce à une stratégie de modélisation quasi expérimentale

Dans la lignée des travaux sur les effets-établissement, les IVAL ont été introduits en France afin de distinguer l'effet de l'action éducative d'un établissement – sa valeur intrinsèque, aussi appelée sa « valeur ajoutée » ou sa « performance » –, par rapport aux effets de facteurs exogènes tels que sa composition socio-académique ou son offre de formation – sa valeur extrinsèque, associée à une réputation sociale –, sur lesquels l'établissement n'a que peu d'emprise et qui relèvent avant tout de déterminismes sociohistoriques constitutifs du marché scolaire et de la carte scolaire (Caro et Hillau, 1997 ; Felouzis, 2005 ; Merle, 2011). Si l'outil de mesure visé est clair, sa réalisation est plus difficile.

En effet, relier un modèle statistique de la réussite scolaire à l'action éducative d'un établissement n'est pas anodin du point de vue de l'inférence causale (Raudenbush et Willms 1995). Du fait de déterminismes multiples (e.g., ségrégation sociale et scolaire des établissements, stratégies de scolarisation des parents, inégalités territoriales dans l'offre de formation), les élèves ne sont pas répartis de manière aléatoire à travers les lycées. Pour parvenir néanmoins à une inférence causale plausible, la méthodologie des IVAL recourt à une stratégie de modélisation quasi expérimentale consistant à contrôler les effets des caractéristiques des élèves et du contexte socioscolaire du lycée par l'intermédiaire de régressions logistiques multiniveaux (voir Encadré 1). Dans cette stratégie, on opère un « calcul des différences » entre un taux prédit par le modèle statistique (e.g., 70 % de réussite au baccalauréat) et un taux effectivement constaté (e.g., 80 % de réussite), afin de dégager une valeur ajoutée (e.g., + 10 points de réussite). Le taux prédit correspond alors à la somme des effets éducatifs associés aux facteurs exogènes qui peuvent être effectivement mesurés (*i.e.*, la valeur extrinsèque), ce qui permet de mettre tous les établissements sur un pied d'égalité en termes de niveau attendu ; la différence entre le taux attendu et le taux constaté au sein d'un lycée est alors interprétée comme la valeur ajoutée de l'établissement, sa performance.

ENCADRE 1. – ***Le calcul des IVAL en France***

Les IVAL ministériels prennent appui sur un modèle statistique permettant d'estimer le taux attendu de réussite au baccalauréat (ou de mention au baccalauréat, ou d'accès au niveau supérieur), en ajustant les caractéristiques individuelles et contextuelles du lycée (Evain et Evrard, 2017 ; MENJS-DEPP, 2020). Dans ce modèle, chaque élève *i* a une probabilité de réussite au baccalauréat qui dépend du lycée *j*, notée P_{ij}, et une chance de réussite, notée Y_{ij}^*, suivant l'égalité $Y_{ij}^* = logit(P_{ij})$. Le modèle linéaire des Y_{ij}^* comprend des effets fixes γ_p et γ_q associés, respectivement, aux variables de contrôle individuelles X_p (*p* = 1, . . . , P ; e.g., âge, origine sociale, sexe, note globale obtenue au diplôme national du brevet – DNB) et contextuelles C_q (*q* = 1, . . . , Q ; e.g., proportions d'élèves issus de chaque catégorie d'origine sociale, notes moyennes au DNB), ainsi qu'une constante, notée γ_0, qui représente un effet fixe moyen sur la propension à réussir. Ces effets sont dits « fixes » pour signifier qu'ils ne varient pas en fonction des lycées. Cependant, le modèle intègre un effet aléatoire spécifique à chaque lycée, noté u_j, qui représente l'action éducative globale du lycée (cet effet aléatoire u_j est supposé gaussien, de moyenne nulle et de variance σ_u^2). (L'effet aléatoire spécifique à chaque élève n'est pas identifié dans le modèle logistique, la variance individuelle étant fixée à une constante $\frac{\pi^2}{3} = 3.29$; Snijders et Bosker, 1999). Par cette stratégie de modélisation, les effets éducatifs des facteurs exogènes à l'action de l'établissement (les effets fixes γ) sont formellement distingués de l'effet éducatif du lycée (l'effet u_j), ce que met en évidence le modèle complet :

$$logit(P_{ij}) = Y_{ij}^* = \gamma_0 + \sum_p \gamma_p X_{pi} + \sum_q \gamma_q C_{qj} + u_j \quad (1)$$

Les chances de réussite au baccalauréat pour un élève *i* peuvent donc être prédites à partir des effets fixes et aléatoires γ et u_{0j}. En ignorant l'effet aléatoire u_j, on ignore également l'effet lycée pour ne retenir que les effets exogènes à l'action de l'établissement.

Les IVAL ministériels utilisent alors les effets fixes de ce modèle pour prédire un taux attendu de réussite au sein d'un lycée, que l'on compare au taux constaté. Plus précisément, le modèle est ajusté à l'ensemble des lycées afin d'obtenir une estimation des effets fixes, notés $\hat{\gamma}$. En remplaçant les effets fixes théoriques γ par leurs estimations $\hat{\gamma}$, et en remplaçant les variables correspondantes par leur valeur pour chaque élève, il devient alors possible d'utiliser l'équation 1 (ci-dessus) pour calculer le taux de réussite qui est prédit par le modèle pour l'élève *i* du lycée *j* et, par extension sur tous les élèves, le taux de réussite prédit pour ce lycée *j*. La valeur ajoutée du lycée *j* est alors obtenue en faisant la différence entre le taux prédit de réussite et le taux effectivement observé de réussite. Par exemple, pour un taux prédit de 90 % et un taux observé de 95 %, la valeur ajoutée sera estimée à 5 points de pourcentage (95 - 90 = 5).

Des IVAL mesurant le pouvoir « qualifiant » des établissements

L'action éducative d'un établissement est multidimensionnelle et peut concerner le développement des compétences cognitives (e.g., savoirs académiques, savoirs professionnels), des compétences non cognitives (normes sociales et comportementales, motivation personnelle, confiance en soi), ou encore des débouchés socioprofessionnels (e.g., accès à l'emploi, niveau de salaire, réseau social ; Cunha et Miller, 2014). Au niveau des lycées, l'enseignement scolaire français vise l'obtention d'un diplôme qualifiant en vue de l'insertion professionnelle ou de l'accès aux études supérieures (*Code de l'éducation*, Art. L121-1 à L123-9). Cet objectif de qualification se justifie, entre autres, par le fait que l'absence de diplôme (*vs*. présence) est associée à un risque plus élevé de chômage et de précarité de l'emploi (*i.e.*, revenus faibles, emploi partiel ou à durée déterminée ; Moncel, 2007). Dans cette optique, les IVAL ont été introduits en France en se centrant sur l'obtention du baccalauréat chez les élèves.

Deux de ces IVAL évaluent l'effet de l'action éducative sur les épreuves du baccalauréat en distinguant des aspects quantitatif et qualitatif (*i.e.*, taux de réussite et taux de mention aux épreuves), tandis que le troisième évalue cet effet sur le fait de réussir aux épreuves au sein du même lycée d'origine (*i.e.*, taux d'accès au bac à l'entrée en classe de 2de), afin d'appréhender à quel point l'action éducative permet de conduire tous les lycéens jusqu'au baccalauréat (MENJS-DEPP, 2020). En somme, la combinaison de ces indicateurs complémentaires vise une appréciation multidimensionnelle et donc plus fine de la valeur ajoutée du lycée dans l'accès à la qualification (baccalauréat), que les personnels d'éducation peuvent utiliser pour mieux comprendre l'efficacité de leurs actions face à cet objectif (Goldhaber *et al.*, 2015).

Critiques des IVAL

D'origine institutionnelle, les indicateurs de performance ont besoin d'être interprétés par les acteurs éducatifs intervenant dans les établissements (e.g., chefs d'établissements, enseignants, conseillers d'éducation, inspecteurs d'académie). En effet, la valeur ajoutée d'un lycée a un caractère global qui, en l'absence d'autres informations, ne permet pas d'appréhender les processus sous-jacents (e.g., impossible de dire quelles dimensions spécifiques de l'action éducative ont été efficaces ou non ; Raudenbush et Willms 1995). La participation des acteurs éducatifs est donc nécessaire pour analyser les dimensions éducatives pouvant expliquer la performance d'un établissement (Goldhaber *et al.*, 2015).

Malgré cette logique « centrée sur l'utilisateur », c'est-à-dire sur les acteurs éducatifs, les IVAL peuvent été perçus comme une évaluation externe de l'institution venant sanctionner – positivement ou négativement – le travail des acteurs éducatifs, sans pour autant leur venir en aide (Felouzis, 2005). Certaines critiques historiques des IVAL permettent d'expliquer ce point de vue et de justifier certaines réticences à leur égard. Par exemple, en ignorant le niveau initial des élèves, les premiers IVAL ministériels ont eu tendance à confondre les effets de ségrégation socioscolaire avec ceux de l'action éducative, et donc à renforcer les rumeurs sur la bonne ou mauvaise réputation des lycées (ou estimations « brutes »), au lieu de les dépasser par des estimations « ajustées » rendant compte du public accueilli (Felouzis, 2005). Des critiques similaires ont été adressées aux indicateurs de performance utilisés à l'étranger (voir par exemple les critiques adressées aux évaluations des *School league tables* en Angleterre ou du *National Assessment of Educational Progress* aux États-Unis, à la fin des années 1990 ; Goldstein et Spiegelhalter, 1996 ; Meyer, 1997). Autrement dit, la validité des IVAL en tant que mesures de performance n'a pas toujours été garantie et, nous le verrons, demande encore à être interrogée. D'un point de vue plus politique, la publicité des IVAL a aussi eu pour effets collatéraux de mettre en concurrence les lycées autour de ces seuls indicateurs et, par suite, de négliger d'autres composantes du système éducatif non évalués par ceux-ci, notamment les inégalités scolaires (e.g., ségrégation socioscolaire des établissements, des filières d'enseignement, des acquisitions scolaires, etc. ; Goldstein et Spiegelhalter, 1996 ; Thrupp, 1995). Ainsi, en mesurant le pouvoir « qualifiant » des lycées en France, les IVAL ont mis l'accent sur le niveau moyen de performance, plutôt que sur l'égalité de résultats entre élèves (Givord et Suarez, 2019).

L'institutionnalisation des IVAL comporte donc le double risque de produire des représentations erronées de l'action éducative des lycées (problème de la mesure des IVAL) et de cautionner, sous couvert d'actions fondées sur la preuve, une politique néolibérale de « croissance » éducative qui néglige les inégalités d'accès à cette croissance (problème de l'utilisation des IVAL ; Duru-Bellat, 2002 ; Goldstein et Spiegelhalter, 1996 ; Thrupp, 1995). Sans minimiser l'importance de ces dérives qui ont accompagné l'introduction de ce genre d'indicateurs, le propos sera porté ci-après sur une limite des IVAL français qui interrogent plus directement leur valeur informative en tant qu'éléments de « preuve », c'est-à-dire de mesure de l'action éducative.

Un problème de « bruit » dans la preuve fournie par les IVAL ministériels

Nature du bruit : une erreur d'échantillonnage amalgamée à l'action éducative

Comme indiqué *supra* (Encadré 1), les indicateurs de performance sont calculés à partir de la différence entre un taux prédit et un taux constaté de réussite (Duclos et Murat 2014 ; Evain et Evrard 2017). Or, ce calcul introduit du « bruit » dans la preuve fournie par ces indicateurs et, donc, dans les représentations de l'action éducative qui en résultent. Cette section explique la nature de ce bruit problématique [1], à savoir l'erreur d'échantillonnage qui entoure le taux constaté de réussite.

Le taux de réussite constaté pour l'année *n* au sein d'un lycée *j* représente la proportion des lycéens qui, étant inscrits dans le lycée *j* et s'étant présentés aux épreuves du baccalauréat l'année *n*, l'ont obtenu (e.g., 90 % des lycéens du lycée *j* ont obtenu le baccalauréat à la session $n = 2016$). Formellement, ce taux décrit la réalisation d'une expérience éducative (la réussite ou l'échec au baccalauréat) parmi l'ensemble des candidats d'un lycée donné, dont la distribution statistique peut être décrite par une loi binomiale à deux paramètres (Agresti, 2002). Le premier paramètre concerne le taux théorique de réussite, qui correspond au taux de réussite observé au sein de la population générale (e.g., les 700 000 candidats au baccalauréat à une session donnée). Le second paramètre concerne la variance d'échantillonnage autour de ce taux de réussite, qui correspond à un écart systématique – le bruit statistique – entre le taux observé dans la population et le taux observé dans un échantillon aléatoire de la population (e.g., 100 candidats tirés au hasard).

En théorie, cet écart systématique est entièrement dû au hasard et, plus précisément, aux variations aléatoires induites par l'échantillonnage. Par exemple, des différences de répartition du genre entre deux échantillons tirés au sort (e.g., 51 % de filles

[1] Une autre source de bruit moins problématique concerne l'erreur d'estimation des paramètres du modèle de prédiction associé au taux prédit. À la différence de l'erreur d'échantillonnage qui joue en faveur (ou défaveur) des petits lycées, ce bruit paramétrique concerne l'ensemble des lycées et n'introduit donc pas d'« inégalités » dans l'évaluation. Ce bruit est en outre minimisé par la stratégie d'estimation par ré-échantillonnage paramétrique (Goldstein et Spiegelhalter, 1996 ; Knowles et Frederick, 2019).

dans l'échantillon 1 contre 48 % dans l'échantillon 2) se traduiront par des différences de réussite au baccalauréat qui seront proportionnelles à l'effet du genre sur la réussite et aux écarts de répartition observés entre les échantillons et la population (e.g., pour 49 % de filles dans la population nationale, on observerait ici un écart de répartition de + 2 points dans l'échantillon 1 et de - 1 point dans l'échantillon 2). Plus l'échantillonnage sera restreint (e.g., 10 élèves), plus les chances seront élevées d'observer des écarts de répartition importants et, par conséquent, des écarts de réussite au baccalauréat par rapport à la population. *A contrario*, un échantillonnage large (e.g., 10 000 élèves) permettra d'approximer avec plus de certitude les paramètres de la population. À un niveau agrégé prenant en compte l'ensemble des variations aléatoires, la variance d'échantillonnage offre un indicateur global qui résume l'écart escompté (l'erreur attendue) entre le taux de réussite constaté dans un échantillon donné et le taux constaté dans la population. Cet écart, connu à travers la loi binomiale, est proportionnel au nombre d'individus dans l'échantillon et au taux de réussite dans la population (voir équation 1, Encadré 2).

Par rapport à la loi binomiale décrite ci-dessus, la méthodologie des IVAL présente l'intérêt de reconnaitre que les lycées ne sont pas représentatifs de la population (en raison des facteurs « exogènes » mentionnés *supra*) et qu'il convient donc de remplacer le taux théorique de la loi binomiale (*i.e.*, le taux de réussite observé dans la population) par un taux prédit par un modèle statistique, qui prenne en compte les caractéristiques du public accueilli dans chaque lycée (voir Encadré 1). Ce faisant, la valeur ajoutée des lycées est associée à l'écart entre le taux constaté et le taux prédit, le raisonnement étant que cet écart reflète l'action éducative du lycée. Pourtant, le fait que le taux observé soit soumis à une variance d'échantillonnage nous indique que cet écart est, en partie du moins, le fruit d'une erreur d'échantillonnage. Cela indique que la méthode de calcul actuellement utilisée (la différence taux constaté-taux prédit) confond deux sources de variance dans la valeur ajoutée, l'une due à l'action éducative, l'autre due au « bruit » de l'erreur d'échantillonnage.

Le fait que les indicateurs de performance contiennent du bruit n'est pas surprenant en soi : analyser un phénomène suppose bien de se centrer sur ses traits essentiels et d'en ignorer d'autres plus secondaires, afin de simplifier la modélisation de l'objet. Toute modélisation statistique comporte ainsi une forme d'approximation qui laisse de la place à l'incertitude dans les estimations. Cette simplification de la réalité peut néanmoins s'avérer problématique lorsque le bruit devient prépondérant par rapport à l'information fournie par le modèle. Dans cette perspective, il convient de mesurer l'ampleur du bruit contenu dans les indicateurs de performance. D'après la loi binomiale, il est possible d'identifier la présence de ce bruit en testant si l'écart entre les taux prédit et constaté de performance d'un lycée (*i.e.*, méthode de calcul des IVAL ministériels) ne diffère pas, pour un seuil de significativité donné (e.g., $p < .05$), de l'écart escompté d'après la variance d'échantillonnage (voir Encadré 2, équation 2). Dans le cas d'un écart non significatif (i.e., $p > ,05$), la valeur ajoutée sera alors considérée comme étant confondue avec l'erreur d'échantillonnage, c'est-à-dire bruitée. La question étant alors de savoir dans quelle mesure les valeurs ajoutées s'avèrent bruitées.

À quel point les IVAL sont-ils bruités ?

Nous appliquons à présent le test binomial pour établir la significativité des IVAL de réussite au baccalauréat tels que calculés à partir des données (Fichiers anonymisés d'élèves pour la recherche et les études – FAERE) et des techniques de modélisation en vigueur au niveau national (modèle de régression logistique multiniveau, Encadré 1 ; voir Tableau A1 en Annexe). Il s'agit ainsi d'estimer la prévalence du bruit statistique contenu dans les IVAL. Pour plus de robustesse, l'analyse porte sur les valeurs ajoutées de plusieurs sessions consécutives (2014, 2015, 2016), se centre sur les baccalauréats généraux et technologiques, et ne prend en compte que les lycées comprenant au moins vingt candidats à chaque session, conformément aux précautions prises dans le calcul des IVAL (Duclos et Murat, 2014 ; MENJS-DEPP, 2020). L'échantillon final comprend 112 lycées et concerne à chaque session plus de 20 000 élèves, le taux de réussite au baccalauréat oscillant autour de 94 % (voir Tableau 1). À de légères variations près, ces données reproduisent bien les statistiques publiques de l'académie de Grenoble (SESPAG, 2019).

En premier lieu, les résultats de l'étude attestent d'un fort risque que les écarts entre les taux prédits et les taux observés, de même que leurs variations au cours du temps, soient dus au bruit (voir Tableau 1). Ainsi, pour chacune des sessions de baccalauréat, près de 80 % des lycées se voient attribuer un IVAL non significatif, indiquant que l'estimation de la valeur ajoutée de leur action éducative est bruitée et s'avère confondue avec l'action du hasard. Cette forte prévalence du bruit s'aggrave au cours du temps pour concerner la quasi-totalité des lycées : au bout de trois sessions (2014, 2015 et 2016), 98 % d'entre eux connaissent au moins une fois l'attribution d'un IVAL non significatif. Outre cette forte prévalence, le bruitage des IVAL apparait permanent pour 61 lycées (54 %), en ceci que leur IVAL est non significatif sur chacune des sessions de baccalauréat. Ces premiers résultats montrent donc que, dans la grande majorité des cas, et parfois de manière continue, l'information fournie par les IVAL concernant l'action éducative des lycées sur la réussite au baccalauréat n'est pas interprétable en soi, en raison de son amalgame avec du bruit statistique.

En deuxième lieu, il apparait que les IVAL non significatifs présentent une hétérogénéité importante, qui contribue à les rendre peu repérables sur le plan de l'intuition. Par exemple, toutes sessions confondues (2014 à 2016), ces IVAL sont représentés de manière assez lissée ou équilibrée sur l'ensemble des quartiles définissant les tailles d'effet des valeurs ajoutées (*i.e.*, Q1 = 31 %, Q2 = 31 %, Q3 = 26 %, Q4 = 11 %) ou encore le nombre de candidats dans le lycée concerné par l'IVAL (*i.e.*, Q1 = 28 %, Q2 = 24 %, Q3 = 23 %, Q4 = 24 %). Plus de 44 % des IVAL non significatifs se trouvent par ailleurs parmi les 50 % des valeurs ajoutées les plus hautes et, simultanément, apparaissent parmi les 50 % des lycées les plus grands en termes d'effectifs. Dans ces conditions, le bruit apparait non seulement prépondérant mais aussi difficile à repérer sur la seule base de la taille d'effet ou des effectifs de lycée. Cette « invisibilité » explique aussi sans doute pourquoi, d'ailleurs, la question du bruit au sein des IVAL n'a pas été formellement soulevée et traitée jusqu'à aujourd'hui.

Enfin, en troisième lieu, si l'ensemble des lycées est ainsi concerné par l'affectation d'une valeur ajoutée non significative, on peut s'interroger sur les 61 lycées pour lesquels la

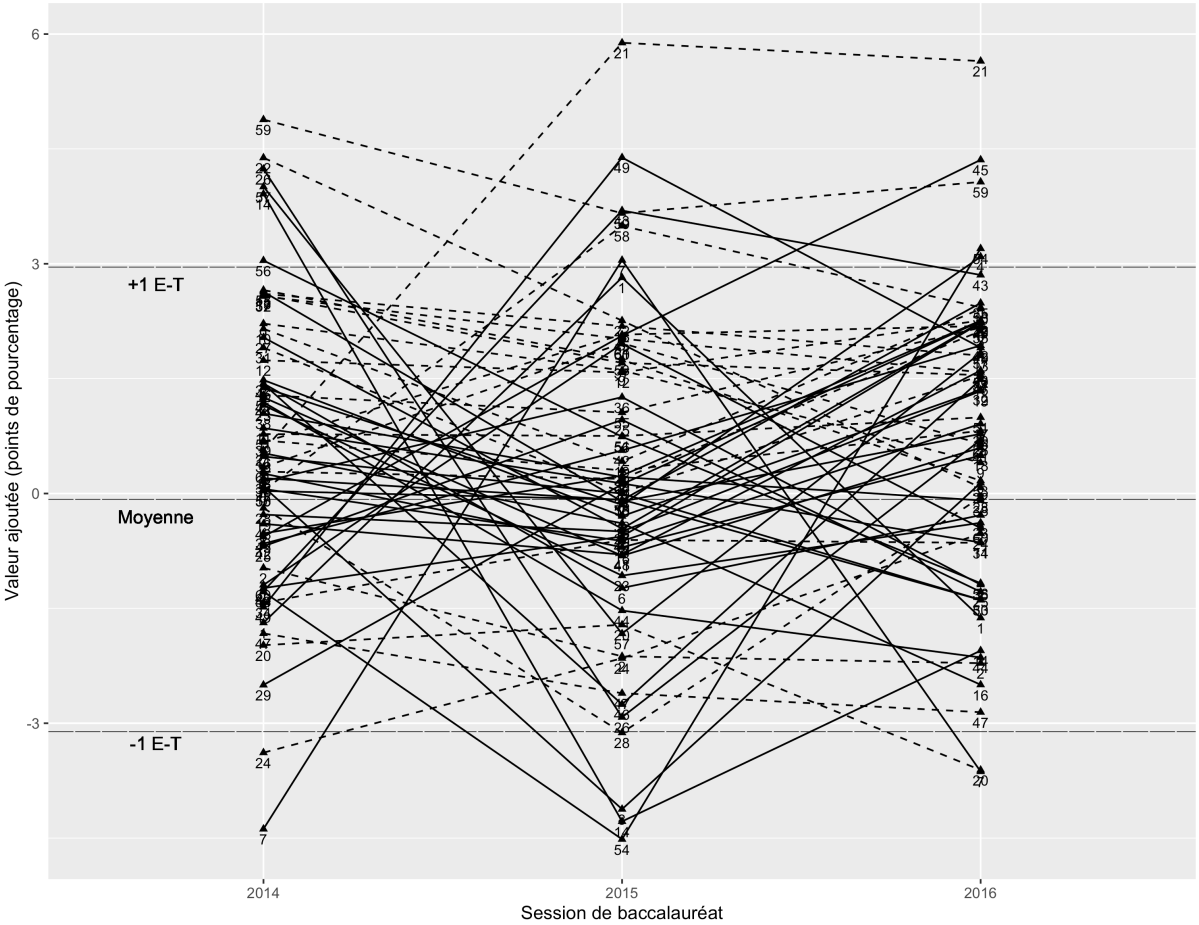
valeur ajoutée n'est jamais significative sur la période observée. L'analyse révèle ici que leur IVAL présente le plus souvent une valeur ajoutée élevée et/ou une valeur ajoutée instable au cours du temps. Plus précisément, 30 % d'entre eux ont des valeurs ajoutées qui peuvent être considérées importantes (*i.e.*, valeurs ajoutées situées 1 écart type au-dessus ou en-dessous de la moyenne ; voir Graphique 1), et qui estimeraient leur action éducative à 3 points de pourcentage ou plus (en valeur absolue) au-delà ou en deçà de la valeur attendue. En outre, 61 % de ces lycées (ou 37,2 % de l'ensemble des lycées) ont des valeurs ajoutées qui changent de signe entre une ou deux sessions de baccalauréat, leur action éducative passant de positive à négative, ou de négative à positive d'une année sur l'autre (voir les trajectoires en pointillés dans le Graphique 1). Ces changements de signe sont souvent conséquents et représentent, en moyenne, entre 2,8 points et 3 points de pourcentage de réussite au baccalauréat, soit l'équivalent d'un écart type de variation (ET = 3 points). Autrement dit, alors même que ces lycées reçoivent un IVAL qui n'est jamais significatif sur la période observée en raison de l'erreur d'échantillonnage, leur interprétation au sein de la méthodologie ministérielle amènerait à penser que l'action éducative de ces lycées évolue de manière importante d'une année sur l'autre.

TABLEAU 1. ***– Statistiques descriptives sur la réussite au baccalauréat et la valeur ajoutée des lycées (indicateur ministériel) au sein de l'académie de Grenoble, sessions 2014 à 2016***

	2014	2015	2016	2014-2016 (≥ 1 fois)	2014-2016 (3 fois)
Taux de réussite	94,4 %	94,2 %	93,7 %	—	—
Nombre de candidats	20 230	22 249	23 116	—	—
Nombre moyen de candidats par lycée (étendue)	184 (24 ; 477)	196 (35 ; 492)	210 (34 ; 578)		
IVAL moyen (écart type)	- 0,06 (3,25)	- 0,08 (2,80)	- 0,08 (3,07)	—	—
Lycées dont l'IVAL n'est pas significatif (p > ,05)	90 (80 %)	88 (79 %)	90 (80 %)	110 (98 %)	61 (54 %)

Note : N = 112 lycées ayant au moins 20 candidats au baccalauréat général et technologiques à chaque session.
Lecture : Entre 2014 et 2016, les valeurs ajoutées (IVAL) calculées pour 110 lycées (soit 98 % des lycées) ne peuvent être différenciés de l'action du hasard (*i.e.*, test binomial non significatif à 5 %) dans au moins une session de baccalauréat.

GRAPHIQUE 1. ***–Valeurs ajoutées de réussite au baccalauréat sur les sessions 2014 à 2016, parmi des lycées de l'académie de Grenoble***



Note : *N* = 61 lycées dont la valeur ajoutée n'est jamais significative au seuil de 5 %. Les triangles indiquent la valeur ajoutée de chaque lycée en fonction de la session de baccalauréat. Aucune des valeurs ajoutées affichées n'est significativement différente de l'action du hasard (test binomial, $p > ,05$). Les droites pleines esquissent la trajectoire des valeurs ajoutées qui changent de signe au cours du temps (de positive à négative, ou de négative à positive), et inversement pour les droites en pointillés (pas de changement de signe).

Conséquences pratiques du bruit : une prise d'information qui risque d'être trompeuse et un pilotage hasardeux

Les résultats de cette étude contribuent à révéler une situation pour le moins équivoque : alors que l'institution scolaire enjoint les acteurs éducatifs à fonder leurs actions sur la preuve des indicateurs de performance, cette preuve s'avère en partie fondée sur du bruit statistique. Cette situation a pour conséquences de rendre incertaine la prise d'information quant à l'efficacité des actions éducatives et, sur cette base, de rendre le pilotage éducatif tributaire du hasard.

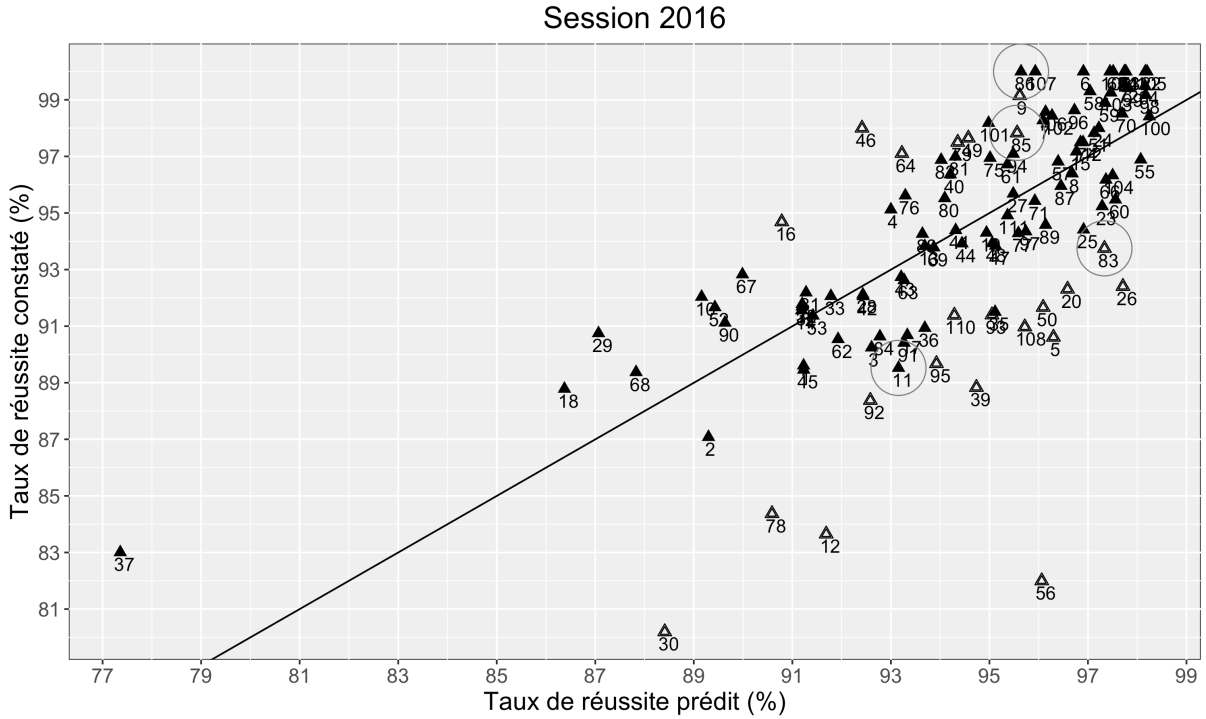
Afin d'illustrer les erreurs de jugement qui peuvent résulter du bruit, nous projetons les taux prédits et observés de réussite à la session 2016, estimés sur le même jeu de données que précédemment. Du côté droit du Graphique 2, nous observons que le lycée « 86 » présente une valeur ajoutée élevée de l'ordre de 4,4 points, signifiant que son taux de réussite observé (*i.e.*, 100 % de bacheliers) est plus de 4 points au-dessus du taux qui

était attendu d'après le modèle de prédiction (*i.e.*, 95,6 % de bacheliers). Au seul vu de cette valeur, l'action de ce lycée parait tout à fait efficace pour soutenir la réussite scolaire. Pourtant, cette valeur ajoutée n'est en fait pas significative d'après le test binomial, car l'effectif de ce lycée est relativement faible (N = 77 élèves) et induit une erreur d'échantillonnage importante. L'écart entre les deux taux ne peut donc pas être attribué à l'efficacité du lycée parce que le risque qu'il soit dû au hasard est trop élevé. À l'inverse, le lycée « 85 » présente – pour un taux prédit de réussite similaire – une valeur ajoutée de 2,3 points de pourcentage qui s'avère significative à 5 % au vu de son effectif (N = 415). Autrement dit, tandis que le premier lycée a une valeur ajoutée deux fois plus importante que celle du second, seule cette dernière peut être associée avec quelque confiance à l'action du lycée. Inversement, le test binomial révèle des différences de significativité entre les lycées « 83 » (N = 256) et « 11 » (N = 105) pour des IVAL négatifs identiques (- 3,6), indiquant que seule la valeur ajoutée du premier reflète son action éducative avec quelque certitude. Plus globalement, ces exemples suggèrent que deux lycées équivalents sur le plan de l'efficacité de leur action éducative risquent de se voir différemment valorisés par l'indicateur IVAL dès lors que l'effet du hasard éloignera chacun des lycées de leur taux attendu de manière différenciée.

En termes de pilotage, une valeur ajoutée bruitée par le hasard (*i.e.*, due à l'erreur d'échantillonnage) est par essence inexplicable et non informative. Pourtant, elle a toutes les chances d'être interprétée par les équipes éducatives comme la conséquence de leurs actions, et cela pour deux raisons principales. D'une part, l'indicateur de performance incorpore bel et bien, pour une part inconnue (mais que nous éclairerons plus loin dans cet article), l'effet de leurs pratiques. D'autre part, les humains ont tendance, spontanément, à ne pas voir l'effet du hasard dans des évènements où, pourtant, il peut avoir une part prédominante. Les travaux sur le jugement humain ont en effet bien montré notre faible niveau d'intuition statistique (défaut de prise en compte du hasard) et, à l'inverse, notre propension à établir des associations causales fondées sur l'accessibilité immédiate de l'information (Kahneman, 2011). Ainsi, les acteurs éducatifs auront tendance à mobiliser des explications causales qui leur sont plus accessibles pour interpréter la valeur ajoutée d'un lycée (e.g., dispositifs ou recrutements réalisés sur la période récente, théories éducatives en vogue ou ancrées dans les pratiques), plutôt qu'une explication difficile d'accès (*i.e.*, l'effet du hasard).

Dans le pire des scénarios – qui concerne 37 % des lycées dans notre étude longitudinale (Graphique 1) –, l'effet du hasard pourra contribuer à produire des IVAL positifs une année, négatifs l'année suivante pour un même lycée, les acteurs éducatifs se retrouvant ainsi contraints d'interpréter les fluctuations positives ou négatives de leur IVAL et, dans une spirale d'auto-évaluation éprouvante, à valider ou invalider successivement leurs pratiques, indépendamment de l'efficacité réelle de ces dernières. Des développements s'imposent donc pour prendre en compte l'effet du hasard et obtenir des IVAL moins équivoques pour le pilotage éducatif.

GRAPHIQUE 2. – ***Significativité de la différence entre les taux de décrochage scolaire prédit et observé (i.e., IVAL brut) pour chaque lycée de l'académie de Grenoble***



Note : N = 112 lycées. La droite indique une égalité entre les taux prédit et constaté de réussite au baccalauréat (X = Y). Les triangles de couleur claire indiquent les IVAL significatifs à 5 % d'après un test binomial, les triangles de couleur noire les IVAL non significatifs.

Vers une mesure moins bruitée de l'action éducative des lycées

Contrôler l'erreur d'échantillonnage : introduction des IVAL brut et corrigé

Les sections précédentes ont montré la nécessité de réformer la stratégie d'estimation de la valeur ajoutée et, en particulier, de contrôler l'effet du hasard associé à l'erreur d'échantillonnage, afin d'obtenir des IVAL plus informatifs. Deux stratégies sont envisagées ici (voir Encadré 2) qui, nous le verrons, offrent des résultats convergents du point de vue de l'évaluation de l'action éducative face à la réussite au baccalauréat.

La première stratégie, appelée « IVAL brut », consiste à utiliser un test binomial pour intégrer l'erreur d'échantillonnage associée à l'estimation de la valeur ajoutée. Le mode de calcul reste le même que dans la méthodologie des IVAL ministériels : elle correspond à la différence entre un taux attendu (le taux prédit par un modèle statistique) et un taux constaté. Cependant, l'IVAL brut teste en outre la significativité de cette différence afin de distinguer les valeurs ajoutées qui sont le fruit du hasard (*i.e.*, de l'erreur d'échantillonnage) et celles qui reflètent une action éducative sur la variable d'intérêt. Par exemple, le taux attendu de réussite au baccalauréat pour le lycée 11 (N = 105, Graphique 2) est de 93,2 %. Le seul effet du hasard pourra faire que le taux constaté diffère du taux attendu en fonction de l'erreur type (*i.e.*, E-T = $\sqrt{Var} = \sqrt{0.932 * (1 - 0.932)/105} = 0{,}025$; voir équation 2 de l'Encadré 2). Afin de s'assurer que la différence entre les deux taux n'est pas due au hasard, on vérifie que le taux constaté du lycée 11 (*i.e.*, 89,5 %), ainsi que l'essentiel des valeurs du

taux constaté qui pourraient être dues au hasard (95 % des valeurs produites par l'erreur d'échantillonnage) ne comprennent pas le taux attendu de réussite. D'après l'erreur type, il est estimé que ces valeurs sont comprises entre 84,6 % et 94,4 %[2]. Or, le taux attendu de réussite étant compris dans cet intervalle, nous pouvons conclure que celui-ci ne diffère pas significativement du taux constaté ou, dit autrement, que leur différence (*i.e.*, $IVAL_{brut,lycée11}$ = 0,895 - 0,932 = - 0,037) n'est pas significative. C'est ce même raisonnement qui a été utilisé précédemment pour évaluer la significativité des valeurs ajoutées (voir Tableau 1 et Graphique 2). En somme, l'IVAL brut produit des informations qualitatives sur la nature positive, négative ou non significative de la valeur ajoutée. En revanche, dans la mesure où le test statistique ne corrige pas l'estimation elle-même de la valeur ajoutée, l'IVAL brut a une valeur limitée en tant qu'indicateur quantitatif. En effet, celui-ci est négativement corrélé au nombre de lycéens dans l'établissement (sur les sessions 2014 à 2016, - ,172 > $r(|IVAL_{brut,j}|, n_{total,j})$ > - ,289, ,002 < *p* < ,069), indiquant que les IVAL bruts des plus petits lycées sont systématiquement surévalués ou sous-évalués.

La seconde stratégie, appelée « IVAL corrigé », permet au contraire de corriger l'estimation de la valeur ajoutée en mobilisant la méthodologie propre à l'analyse multiniveau (Raudenbush et Bryk, 2002), qui est aussi celle utilisée au niveau international pour le calcul des indicateurs de performance (Goldstein et Spiegelhalter, 1996 ; Perry, 2016 ; Raudenbush, 2014). Dans cette approche, la valeur ajoutée équivaut à un effet résiduel spécifique à chaque lycée, « corrigé » par un paramètre de fidélité qui incorpore l'erreur d'échantillonnage (voir Encadré 2). Plus l'erreur d'échantillonnage est élevée, et plus elle « contamine » l'effet spécifique du lycée contenu dans la valeur ajoutée. La correction par le paramètre de fidélité permet alors de ramener l'effet du lycée depuis sa valeur contaminée par l'effet du hasard à une valeur débarrassée de cet effet, qui reflète mieux, en moyenne, l'action spécifique du lycée. Grâce à cette pondération, l'IVAL corrigé parvient à gommer, au sein de la taille d'effet « brute » de l'action éducative du lycée (l'effet résiduel non corrigé), l'effet qui est en fait dû à l'erreur d'échantillonnage (elle-même liée au nombre de lycéens), pour ne retenir que l'effet « ajusté » du lycée (*i.e.*, non dû au hasard). Autrement dit, l'IVAL corrigé permet une quantification adéquate de la valeur ajoutée d'un lycée qui neutralise l'effet du hasard. L'efficacité de cette stratégie est d'ailleurs apparente dans le fait que, sur toutes les observations considérées ici (2014 à 2016), l'IVAL corrigé ne corrèle jamais de manière significative avec les effectifs de lycéens (- ,026 < $r(|IVAL_{corrigé,j}|, n_{total,j})$ < ,133, ,164 < *p* < ,788), contrairement à ce qui était le cas pour l'IVAL brut (voir *supra*).

[2] Par souci de simplicité, nous recourons ici à l'approximation de Wald utilisée dans le test de proportions, qui permet d'estimer l'intervalle de confiance du taux constaté à partir de l'erreur type (σ_j), du nombre d'élèves ($n_{total,j}$) et d'un quantile de la loi normale (score *z*), selon la formule : $\pi_{constaté,j} * n_{total,j} \pm z * \sigma_j$ (Agresti, 2002). Ainsi, l'intervalle de confiance à 95 % (|*z*| = 1,96) pour le taux constaté du lycée 11 vaut $IC_{Wald} = 0{,}895 \pm 1{,}96 * 0{,}025 = (0{,}846\ ;\ 0{,}944)$. Ces valeurs sont proches de celles calculées selon la méthode exacte utilisée dans le test binomial, $IC_{exact} = (0{,}820\ ;\ 0{,}947)$, dont le calcul – plus complexe – est fonction des distributions *beta* et *F*. Les deux méthodes diffèrent néanmoins, car le test de proportions perd en exactitude pour des taux proches de 0 ou 1. Par précaution, il convient d'utiliser la méthode exacte. Dans cette étude, des tests binomiaux (méthode exacte) sont utilisés.

ENCADRE 2. – ***Deux stratégies pour contrôler l'erreur d'échantillonnage dans le calcul des IVAL***

« IVAL brut » : établir la significativité de l'IVAL suivant un test binomial

La stratégie de l'IVAL brut consiste à évaluer la significativité de la différence entre le taux constaté de réussites au baccalauréat d'un lycée (ou de mention, de décrochage, etc.) et le taux prédit par un modèle multivarié valable pour l'ensemble des établissements (Encadré 1), en recourant à un test binomial (Agresti, 2002). Le raisonnement est le suivant. Si le lycée *j* était représentatif de la population générale, le nombre constaté de réussites au baccalauréat ($n_{constaté,j}$) suivrait une loi binomiale caractérisée par le nombre total de lycéens en *j* ($n_{total,j}$) et par un taux de réussite théorique ($\pi_{théorique}$), correspondant au taux de réussite dans la population. Cependant, afin d'ajuster les attributs non représentatifs des élèves et du contexte du lycée *j*, le taux de réussite théorique est remplacé par le taux de réussite prédit pour le lycée *j* ($\pi_{prédit,j}$). La distribution du taux de réussite au baccalauréat au sein du lycée *j* se caractérise alors par le taux prédit de réussites et par une variance d'échantillonnage σ_j^2, définis par l'équation 2 :

$$E\left(taux_{constaté,j}\right) = n_{prédit,j}/n_{total,j} = \pi_{prédit,j}$$

$$Var\left(taux_{constaté,j}\right) = \sigma_j^2 = \pi_{prédit,j}(1-\pi_{prédit,j})/n_{total,j} \quad (2)$$

La valeur ajoutée est considérée significative lorsque le taux constaté de réussites diffère suffisamment du taux prédit (e.g., à 95 % de la variance d'échantillonnage).

« IVAL corrigé » : corriger le calcul de l'IVAL en fonction d'un paramètre de fidélité

Au sein de la stratégie de l'IVAL corrigé, la valeur ajoutée d'un lycée *j* vis-à-vis de la réussite au baccalauréat (ou d'un autre phénomène éducatif) correspond à son effet résiduel au sein du modèle de prédiction valable pour la population (voir Encadré 1), étant donné le degré de fidélité de cet effet résiduel (Raudenbush et Bryk 2002). Notée λ_j, cette fidélité conditionnelle (ajustée) est définie par l'équation 3 :

$$\lambda_j = \frac{\sigma^2_{\text{lycées}}}{\sigma^2_{\text{lycées}} + \sigma^2_{\text{élèves}}/n_{total,j}} \quad (3)$$

où $\sigma^2_{\text{lycées}}$ est la variance de réussite située au niveau des lycées et $\sigma^2_{\text{élèves}}/n_{total,j}$ est la variance de réussite située au niveau des élèves, qui modélise la variance d'échantillonnage de l'effet-lycée. Le résidu statistique « corrigé », noté précédemment u_j (Encadré 1), est alors égal au résidu « brut », noté ici $u_{j,brut}$, pondéré par sa fidélité, suivant l'équation 4 :

$$u_j = \lambda_j * u_{j,brut} \quad (4)$$

Ainsi, les lycées ayant une fidélité λ_j plus faible (une erreur d'échantillonnage plus élevée) que ceux ayant des effectifs élevés (équation 3) subissent une correction plus forte de leur effet spécifique $u_{j,brut}$ sur la réussite au baccalauréat (équation 4). Dans cette approche, la contribution de u_j peut s'exprimer en points de pourcentage en faisant la différence entre un taux prédit par l'ensemble des effets fixes (γ) et un taux prédit par l'ensemble des effets fixes et aléatoires (γ et u_j) du modèle statistique (Encadré 1). L'IVAL corrigé permet ainsi d'obtenir une quantification adéquate de la valeur ajoutée qui neutralise l'effet du hasard associé à l'erreur d'échantillonnage.

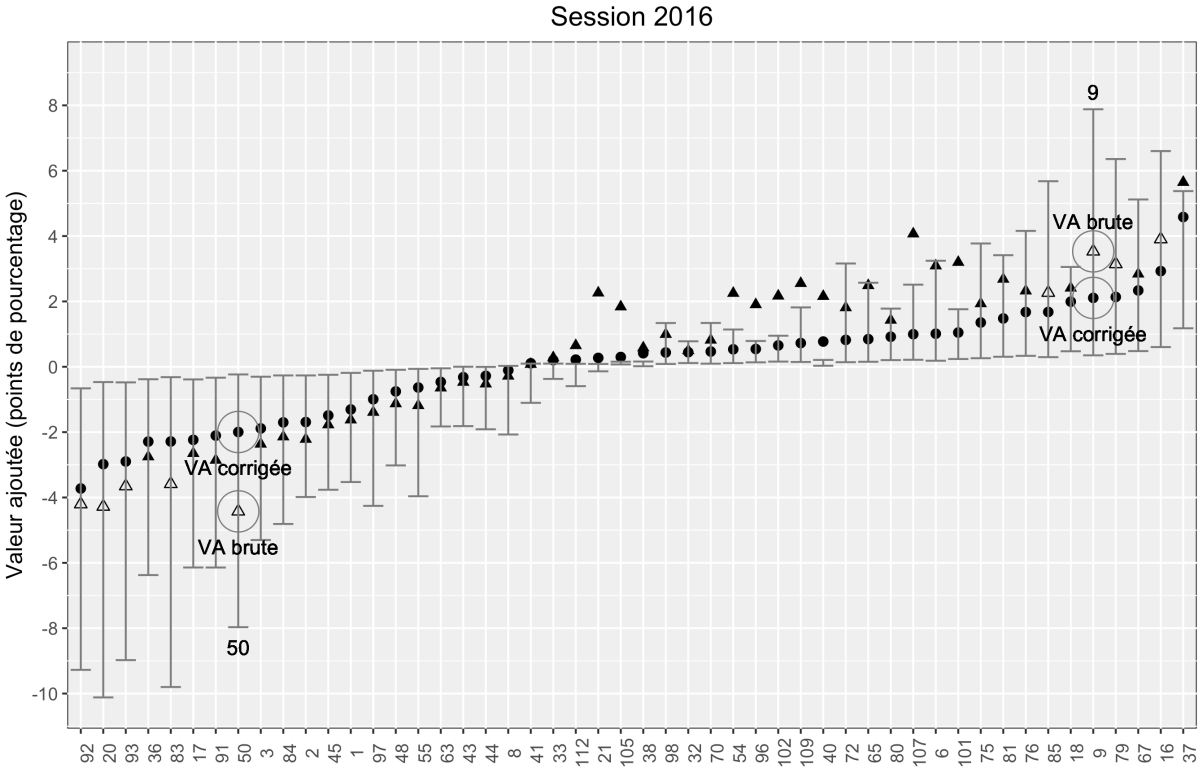
Illustration empirique : la valeur ajoutée corrigée des lycées de l'académie de Grenoble

La stratégie de l'IVAL brut, qui permet de tester la significativité des IVAL ministériels, a été illustrée précédemment pour déceler la présence de bruit statistique au sein de la valeur ajoutée face à la réussite au baccalauréat (Tableau 1 et Graphique 2). Le propos sera donc porté ici sur l'illustration de la stratégie de l'IVAL corrigé, en mobilisant toujours le même jeu de données et en focalisant l'analyse sur la session 2016 du baccalauréat.

Pour rappel, l'IVAL corrigé est égal à l'IVAL brut multiplié par sa fidélité, les deux étant donc quasiment identiques pour une fidélité élevée proche de 1. Une analyse sur l'ensemble des lycées montre que 36 % des lycées ont une fidélité inférieure à ,900. Ceci indique que l'estimation de la valeur ajoutée à travers l'IVAL brut risque d'être surévaluée tout particulièrement pour ces lycées. À un niveau agrégé, nous observons en effet que les estimations de l'IVAL brut sont, en moyenne, deux fois supérieures aux estimations de l'IVAL corrigé [*i.e.*, $\text{Moy}\left(\frac{|IVAL_{brut}|}{|IVAL_{corrigé}|}\right) = 2{,}1$] et que la différence d'ampleur entre les deux IVAL est significative à 0,1% [$t(111) = -9{,}243$, $p < {,}001$]. Nous pouvons donc dire que l'IVAL brut est surestimé par rapport à l'IVAL corrigé, cela étant dû au fait que seul l'IVAL corrigé tient compte de l'effet du hasard au niveau quantitatif et, plus précisément, des effectifs de lycéens à travers la pondération exercée par la fidélité conditionnelle (voir équation 4, Encadré 2). Nonobstant ces variations en termes d'estimations, les deux IVAL sont positivement et fortement corrélés entre eux ($r = {,}48$, $p < {,}001$), le classement des lycées restant en grande partie le même quel que soit l'indicateur retenu.

Afin de matérialiser ces observations, nous projetons la valeur ajoutée de 50 lycées tirés au hasard en fonction des deux IVAL (Graphique 3). Cette projection donne à voir l'association positive entre les deux IVAL, mais aussi le fait que la valeur ajoutée d'un lycée varie selon l'indicateur retenu. Par exemple, le lycée « 50 » (N = 96 lycéens) a une valeur ajoutée brute significative de - 4,4, signifiant que son action en propre contribue à diminuer le taux de réussite de 4,4 points par rapport à ce qui était attendu d'après ses caractéristiques sociodémographiques et académiques. Cette valeur situe ce lycée au 50^{e} rang des 50 lycées en termes de capacité à faire réussir les élèves aux épreuves du baccalauréat. Cependant, cette contribution est estimée à - 2,0 points au sein de l'IVAL corrigé, soit une valeur négative deux fois moins importante qui le situe au 43^{e} rang. Inversement, le lycée « 9 » (N = 235 lycéens) contribuerait à augmenter le taux de réussite de 3,5 points d'après l'IVAL brut (3^{e} rang), mais de 2,1 points seulement d'après l'IVAL corrigé (5^{e} rang ; Graphique 3). Ces exemples illustrent bien les limites associées à l'interprétation de l'IVAL brut, notamment sa forte surestimation de l'effet-lycée à laquelle il conduit. Cette surestimation est neutralisée par l'IVAL corrigé, ce qui induit des variations de rang modérées. Nous observons par ailleurs que le test binomial permet d'identifier assez précisément les lycées qui ont une contribution relative élevée, les IVAL bruts significatifs faisant partie des IVAL corrigés les plus élevés (en valeur absolue) ; cela confirme l'intérêt de ce test pour identifier les lycées ayant un effet positif ou négatif sur la réussite scolaire.

GRAPHIQUE 3. **– *Valeur ajoutée des lycées dans la réussite au baccalauréat (points de pourcentage), en fonction de la définition brute ou corrigée de la valeur ajoutée***



Note : N = 50 lycées tirés au sort à partir de la population des lycées (N = 112). Les triangles renvoient à l'IVAL brut, les points à l'IVAL corrigé. Les triangles de couleur claire indiquent les IVAL bruts significatifs à 5 % d'après un test binomial, les triangles de couleur noire les IVAL bruts non significatifs. L'intervalle de confiance à 95 % correspond aux valeurs de l'IVAL corrigé à travers les simulations.

*

* *

Bien qu'historiquement mis au service de la mise en concurrence des établissements scolaires (e.g., en Angleterre), les indicateurs de performance se sont aujourd'hui développés au service du pilotage éducatif (Goldhaber *et al.*, 2015 ; MENJS-DEPP, 2020). Dans un contexte national où la ségrégation socioscolaire des lycées est une réalité, les acteurs éducatifs peuvent ressentir le besoin de juger de l'efficacité de leurs interventions, au-delà des inégalités territoriales ou entre établissements qui font la plus ou moins grande réussite des élèves. Les indicateurs de performance répondent à ce besoin d'information en proposant une mesure tangible de l'action éducative sur plusieurs indicateurs de réussite (e.g., accès ou mention au baccalauréat). En tant qu'outils statistiques, ces indicateurs restent des représentations simplifiées de l'action éducative dont la pertinence demande à être évaluée en elle-même, tant sur le plan de la mesure que de l'utilisation qui en est faite (Felouzis, 2005 ; Givord et Suarez, 2019 ; Goldstein et Spiegelhalter, 1996 ; Perry, 2016 ; Raudenbush et Willms, 1995 ; Thrupp, 1995). Dans la lignée de ces critiques, la présente

recherche a montré que les IVAL utilisés en France peuvent encore se perfectionner de manière importante sur le plan de la mesure.

Les résultats de notre étude portant sur la réussite au baccalauréat ont montré que la valeur ajoutée estimée à partir de la méthodologie en vigueur est en partie contaminée par du bruit statistique, en raison d'une erreur d'échantillonnage non prise en compte. Sur la base de 112 lycées évalués sur trois sessions d'examens consécutives (2014 à 2016), il est apparu que le bruit statistique était important en termes de prévalence (e.g., 80 % de valeurs ajoutées étaient non significatives et donc susceptibles de ne différer de la moyenne que par l'effet de variations aléatoires), d'étendue (e.g., 98 % des lycées sont concernés par l'attribution d'une valeur ajoutée non significative au bout de trois sessions) et de constance (e.g., 54 % des lycées se voient attribuer une valeur ajoutée qui est bruitée à chacune des sessions). Qui plus est, cette contamination par l'erreur d'échantillonnage n'était pas repérable à partir de critères habituellement utilisés pour juger de la robustesse de la preuve fournie par les IVAL (e.g., taille d'effet de la valeur ajoutée, effectifs du lycée concerné par l'évaluation), les IVAL contaminés s'avérant hétérogènes sur ces aspects. Les personnels d'éducation se retrouvent donc à valider ou invalider leurs pratiques sur la base d'indicateurs « bruités » voire, dans le pire des scénarios, tout à fait aléatoires (lorsque l'effet du hasard est prépondérant). Au total, parce que ce bruit est susceptible d'affecter une grande majorité des estimations de valeur ajoutée (80 % d'entre elles à chaque session, d'après notre étude), il est difficile de dire, à ce jour, dans quelle mesure les IVAL ministériels reflètent l'action éducative des lycées plutôt que l'effet du hasard.

L'ensemble de ces résultats permet donc d'affirmer que les indicateurs de performance français nécessitent sans doute une révision de leur stratégie d'estimation, qui puisse offrir des indicateurs plus informatifs quant à l'action réelle des lycées. En vue de pallier le bruit associé à l'erreur d'échantillonnage, nous avons ainsi proposé deux nouvelles stratégies d'estimation, l'une augmentant la méthodologie ministérielle des IVAL (stratégie IVAL brut associée à un test de significativité), l'autre adoptant la méthodologie recommandée au niveau international pour ce genre d'indicateurs (stratégie IVAL corrigé). Dans la présente étude, la première stratégie permet de tester si les écarts entre taux prédit et taux constaté de performance peuvent être attribués avec quelque confiance à l'action éducative des lycées ou, au contraire, à une contamination par l'effet du hasard. Allant plus loin, la seconde stratégie permet de neutraliser l'effet du hasard par un critère de fidélité dépendant de la taille des lycées, et de parvenir ainsi à des valeurs ajoutées exploitables pour l'ensemble des lycées. La comparaison des deux stratégies montre que celles-ci convergent pour identifier les lycées à valeurs ajoutées positives ou négatives, mais qu'elles diffèrent par rapport à la quantification de ces valeurs ajoutées, surestimées dans la première stratégie (*i.e.*, d'un facteur 2 par rapport à la seconde stratégie). Au total, si la stratégie de l'IVAL brut s'aligne mieux avec la tradition ministérielle, la stratégie de l'IVAL corrigé est plus performante en termes de champ (tous les lycées obtiennent une valeur ajoutée exploitable) et de précision (quantification plus réaliste de l'action éducative).

Les résultats présentés dans cette recherche contribuent donc à renforcer la manière de mesurer l'action éducative des lycées en France, en prenant en compte les effets du hasard. Ce faisant, ils pourraient participer à la réduction du « bruit » produit par le système d'information que représentent les IVAL, et participer ainsi à une meilleure

interprétation tant pour les équipes éducatives que pour les cadres de l'Éducation nationale et les familles (Kahneman *et al.*, 2021).

Un renforcement plus substantiel consisterait néanmoins à interroger les objectifs institutionnels visés par les indicateurs de performance. Par exemple, les IVAL mesurent le pouvoir qualifiant des établissements, mais ils ne nous informent par sur leur capacité à réduire les inégalités scolaires. Des indicateurs prometteurs ont été proposés pour mesurer ces effets égalitaires au niveau des lycées (Givord et Suarez, 2019), qui pourraient être peaufinés par l'intégration des stratégies d'estimation proposées ici. On peut également s'interroger sur la pertinence de l'accès au baccalauréat comme unique dimension éducative visée par les IVAL. D'un côté, l'obtention du baccalauréat atteste de compétences cognitives et non cognitives devenues nécessaires pour la bonne insertion sociale des jeunes (e.g., acquisitions scolaires, respect du règlement intérieur, intérêt pour les apprentissages, assiduité en cours), le « signal » négatif de l'absence de diplôme les maintenant durablement éloignés de l'emploi (Moncel, 2007). Mais, de l'autre, le diplôme du baccalauréat ne suffit pas à garantir l'accès à un emploi dans un contexte de massification scolaire (e.g., 80 % d'une génération accède au baccalauréat en France ; Núñez-Regueiro, 2018). L'insertion professionnelle, qui constitue l'une des missions incombant aux systèmes éducatifs, n'est donc qu'imparfaitement évaluée à travers les IVAL. D'autres indicateurs mesurant l'action éducative sur l'insertion professionnelle seraient donc nécessaires pour compléter l'évaluation des lycées, mais cela supposerait de surmonter des difficultés techniques importantes (e.g., suivi des élèves depuis le lycée jusqu'à leur sortie de formation initiale) et de prendre en compte, là encore, les inégalités d'accès à cette insertion professionnelle, notamment à travers les mécanismes de sélection dans l'enseignement supérieur (Duru-Bellat et Kieffer, 2008).

En somme, les interrogations autour des types d'IVAL et de leur utilisation sont nombreuses et demandent sans doute une réflexion plus large sur les objectifs visés par le système éducatif, que nous n'avons pas abordés ici en détails. Ce travail s'est attelé plutôt à traiter l'autre question épineuse des IVAL, à savoir la qualité de leur mesure, en prenant pour exemple les IVAL ministériels. Nous espérons que des travaux futurs pourront concilier les deux facettes du problème en proposant de nouveaux indicateurs de performance mesurant, de manière fiable, la réduction des inégalités sociales en contexte scolaire et, pourquoi pas, en contexte professionnel.

Fernando NÚÑEZ-REGUEIRO
fernando.nunez-regueiro@univ-grenoble-alpes.fr

Pascal BRESSOUX
pascal.bressoux@univ-grenoble-alpes.fr

Laboratoire de recherche sur les apprentissages en contexte (LaRAC)
Université Grenoble Alpes
1251, avenue Centrale
38400 Saint-Martin-d'Hères

ANNEXE

TABLEAU A1. **– *Modèles de prédiction de réussite au baccalauréat***

	Estimation	Erreur type	z	p-valeur
Effets fixes				
Constante	5,484	1,071	5,122	< ,001
Age	- 0,426	0,045	- 9,383	< ,001
Sexe (réf. Féminin) : Masculin	0,300	0,058	5,139	< ,001
Origine sociale (réf. : Défavorisée)				
Moyenne	0,039	0,073	0,539	,590
Favorisée B	0,196	0,090	2,193	,028
Favorisée A	0,204	0,090	2,252	,024
Boursier (réf. Oui) : Non	- 0,222	0,073	- 3,043	,002
DNB (réf. Refusé ou absent)				
Admis sans mention	0,451	0,209	2,155	,031
Admis mention assez bien	1,293	0,213	6,075	< ,001
Admis mention bien	2,771	0,237	11,701	< ,001
Admis mention très bien	4,106	0,382	10,763	< ,001
Niveau moyen d'origine sociale	0,526	0,264	1,996	,046
% Elèves non boursiers	0,043	0,007	5,988	< ,001
Niveau moyen au DNB	- 0,540	0,231	- 2,334	,020
Effets aléatoires				
Constantes des lycées (variance)	0,207			
Spécificité des prédictions	,765			
Sensibilité des prédictions	,712			
Aire sous la courbe (AUC)	,813			

Champ : Elèves présents aux épreuves du baccalauréat 2016 et inscrits dans un lycée public ou privé sous contrat de l'académie de Grenoble, 2015-2016.
Source : DEPP, base FAERE.
Note : N = 23 116 lycéens, 112 lycées. Les coefficients rapportés correspondent au logarithme des rapports de chances (log odds ratio).

REFERENCES BIBLIOGRAPHIQUES

AGRESTI A., 2002, *Categorical Data Analysis*, Hoboken (NJ), John Wiley & Sons Inc.

AGRESTI A., FINLAY B., 2008, *Statistical Methods for the Social Sciences*, Upper Saddle River (NJ), Pearson Prentice-Hall [4e éd.].

BRESSOUX P., 1994, « Note de synthèse [Les recherches sur les effets-écoles et les effets-maîtres] », *Revue française de pédagogie*, *108*, 1, p. 91-137.

BRESSOUX P., 2010, *Modélisation statistique appliquée aux sciences sociales*, Bruxelles, De Boeck Université [2e éd.].

BROOKOVER W., SCHWEITZER J., BEADY C., FLOOD P., WISENBAKER J., 1979, *School Social Systems and Student Achievement: Schools Can Make a Difference*, New York, Praeger.

CARO P., HILLAU B., 1997, « La logique dominante des publics scolaires. Offre de formation et environnement local », *Formation emploi*, *59*, p. 87-103.

CODE DE L'EDUCATION, « Titre II : Objectifs et missions du service public de l'enseignement (Articles L121-1 à L123-9) ».

COLEMAN J. S., CAMPBELL E., HOBSON C., MCPARTLAND J., MOOD A., WEINFELD F., 1966, *Equality of Educational Opportunity Study*, Washington (DC), United States Department of Health, Education, and Welfare.

CUNHA J. M., MILLER T., 2014, « Measuring Value-Added in Higher Education: Possibilities and Limitations in the Use of Administrative Data », *Economics of Education Review*, *42*, p. 64-77.

DUCLOS M., MURAT F., 2014, « Comment évaluer la performance des lycées », *Éducation & formations*, 85, p. 73-84.

DURU-BELLAT M., 2002, « Note critique de Thrupp (Martin), *School Making a Difference. Let's be realistic!* », *Revue française de pédagogie*, 139, p. 173-175.

DURU-BELLAT M., KIEFFER A., 2008, « Du baccalauréat à l'enseignement supérieur en France : déplacement et recomposition des inégalités », *Population*, *63*, 1, p. 123-157.

DURU-BELLAT M., LE BASTARD-LANDRIER S., PIQUEE C., SUCHAUT B., 2004, « Tonalité sociale du contexte et expérience scolaire des élèves au lycée et à l'école primaire », *Revue française de sociologie*, *45*, 3, p. 441-468.

EVAIN F., EVRARD L., 2017, « Une meilleure mesure de la performance des lycées : refonte de la méthodologie des IVAL (session 2015) », *Education & formations*, 94, p. 91-116.

FELOUZIS G., 2005, « Performances et "valeur ajoutée" des lycées : le marché scolaire fait des différences », *Revue française de sociologie*, *46*, 1, p. 3-36.

GIVORD P., SUAREZ M., 2019, « Excellence for All? Heterogeneity in High-schools' Value-Added », Insee, *Document de travail, N° G2019/14.*

GOLDHABER D., HARRIS D. N., LOEB S., MCCAFFREY D. F., RAUDENBUSH S. W., 2015, « Carnegie Knowledge Network Concluding Recommendations », Stanford (CA), Carnegie Foundation for the Advancement of Teaching.

GOLDSTEIN H., SPIEGELHALTER D. J., 1996, « League Tables and their Limitations: Statistical Issues in Comparisons of Institutional Performance », *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *159*, 3, p. 385-409.

KAHNEMAN D., 2011, *Thinking, Fast and Slow*, New York, Farrar, Straus and Giroux.

KAHNEMAN D., SIBONY O., SUNSTEIN C. R., 2021, *Noise: A Flaw in Human Judgment*, New York, Little, Brown Spark.

KHOUAJA E.-M., MOULLET S., 2016, « Le rôle des caractéristiques des établissements dans le décrochage scolaire », *Formation emploi*, 134, p. 7□26.

KNOWLES J. E., FREDERICK C., 2019, *merTools: Tools for Analyzing Mixed Effect Regression Models*, version R package version 0.5.0.

LI M., 2007, « Bayesian Proportional Hazard Analysis of the Timing of High School Dropout Decisions », *Econometric Reviews*, *26*, 5, p. 529-556.

MEN-DEP, 1994, « Trois indicateurs de performance des lycées. Tome 1 : Baccalauréat général et technologique 1993 », *Les dossiers d'Education et formations*, 41.

MENJS-DEPP, 2020, *Méthodologie des indicateurs de résultats des lycées*, DEPP.

MERLE P., 2011, « Concurrence et spécialisation des établissements scolaires. Une modélisation de la transformation du recrutement social des secteurs d'enseignement public et privé », *Revue française de sociologie*, *52*, 1, p. 133-169.

MEYER R. H., 1997, « Value-Added Indicators of School Performance: A Primer », *Economics of Education Review*, *16*, 3, p. 283-301.

MONCEL N., 2007, « Recent Trends in Education and Labor Market Policy for School-to-Work Transition of Secondary Education School Leavers in France », *The Japan Institute for Labor Policy Training Report*, 5, p. 39-59.

NÚÑEZ-REGUEIRO F., 2018, *Le décrochage scolaire au lycée : analyse des effets du processus de stress et de l'orientation scolaire, et des profils de décrocheurs*, Thèse de doctorat, Grenoble, France, Université Grenoble Alpes.

NÚÑEZ-REGUEIRO F., BRESSOUX P., LARBAUD J.-C., 2022, « Elaboration d'un indicateur de valeur ajoutée des lycées (IVAL) dans la lutte contre le décrochage scolaire », *Revue française de pédagogie, 216*.

PERRY T., 2016, *The Validity, Interpretation and Use of School Value-Added Measures*, Doctoral dissertation, University of Birmingham.

RAUDENBUSH S. W., 2014, « What Do We Know about the Long-Term Impacts of Teacher Value-Added? », Stanford (CA), Carnegie Foundation for the Advancement of Teaching.

RAUDENBUSH S. W., BRYK A. S., 2002, *Hierarchical Linear Models: Applications and Data Analysis Methods*, London, Sage.

RAUDENBUSH S. W., WILLMS J. D., 1995, « The Estimation of School Effects », *Journal of Educational and Behavioral Statistics*, *20*, 4, p. 307-335.

RUMBERGER R. W., PALARDY G., 2005, « Does Segregation Still Matter? The Impact of Student Composition on Academic Achievement in High School », *The Teachers College Record*, *107*, 9, p. 1999-2045.

RUMBERGER R. W., THOMAS S. L., 2000, « The Distribution of Dropout and Turnover Rates among Urban and Suburban High Schools », *Sociology of Education*, *73*, 1, p. 39-67.

RUTTER M., MAUGHAN B., MORTIMORE P., HOUSTON J., 1979, *Fifteen Thousand Hours: Secondary Schools and Their Effects on Children*, Cambridge (MA), Harvard University Press.

SESPAG, 2019, « Résultats définitifs de la session 2018 du baccalauréat - Académie de Grenoble », *Fiche Stats*, Service d'études statistiques, de la performance, et de l'analyse de gestion, Rectorat de l'académie de Grenoble.

SNIJDERS T., BOSKER R., 1999, *Multilevel Analysis: An Introduction to Basic and Applied Multilevel Analysis*, London, Sage.

THRUPP M., 1995, « The School Mix Effect: The History of an Enduring Problem in Educational Research, Policy and Practice », *British Journal of Sociology of Education*, *16*, 2, p. 183-203.

ABSTRACT

Assessing the Educational Impact of High Schools: When "Noise" Interferes with Evidence

Each year in France, the ministry of Education publishes value-added indicators of high schools (*IVAL*) that seek to measure the effectiveness of educational practices. Regarded as reliable managerial tools, these IVAL play an important role in school educators' decision-making processes. Yet, IVAL are limited because they do not account for sampling variance, which confounds high school effects with statistical "noise" (i.e., chance effects). Retracing the history of these indicators—from their origins in school effectiveness research, to their dissemination in France—, this study shows why this limitation poses a measurement problem that can alter a reliable evaluation of school effectiveness. To address this issue, alternative estimation strategies are proposed. The relevance of this critique is illustrated in a study of 112 high schools from the academic region of Grenoble and of their educational effectiveness in facilitating graduation from high school.

Keywords. *IVAL* – VALUE-ADDED – ASSESSMENT – HIGH SCHOOL – GRADUATION – MULTILEVEL ANALYSIS