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

Pauline Givord* and Milena Suarez Castillo**

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.

JEL Classification: I20, C21, C50
Keywords: quantile regression, school value-added, student growth percentile

**INSEE-LIEPP (pauline.givord@travail.gouv.fr); **CREST-INSEE (milena.suarez-castillo@insee.fr)*

We would like to thank Fabrice Murat, Cédric Afsa, Caroline Simonis-Sueur and Fabienne Rosenwald for authorising access to the data, as well as Olivier Monso, Thierry Rocher, Franck Evain and Laetitia Evrard for hosting us at DEPP and for their advice. We would like to thank participants in INSEE's seminars, in particular Marco Paccagnella, for their stimulating discussion of an earlier version of this article and two anonymous reviewers for their careful and accurate remarks which helped improve it. The authors remain solely responsible for any remaining errors or approximations.

Received in October 2020, accepted in May 2021. Translated from: "Qu'est-ce qu'un bon lycée ? Mesurer les effets établissements, au-delà de la moyenne".

The opinions and analyses presented in this article are those of the author(s) and do not necessarily reflect their institutions' or Insee's views.

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

What makes a good high school? In recent decades, a large number of research projects have focused on measuring the effect of schools on the success of students with a view to improving the information available, accompanying a growing demand and interest in evaluating this subject. In France, the evaluation of the education system has long been identified, and regularly reaffirmed, as key to improving the quality of the public education service, from the early thoughts on the need for a culture of evaluation (Thélot, 1994a; 1994b) to the creation, in 2019, of the *Conseil d'évaluation de l'école* (School Evaluation Council). In the case of high schools, this has resulted in particular in the regular publication of *Indicateurs de Valeur Ajoutée des Lycées* (Value-Added indicators for high schools, or IVALs) by the *Direction de l'évaluation, de la prospective et de la performance* (DEPP, the statistical office of the Ministry of National Education). The IVALs provide a set of indicators on the performance of French high schools in terms of the success of their students in the baccalaureate, but also with regard to their ability to support them right through to their final examination, taking account in particular of the profile of the students taught (see Box 1). In the United States, the evaluation of schools, primarily on the basis of quantitative criteria, dates back much further, to the development of a results-based culture in the 1980s and the idea of handing the responsibility for the success of students to the schools themselves (with a view in particular to improving school selection). The most symbolic example of this development was the adoption of the "No Child Left Behind" federal law in the United States in 2001, which required states to subject all students to annual tests and offered strong incentives for schools to meet student success targets. In order to meet these objectives, the majority of states developed measurement tools for schools and even teachers.

Such assessments can serve at least two main purposes, which raise various measurement issues. The first, which underlies the development of measures of this type in the United States, aims to provide the public authorities in charge of managing schools with instruments for assessing their effectiveness and efficiency. This could, for example, be a case of comparing the good results achieved by a high school (or, conversely, the disappointing results) with the practices and resources implemented. As pointed out by Raudenbush & Wilms (1995), for example, this objective is especially complex, since schools have no control over some of the elements that can influence the success of their students, such as the impact that other students have on individual achievements. Such "peer effects" on success are complex and, above all, very difficult to measure (for a recent overview, see e.g. Monso *et al.*, 2019). It is therefore generally impossible to distinguish between the elements of the measurement of the school's effect on student success that are the result of the actions of the school and those that result from interactions between the students. However, measures of high school effects can be useful for fulfilling a second, more modest purpose, which is to provide families with an indication of the expected effect of attending one school over another, whether that effect is related to the school's practices or the contextual effects linked to the interactions with other students.

Even if we limit ourselves to this objective of providing information to families, it is difficult to identify relevant measurement tools. Firstly, because parents may have different criteria for what makes a good high school. Of course, for the majority of parents, a good school is one that is capable of providing their children with support right through to the baccalaureate, guaranteeing them a problem-free education, while also ensuring that they are as well prepared as possible for the future. Nevertheless, the assessment of how well a school meets these objectives could vary depending on students. Some teenagers may thrive in schools that encourage competition and academic excellence, while others may suffer if faced with an overly competitive atmosphere. Beyond pure academic performance, some parents may value the ability of teachers to instil a taste for learning and self-confidence in their students, the quality of the atmosphere within the school or the assistance provided to students in establishing their future direction and making it a reality.

Regardless of what defines a good high school, identifying a school that meets the criteria is even more complex. It would require to determine what a student's education would have been like at a school other than the one they attended, which is difficult if not impossible. In general, parents have little to go on when judging a school. Past experiences of acquaintances or siblings and the baccalaureate pass rate are certainly useful information, but they only provide indirect information with regard to how a particular student will ultimately adapt to a high school. The success demonstrated by a high school is first and foremost a reflection of the characteristics of the students who are educated there, but not accounting for selection effects can give a distorted picture of the quality of schools and therefore provide information that is of little relevance for families. That is why indicators, such as those developed by the DEPP, take account of students' starting levels.

The most frequently used indicators focus on average effects. However, such averages may mask disparities: the same average effect could be measured for a high school that helps all of its pupils to make a small amount of progress and another that helps a small minority of students to make significant progress. The degree to which the information provided by the indicator is relevant will vary, particularly for those parents who would use these measures to enrol their children in the high school offering the best education. This article therefore aims to enrich the description of the high school effect by providing indicators that aim to characterise high schools based on their propensity to amplify or, conversely, reduce, inequalities in baccalaureate examination performance when compared with what is expected in view of the characteristics of the students.[1] The remainder of the article starts by proposing a review of the extensive literature on high school effects measurement before going on to detail the approach used here for French high schools in the general and technological streams, based on the baccalaureate results achieved in 2015.

[1] *A full evaluation of a school, which would require information on the financial resources, exceeds the scope of this article, which seeks to measure the effect that a school has on improving the academic success of its students.*

Box 1 – **Value-Added Indicators for High Schools (IVAL)**

Since 1993, IVALs have been disseminated by the statistical office, now the DEPP, of the Ministry of National Education (for a detailed presentation, see Evain, 2020).

While the methodology used for their construction has changed over time, their aim is to allow comparisons to be made between schools, taking account of the initial differences between the students that they educate. The "value-added" of high schools is highlighted by comparing the expectations given the characteristics of their students (particularly in terms of their educational level and social background), as predicted by a model, and the student results actually observed within that school.

To take account of the difficulty of evaluating the action of a school on the basis of a single indicator, several indicators are proposed. The first looks at the probability that a student enrolled at the school will pass the baccalaureate exam: this is the indicator that most directly resembles the rankings published by the media, but in this case it also takes account of the initial composition of the schools.

This indicator of successfully passing the baccalaureate is supplemented by the probability of passing the exam having attended the school since year 11 or 12, the "access rate". Analysis of the value-added of the access rates makes it possible to avoid overvaluing schools with a "skimming" policy by means of which they select the best students as they progress through high school: these schools may have very good final examination results, but at the expense of the less promising students who find themselves being dropped. Conversely, increased value-added for the access rate reflects the school's ability to support its students throughout their schooling.[i]

Finally, since 2017, the value-added has also been calculated for the probability of achieving a distinction in the exam. This makes it possible to better account for the disparities between the different levels at which the students find themselves, going beyond the mere fact of passing the exam. Indeed, the baccalaureate success rate has become fairly indistinguishable given the very high levels observed, particularly in the general and technological streams: in the 2019 baccalaureate session, the average pass rate was 91% for the general stream, 88% for the technological stream and 82% for the vocational stream. Looking at the probability of achieving a distinction (i.e. having obtained an average of at least 12/20 in the examination) makes it possible to draw finer distinctions between schools.

In practice, value-added is calculated on the basis of the logistic modelling of the probability of passing the examination, using random effects modelling to take account of the high school effects (for details, see Duclos & Murat, 2014 and Evrard & Evain, 2017). The model incorporates individual student variables: academic level, social position index,[ii] age and gender.[iii] The correlations observed between these individual characteristics are used to estimate the probability of students passing their examinations predicted by the model, which when aggregated at the high school level, allows for the calculation of an "expected" pass rate. The value-added corresponds to the difference between the observed rate and the expected rate.

[i] *However, one limitation presented by the indicator measuring the access rate is that it does not allow distinguishing between moves made voluntarily by the students and specific practices implemented by the schools.*
[ii] *This index is a synthetic measure of the social, economic, and cultural dimensions associated with school success, by parental occupation and social class (Rocher, 2016).*
[iii] *In addition, the means of these variables are added to the model (see the discussion in Box 2), which allows to account for the fact that these estimates of individual variables may be biased if they are correlated with the unobserved characteristics of the high school.*

1. Measuring the Effectiveness of a School or a Teacher: Methodological Issues and Challenges for Interpretation

1.1. Selection Effects Make it Difficult to Measure the Effects of Individual Schools or Teachers

One of the key difficulties in measuring the ability of a school or a teacher to help their students to progress is the existence of significant selection effects (Felouzis, 2005). For example, a high school that selects its students on the basis of their academic record at entry (the *classe de seconde*, which is the first year at *lycée* – equivalent to the 10th grade in the US, and year 11 in UK) will obviously have a very high pass rate for the baccalaureate. This does not mean that it can be credited with making any particular effort to help its students to progress. It also does not mean that any student who is educated in such a school would be guaranteed to achieve equally good results, regardless of their starting level. Generally speaking, schools do not educate the same students, and students do not have the same teachers within schools. The apparent success of some may simply reflect differences in the initial level of their students. These same questions arise if the aim is to measure "teacher effect" or, in other words, to assess the extent to which a teacher's actions could influence their students' outcomes, either positively or negatively. These are central issues within school systems that have institutionalised performance-related pay, as is the case in certain US states. For this reason, a large body of literature has focused in particular on the issue of measuring teacher effects (see, for example, Chetty *et al.*, 2014). Although the underlying factors determining the effect of schools or teachers are clearly different, both raise identical methodological issues from a statistical point of view.

In order to compare two teachers or two high schools, you would ideally want to compare their ability to help the same types of students to progress. Measuring the specific effect of a school would, in theory, require the ability to randomly assign students with identical profiles to high schools and classes; however, the feasibility of such an exercise is very limited for both practical and ethical reasons. Most models developed to measure the effect of schools aim to reduce the biases linked to the effects of different school or class compositions by controlling for the initial level of students. Two main types of models have been developed within this framework: Value-Added models and Student Growth Percentile models.

1.1.1. Two Statistical Models: Value-Added Models and Student Growth Percentile Models

In their simplest form, Value-Added models (VAMs) assume that the variable of interest (e.g. the average baccalaureate examination results) depends on the results achieved previously by each student, a certain number of observable characteristics, such as their initial level or background, and an effect specific to the school. The latter is captured by introducing an indicator that is common to all of the students at the school. This type of model is used by DEPP to measure the value-added of high schools for a set of indicators, including in particular the probability of passing the baccalaureate or of earning a distinction, as well as the probability of a student who has completed their entire education within the school of passing the baccalaureate (cf. Box 1).

Student Growth Percentile (SGP) models were most notably developed by the US State of Colorado (Betebenner, 2007), followed by 18 other US states, while VAMs are used in 15 states, having been pioneered by Tennessee (see Kurtz, 2018 for a review). SGP models offer the advantage, for operational use, of being fairly simple to interpret. Their principle is based on the following question: how well did a student perform compared with students who had achieved comparable results in previous tests? Students are "ranked" according to their test results, with their position in this ranking being represented by the percentile in the grade distribution. For example, if a student performs better in an end-of-year test than 80% of students who were at a similar level to them at the start of the year, a positive effect of 80 is attributed to the high school for that student. The effectiveness of the school (or the teacher) will then correspond to the mean (or the median) of these effects, measured across all of the students enrolled at the school (or taught by the teacher). In practice, these estimates are derived from quantile regressions, which allow the distribution of test scores to be modelled conditionally based on the results of previous tests (see Box 2).

1.1.2. Statistical Limitations of the Two Models

The measurement of school effects and teacher effects has been the subject of intense methodological research. This level of interest can be explained by the high stakes that may be associated with these indicators. While the perceived quality of schools can be an important factor in the decision by parents as to which school they enrol their children in, the publication of "league tables" can contribute to widening the initial gaps – particularly as the parents who are better informed or who have the means to choose the school at which their children will be educated often have greater academic capital. More radically, these methods are sometimes used, for example in the United Kingdom and certain US states, to measure the "effectiveness" of schools or teachers, with consequences that can be significant for those being evaluated: financial incentives for teachers based on their performance or closure of schools – or dismissal of teachers – whose effectiveness is assessed as inadequate.[2] Given the high stakes involved for those concerned, it is crucial that the instruments used are valid and relevant.[3] However, the tools available attract criticism from several sides.

Firstly, the majority of contributions highlight the difficulty that these models face in overcoming the limitations associated in particular with the absence of randomisation (for a summary, see, for example, Everson, 2016). In particular, the measurement of teacher effects or school effects is extremely sensitive to the variables used to control for composition effects. Failure of the models to take account of some of the characteristics of students that may influence their academic progress, such as their social background, significantly reduces the ability of these models to differentiate between effective teaching and teaching students from backgrounds that are more conducive to academic success. SGP models, which are used routinely, do not take account of these dimensions and are often criticised for this (Guarino *et al.*, 2015a). The various comparisons suggest that these indicators tend to penalise teachers dealing with students from disadvantaged social backgrounds or with special needs when compared with VAMs, which take these dimensions into account (Walsh & Isenberg, 2015). Since the information that would be needed is not always available, this issue also arises for VAMs. The type of variables used to control for composition effects within this other model type can also affect the conclusions that can be drawn from it (Ehlert *et al.*, 2014; Sass *et al.*, 2014), as can the statistical specification used (Guarino *et al.*, 2015b; Soland, 2016). In addition, as was discussed in the introduction, some of the effects that composition has on success stem from the interactions between pupils, which are especially difficult to measure (for an example of a measurement of this in French high schools, see Boutchenik & Maillard, 2019), and the respective effect of which is generally impossible to separate from the school effect.

More generally, certain authors are highly sceptical of the possibility of reducing selection biases, which are linked in particular to the fact that the characteristics of the students and the teachers that educate them are not independent of one another (Rothstein, 2010; Sass *et al.*, 2014), although others have more confidence in the possibility of relying on factors such as the mobility of teachers between schools and between classes to evaluate these effects (Chetty *et al.*, 2014; Koedel *et al.*, 2015). In addition, the effects measured by these models can be very imprecise, especially since they are estimated on the basis of a small number of observations. A recent study observed, for example, that these models can be used to identify pseudo teacher effects on elements such as the height of their students, a characteristic that is not likely to be altered by teaching practices (Bitler *et al.*, 2019). The authors demonstrate that this paradoxical finding can be explained by the small sample sizes from which the estimates were derived, which leads to factors being incorrectly attributed to the teacher that are nothing more than statistical "noise". Although this effect disappears when the observations used are gathered over several years, this solution is not always used when assessing, for example, the value-added of teachers.

1.2. Back to the Question: Can What Makes a Good High School be Measured?

Looking beyond these methodological issues, the use of instruments of this type to evaluate teachers has also been criticised for the fact that it tends to focus on what we know how to measure best (student success in academic tests) to the detriment of more fundamental competencies, such as the ability of teachers to instil self-confidence, the desire to learn or critical thinking in their students, dimensions that only partially overlap with cognitive competencies. For example, an American study randomly assigned students to classes as part of a randomised study aimed at comparing teacher effects on standardised test scores with those obtained *via*

[2] *One of these was the No Child Left Behind act mentioned in the introduction, which required all public schools to demonstrate "adequate annual progress" in the performance of their students, as measured by yearly tests, with a set of sanctions and incentives in the event that this was not achieved. Repeated failure to meet these targets for six consecutive years would lead to the establishment of a plan for the complete restructuring of the school, which could go as far as its closure, the dismissal of all of its staff or its conversion into a charter school (see Gamoran, 2012 for an explanation). This law was repealed in 2015.*

[3] *For an example of a critique of these practices, particularly in view of the inherent limitations of the underlying measurement, see Jacob (2005).*

open-ended questions, or on the effort put in by students and their motivation. It found that the correlation between these various dimensions is very weak (Kraft, 2019). Another study also demonstrated that the effect that teachers have on the success of their students during tests shows very little correlation with the effect that they have on the behaviour of those students (for example, absenteeism or having to repeat a year), even though these are the dimensions that are better suited to predicting the future success of students (Jackson, 2018).

In addition, since there are high stakes associated with the assessments – which is the case in particular where they are linked to financial schemes (performance bonuses) for teachers, or simply to the reputation of a school, which is of importance for the quality of the students it will educate in the future – they can induce strategic behaviour on the part of the people concerned, which can have the opposite effect to that intended (for a recent contribution, see Fryer, 2013 and for a review, see Jacob, 2005). In particular, there are frequent attempts to manipulate the indicators. This could involve devoting a disproportionate amount of teaching time to preparing students for the tests (the "teaching to the test" phenomenon, see Wall, 2000). SGP models are a priori less likely to bring about such cramming phenomena (Barlevy & Neal, 2002), since the measurement of the school or teacher effect is based on a relative metric (the progress made by students when compared with those with the same initial level), whereas value-added models require the use of standardised tests, the format and content of which are subject to little variation, to allow reliable and fair comparisons to be made over time. Nevertheless, the sensitivity of these two models to student characteristics other than their initial level could compel the schools or the teachers being evaluated by this measure to take steps to minimise risk. For example, schools can select the most promising students or exclude those who are not achieving adequate results. When they have a choice in where they are assigned, teachers tend to avoid schools with the highest proportions of disadvantaged students (Walsh & Isenberg, 2015), which means that it is often the teachers who have no choice in this regard (often the least qualified or the most inexperienced) who find themselves teaching the students with the greatest needs.

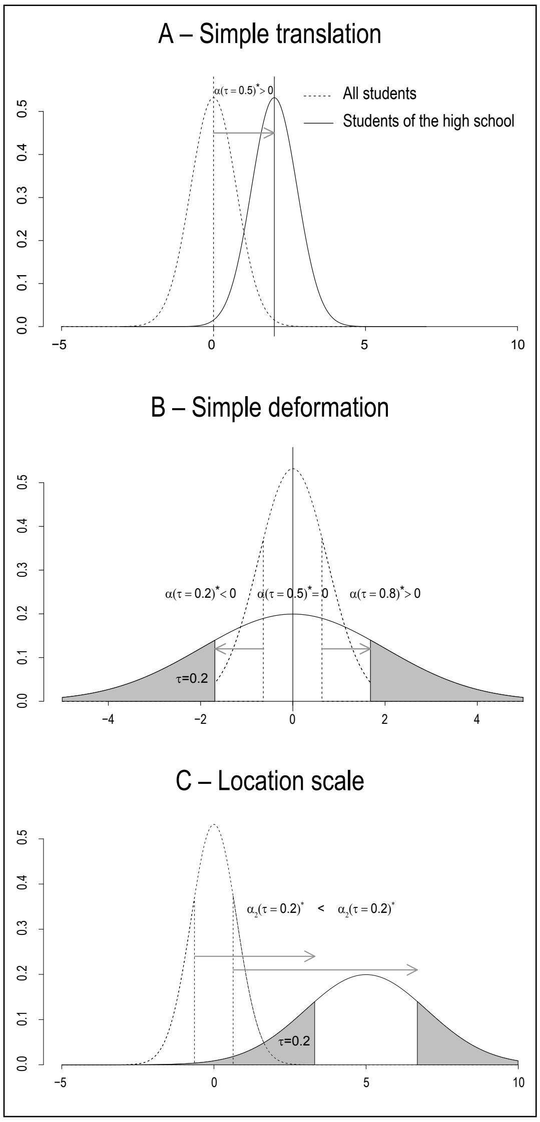
2. Measuring the Dispersion beyond the Average

Even when considering only academic performance indicators, the quality of the schools can be questioned beyond the traditionally measured average effects. An apparent similarity between two average effects could mask very different realities: the same positive average effect could result from the actions of a school in which all of the students are making progress or one in which only a minority of the students are performing extremely well, while the rest are performing much more poorly than expected in view of their characteristics.

This article therefore aims to enrich the description that can be made using traditional means of measuring high school effect by providing indicators that look beyond the average. The aim here is not only to measure the effect of high schools on the average distribution of grades, but also to assess the extent to which a high school tends to have more dispersed or homogeneous baccalaureate results when compared with identical high schools that are similar in terms of the characteristics of the students enrolled. The intuition is illustrated in Figure I, based on a fictitious example representing the theoretical densities of grades, i.e. as expected based on student characteristics, and taking account of the effect of the high school, in three separate cases. The first (Figure I-A) shows a situation in which the high school effect is the same for all students: when compared with the expected distribution of grades, the observed distribution in this school shifts slightly to the right if the effect is positive and slightly to the left if it is negative, but the shape remains unchanged. The second case (Figure I-B) represents the opposite situation, i.e. one in which the high school has a very different effect depending on its students: the weakest students obtain lower grades than expected and the strongest students achieve higher grades than expected. In this fictitious case, the effect is completely symmetrical and there is therefore no impact on average grades (the estimated average effect will be zero); however, the dispersion of the grades observed is much wider. Finally, the third case (Figure I-C) is a combination of the two previous figures: the effect of the high school is positive on average and also tends to increase the dispersion of the grades.

The intention here is to model the effect of the high school at various levels of the distribution of grades within the high school. This is done using a statistical technique known as quantile regression, which is explained in Box 2. Modelling in this way allows us to look beyond this fictitious case, which assumes that the effects are perfectly symmetrical (greater success at the top end of the distribution is "paid

Figure I – **Illustration of the effects of a high school on the dispersion and the mean distribution of grades (scale transposition model)**



A – Simple translation

B – Simple deformation

C – Location scale

for" by lesser success at the bottom end of the distribution): indeed, we will model the top and bottom ends of the distribution of grades within the high school separately, without assuming that the effects are symmetrical. Comparing the effects at the top and bottom ends of the distribution also makes it possible to estimate the extent to which certain schools are able to amplify or reduce the dispersion of the grades achieved by their students when compared with what is expected given their characteristics. The aim is therefore to observe whether certain high schools are able to achieve more homogeneous results or, conversely, more unequal results, than high schools in which the initial characteristics of the students (including in terms of their educational level upon completing middle school) are similar.

The statistical method used here is therefore a hybrid model combining elements of the SGP and VA models. Like the former, it is based on the modelling of the high school effect on the distribution of grades on the basis of quantile regressions, but, like the latter, it takes account of all of the observable characteristics of the students, in particular their initial level and their social background, with a view to trying to reduce selection bias to the greatest possible extent.

In order to estimate the effects specific to each high school, beyond the effects linked to their initial composition, indicators are introduced into the model for each high school, with a standardisation condition. This method, which is traditionally referred to as "fixed effects models" in the econometric literature, offers the advantage of requiring very few assumptions with regard, on the one hand, to the distribution of these fixed effects (they can differ greatly across high schools, without any particular form being specified for these differences) and, on the other hand, to the possible links between these high school effects and the characteristics of the students on whom we wish to measure the effects. More precisely, it is possible to establish an unbiased estimate of the effects that the characteristics of the students have on baccalaureate results, even if the distribution of students within the high schools is based on a combination of these characteristics (e.g. their academic level) and the unobserved characteristics of the high schools. One of the risks posed by this type of model is that the effects can be poorly estimated in schools that only have a small number of students:[4] this is why the analysis is limited here to high schools with an "adequate" number of students (at least 65 in the general stream and 25 in the technological stream, with these thresholds having been selected by striking a balance between not excessively limiting the sample – and its representativeness – and reducing the risk of obtaining biased estimators).

It should be noted that the high school fixed effects "capture" all of the characteristics of high schools: it is therefore impossible to provide an estimate for a single characteristic (such as the seniority of the teachers or the average level of the other students). Moreover, the effect of these variables is generally very difficult to estimate

[4] This problem is especially crucial when modelling non-continuous variables (for example where the variable of interest is passing the baccalaureate rather than the average grade for the baccalaureate), since the poor approximation of fixed effects "contaminates" the estimation of the coefficients that correspond to the individual characteristics of the students.

Box 2 – **Quantile Regression and the Measurement of High School Effects**

Quantile regression is a statistical method of linear regression used to describe how a variable of interest varies on the basis of co-variates (for a detailed description, see, for example, D'Haultfoeuille & Givord, 2014). While standard linear regression models how the mean of the variable of interest varies on the basis of observable variables, quantile regression consists of modelling the quantile of that variable conditional on its observables, with both methods being reliant on a linear approximation. For the quantile $q_\tau(X)$ of order τ (between 0 and 1) of the distribution of the variable of interest Y conditional on co-variates X, the following assumption is therefore made:

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

where β_τ is the coefficient to be estimated.

It is therefore possible, on the basis of multiple quantile regressions, to enrich the description of how one variable is dependent on another by modelling the effect of the co-variates for several percentiles – for example, the first decile, the median and the last decile.

In practice, it has been demonstrated that this coefficient can be obtained without bias by the linear program:

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

where the function $\rho_\tau(.)$ is defined by $\rho_\tau(u) = (\tau - 1_{u<0})u$.

The coefficient $\beta_{j\tau}$ for a variable X_j reflects the way in which the quantile of order τ of the variable of interest varies on the basis of a variation in the variable X_j. If X_j is a continuous variable, the coefficient $\beta_{j\tau}$ is interpreted as the manner in which the quantile varies as a result of a marginal variation of X_j.

In some cases, a linear approximation may be an oversimplified form of the real relationship between the variable of interest and the variables that are observable based on the data and, as with a linear regression, it is possible to use a more complex form, either by means of a transformation (e.g. the logarithm of the variable under consideration), or by a polynomial form in the variable X_j. For the data in this case, it appeared more appropriate to use a quadratic form to approximate the relationship between baccalaureate grades and the grades obtained in the junior secondary education certificate (*Diplôme National du Brevet*, DNB). In practice, this means that, in order to interpret the effect of the DNB grade on the distribution of baccalaureate grades, account must be taken of the two coefficients that correspond to the average baccalaureate grade $\beta_{B\tau}$ and its square, respectively $\beta_{B^2\tau}$. For example, for $\tau = 0.80$, the last quintile for students who passed their DNB with grade N (the level obtained by no more than 20% of students with this initial level) is $\beta_{B\tau} + \beta_{B^2\tau} * (2N+1)$ lower than the quintile of the distribution of baccalaureate grades for students who obtained grade N+1 in the DNB. In the case of this variable, both coefficients are positive for the three quantiles observed. This takes account of the fact that a high DNB grade is a predictor of good baccalaureate results, but also means that those students who performed very well in their DNB can make an even greater difference with the baccalaureate.

It is also possible to compare the effects of the same variable across the various different quantiles. The fact that a variable X has a greater effect on a low grade distribution quantile than on a high grade distribution quantile can be interpreted in terms of the dispersion of the baccalaureate grades. For example, the fact that the coefficient of the indicator that indicates that a student is female is higher for the first quintile than for the last quintile means that girls achieve better results, but with grades that are less dispersed than those of their male counterparts.

It is this type of comparison that is used to interpret the high school effects.

when there are selection effects at play (for example, where the most experienced teachers are more likely to be assigned to the high schools that have the most privileged students, or where the students tend to be grouped by level). So-called "random effects" models, which require the use of a specific distribution (generally normal distribution) to model the effects specific to the schools, make it possible to also estimate coefficients for the variables at the level of the high schools at the same time, together with the effects for each high school (this is the model used by Page *et al.*, 2017, for example). However, where there is a link between the characteristics of the students and the high school effects, the estimated coefficients are likely to be biased (for a general discussion of these types of model in the context of the data used here, see Givord & Guillerm, 2016, for example).[5] This is why fixed-effect models are preferred in this case.

Quantile regressions are used to estimate the fixed effects at the high school level for the weakest students (this level is defined here as the first quintile of the distribution of grades

[5] It is possible to demonstrate that unbiased coefficients can be obtained for the effects brought about by the individual characteristics of the students, provided that the averages of these characteristics, aggregated at the level of the high school, are added into the model (this is known as the "Mundlak regression"). However, this correction does not allow for the correction of possible bias in the variables estimated at the level of the high schools. Therefore, adding the average level of all students in a high school to the score achieved by an individual student allows for an unbiased estimate of the effect that the level of an individual has on success; however, the coefficient obtained for the average cannot be causally interpreted as the effect that the level of these peers has on a student's level (see Castellano et al.*, 2014).*

within the high school, i.e. such that only 20% of students obtain poorer results) and for the strongest students (defined here as the last quintile, i.e. the level at which just 20% of students have a higher level), once account has been taken of their composition, particularly in terms of the initial academic level and the social background of the students.

3. An Application Based on the 2015 Baccalaureate Results

3.1. The Data

We draw upon the comprehensive database of the results of the 2015 national baccalaureate examination. This database provides all of the grades obtained in the various tests, but in this case we use the average grades for the various subjects (weighted by their coefficients in the chosen series), obtained the first time the examination was sat.[6] These results are supplemented by the anonymised files for studies and research (FAERE), which are produced and made available by DEPP. This database, which has been compiled for research purposes on the basis of administrative files that monitor students' education, contains personal information, such as the gender and age of the student, the socio-professional category of their parents, and the schools attended. It also contains the individual results for the junior secondary education certificate (*Diplôme National du Brevet*, DNB), which provide an indicator for the academic level of the student upon starting high school.

Table 1 illustrates the strong compositional effects that the model aims to take into account. It shows the average characteristics of the schools, estimated for three separate groups of schools defined using the average baccalaureate grades obtained by their students. They distinguish between the 20% of high schools with the poorest examination results (352 general high schools and 310 technological high schools), the 20% of high schools with the best examination results and a third group made up of the high schools that fall between these two extremes (1,055 general high schools and 929 technological high schools). By design, this ranking is based on the average grades observed for each school: therefore, while the average grade for all general high schools is 12.2/20, it is just 10.5 for the group of schools with the poorest grades, 12.2 for the middle group and 13.8 for the high schools in the final group. The first observation is that, on average, the vast majority of high schools reproduce the level of their students upon leaving middle school, particularly in the general stream. The average DNB grade for high school students taking the technological baccalaureate was lower, but this same "gradient" is also found in the other direction: the high schools returning the best baccalaureate results are also those that are most likely to educate the students who performed best upon leaving middle school. These differences in performance can also be linked to the socio-economic level of the students, which is one of the most important factors in determining academic success, and the hierarchy of which can be found here. In addition, the high schools that educate the best students on average also have more homogeneous students from a social and academic point of view, as can be seen from the reduced variance in these two indicators for this group of high schools. This means in particular that these high schools are less likely to enrol disadvantaged students, which can be seen from the number of pupils who repeated at least one year during their schooling (referred to as "repeaters" in Table 1 and below). In the general stream, only 3% of the students enrolled in the "best" high schools are repeaters, compared with 11% in the high schools showing the poorest performance.

The estimation of the fixed effects for each high school makes it possible to assess the effects attributable to the schools in the success of their students above and beyond these composition effects. These estimations are made separately for the general and technological streams. In order to reduce the variance in the estimators obtained, the sample is restricted to the schools that had at least 65 students enrolled in 2015 for the general stream and 25 students for the technological stream. These thresholds were selected to retain 95% of students in the two streams and are the result of a compromise. On the one hand, it is a question of keeping enough students per high school to ensure that the pupils who could have highly atypical profiles do not carry too much weight when estimating high school effects. On the other hand, it is important that the overall sample of students remains large enough to not reduce the ability to generalise the results, which could be the case, for example,

[6] *The grades for the first session of the examination correspond to the grades after the harmonisation sessions on marking, but before the catch-up tests. These tests are offered to students whose average score was between 8 and 10, to provide them with an opportunity to repeat an oral examination for certain tests and ultimately increase their average to above 10, which is the score required in order to pass the DNB. For this reason, the distribution of grades after the second session is highly irregular (Givord & Suarez Castillo, 2019), with a significant accumulation point just above 10/20 (this is also the case, but to a lesser extent, for the grades for the first session of the examination) and a mass deficit between 8 and 10. In addition, using the grade from the second session means comparing students' results on two very different scales, since the grades also relate to tests that are not identical for all students.*

Table 1 – **Initial characteristics of high schools by average baccalaureate performance groups**

	General Stream				Technological Stream			
	Total	Lowest 20%	Median group]20,80[	Highest 20%	Total	Lowest 20%	Median group]20,80[	Highest 20%
Number of high schools	1,759	352	1,055	352	1,549	310	929	310
Average baccalaureate grades (1st session)	12.3	10.9	12.2	13.7	11.6	10.5	11.6	12.7
Average DNB grades	12.3	11.2	12.3	13.3	9.7	8.9	9.8	10.5
Average Social Position Index[(a)]	120.7	107.9	120.6	133.6	105.3	95.9	105.5	114.0
Variance in baccalaureate grades (1st session)	6.4	6.6	6.6	5.6	4.4	5.4	4.3	3.8
Variance in DNB grades	4.3	4.7	4.4	3.6	3.1	3.4	3.1	3.0
Variance in the social position index	1,048.7	1,144.6	1,085.6	842.1	975.0	975.1	981.1	956.5
Proportion of students repeating a year (%)	6	11	5	3	18	24	17	14
Proportion of private high schools (%)	26	3	19	70	20	6	16	45

[(a)] see Box 1.
Notes: The high schools are grouped by stream (general and technological) according to the average grades obtained by their students during the first session of the baccalaureate.
Sources: MENJ-DEPP, anonymised files for studies and research (FAERE).

if the students enrolled in "large" high schools differ from those in smaller high schools. Details of how these grades are used can be found in Givord & Suarez Castillo (2019). Individual baccalaureate results are regressed on the basis of the observable individual characteristics of the students: whether they are male or female, their social background,[7] whether they repeated a year during the course of their schooling and results of their final DNB examinations (with a quadratic specification), along with a fixed effect for all students at the same high school. The effect of these variables is estimated at three levels of the baccalaureate grade distribution – first and last quintiles and median. The estimates relate to general and technological high schools, with the two streams being separated. Effects specific to each series (three in the general stream, eight in the technological stream) are also added. They make it possible to take account of the fact that marking practices differ between the various disciplines, the weighting for which differs from one series to the next.

3.2. The DNB Score is the Variable that Best Correlates with Baccalaureate Results

The correlations between the estimated variables and other variables are in line with the results obtained by more conventional means (Table 2). As has already been pointed out by Evain & Evrard (2017) in connection with similar data, there appears to be a high correlation between average baccalaureate grades and average DNB scores. The estimates made here suggest that this dependence is observed at all levels of the distribution, and also that this dependence is non-linear: the quadratic term is positive for the three deciles studied (cf. Box 2). This result can be explained by the fact that the vast majority of very good students generally have very good results upon completing middle school, whereas students with poorer DNB grades can have more variable results.

As regards the impact of repeating a year, the conditional distribution of the baccalaureate results of students who have repeated a year is significantly lower than that of non-repeaters, with the gap being wider at the bottom end of the distribution. Girls generally achieve better results than boys, and their results are also less dispersed, as illustrated by the fact that the "girl" effect is greater at the bottom end than at the top end of the distribution. Unlike the other explanatory variables studied here, the social background (captured by the indicator that looks at the social background of the parents) has an almost identical effect at the three levels of the distribution of baccalaureate grades studied here. Moreover, this is also the only variable in the model for which the correlation with baccalaureate grades is very significantly reduced when high school-specific fixed effects are introduced,

[7] *As captured by DEPP's Social Position Index (cf. Box 1).*

Table 2 – **Impact of explanatory variables on the distribution of average baccalaureate grades (with high school fixed effects)**

	Q20		Q50		Q80	
	Coeff.	Std-E	Coeff.	Std-E	Coeff.	Std-E
General stream (*N*=318,222)						
Mean DNB grade (level)	0.593***	(0.002)	0.632***	(0.002)	0.646***	(0.002)
Mean DNB grade (square)	0.107***	(0.001)	0.105***	(0.001)	0.082***	(0.001)
Social position index	0.079***	(0.002)	0.079***	(0.002)	0.079***	(0.002)
Repeater (*ref.: non-repeater*)	-0.271***	(0.008)	-0.245***	(0.007)	-0.193***	(0.008)
Girl (*ref.: boy*)	0.08***	(0.004)	0.052***	(0.003)	0.032***	(0.004)
L series (*ref.: ES*)	0.074***	(0.005)	0.086***	(0.005)	0.088***	(0.006)
S series	-0.194***	(0.004)	-0.172***	(0.004)	-0.147***	(0.004)
Technological stream[a] (*N*=122,286)						
Mean DNB grade (level)	0.358***	(0.004)	0.392***	(0.003)	0.408***	(0.004)
Mean DNB grade (square)	0.018***	(0.002)	0.025***	(0.002)	0.034***	(0.002)
Social position index	0.034***	(0.004)	0.027***	(0.003)	0.027***	(0.004)
Repeater (*ref.: non-repeater*)	-0.285***	(0.010)	-0.258***	(0.007)	-0.228***	(0.008)
Girl (*ref.: boy*)	0.241***	(0.007)	0.211***	(0.007)	0.189***	(0.008)
ST2S (*ref.: 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] The series of the technological stream, as of 2015, are related to management (STMG), health and welfare (ST2S), laboratory (STL), manufacturing (STI2D), design and applied arts (STD2A), hostelry (HOT).
Notes: Effects of explanatory variables on the results of baccalaureate grades (average of all grades) obtained by quantile regressions for the first quintile (Q20), the median and the last quintile (Q80). Standard errors in brackets: *** significant at 1%;** significant at 5%.
Sources: MENJ-DEPP, FAERE files.

as suggested by the comparison with estimates that do not include these fixed effects (see Givord & Suarez Castillo, 2019). This statistical effect highlights the significant differences in social intake from one high school to the next.

Finally, large gaps can be seen in the distribution of grades between streams. These gaps can be explained by differences in grading for the dominant subjects in each stream, as well as by compositional effects. It can therefore be observed that students in the S series[8] obtain lower average baccalaureate grades than those observed for the two other series in the general stream, once account has been taken of the initial level of the students and their other individual characteristics.[9]

3.3. Widely Dispersed School Effects

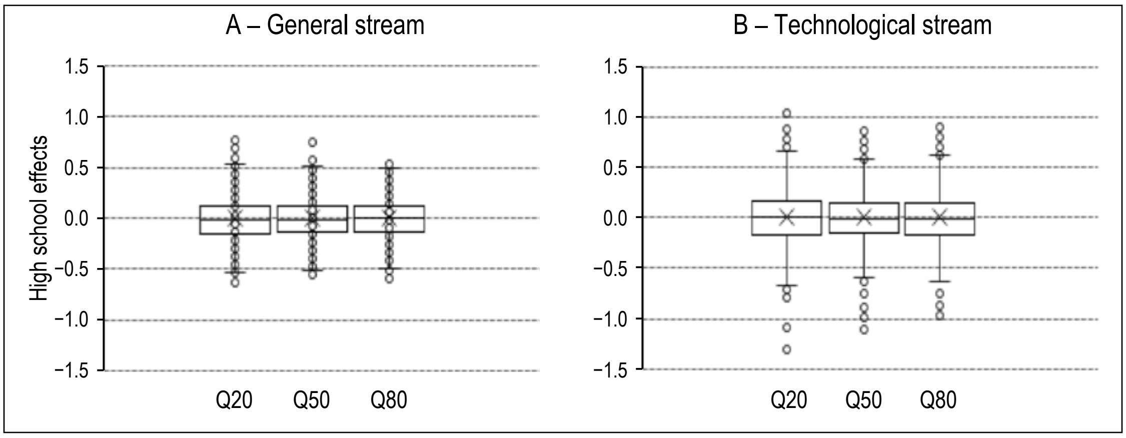
The fixed effects specific to the high school also make it possible to capture the school effects. However, it is necessary to set an identification constraint – within a linear model, it is not possible to estimate the constant and the coefficients separately for all high schools. By convention, the average coefficient for the high schools must be set to zero, which means that for each high school, the estimated fixed effect corresponds to a deviation by this high school from the average effect observed for all high schools. The dispersion of these high school fixed effects is slightly higher in the technological stream than in the general stream (Figure II). This can be at least partly explained by the fact that, for technological high schools, the fixed effects are sometimes estimated on the basis of fewer students, and are therefore less precise. In both streams, it is also possible to observe that the dispersion is slightly greater for the effects of high schools at the bottom end of the distribution (at the level of the first quintile) than at the top (at the level of the last quintile), with extreme values that are far removed from the mean.

Figure III illustrates a case involving two high schools. It represents, for each school, the relationship estimated by quantile regressions between average baccalaureate grades and DNB grades (each observation relates to one student)

[8] *There are 3 series in the general stream (and baccalaureate): S for 'scientific', L for 'literature' and ES for 'economic-social'.*

[9] *This finding suggests that it could be useful to look at the interaction of each individual variable for each series to take account of the differences in testing in each series, and to introduce differentiated expectations for each series according to the characteristics of the students. This option has not been used here, since it greatly increases the number of coefficients that need to be estimated, even though the number of students per series in each high school can be small, bringing with it a risk that the models will be "over-adjusted", which also has consequences for the estimation of the high school fixed effects. It would be prudent to estimate this type of model on the basis of several consecutive years (which was not possible with the data available for this study).*

Figure II – **Characteristics of high school fixed effect distributions**

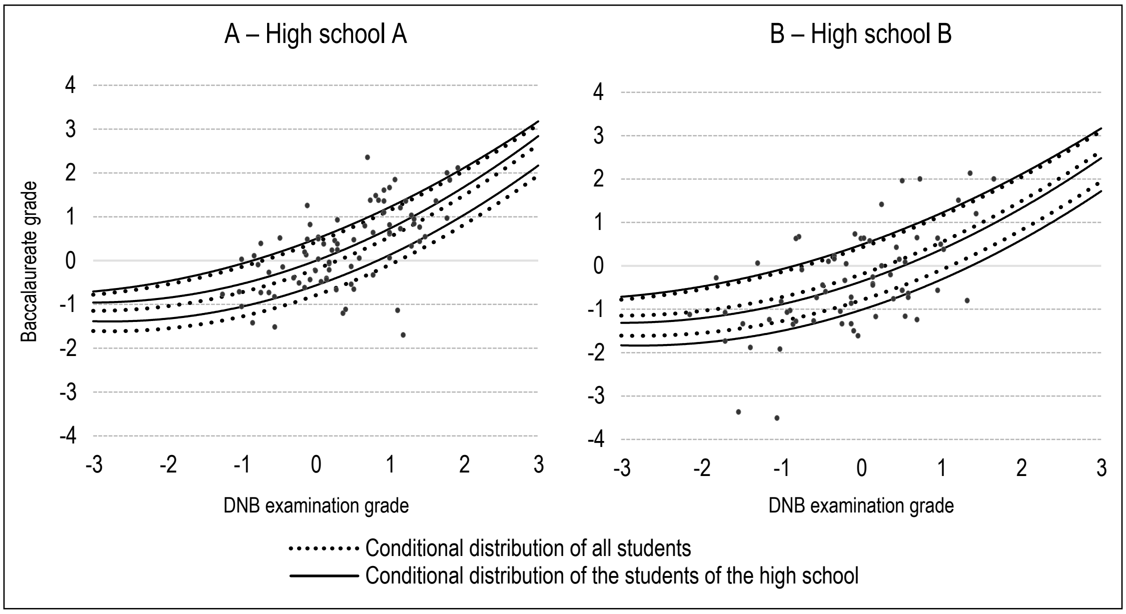


Notes: High school fixed effects obtained by quantile regressions for the first quintile (Q20), the median (Q50) and the last quintile (Q80).
Sources: MENJ-DEPP, FAERE files.

for the three quantiles studied. The solid lines represent the estimates, taking account of the fixed effect of the school – they correspond to a split of students at the high school according to the distribution level of interest. The lowest line represents the first quintile; this is therefore the line that 20% of students at the high school find themselves below and 80% above. Similarly, the other two solid lines represent a split that 50% (for the median) and 80% (for the highest quintile) of students fall below. The dotted lines are the same as the solid lines; however, they do not take account of the fixed effects of the high schools – in other words they represent the expected effects according to the correlations observed across all students who sat the baccalaureate in this stream.

In the two cases illustrated here, the best students from each of the high schools studied do not perform worse than expected (the line representing the highest quintile is slightly above the corresponding dotted line, but the differences are not significant, as is discussed below). Nevertheless, the results obtained by the students at these two high schools are very different for the remaining distribution levels. In high school A, both the median and the lowest

Figure III – **DNB and baccalaureate grades in two high schools and estimates obtained and predicted from quantile regressions**



Notes: High school fixed effects obtained by quantile regressions for the first quintile (Q20), the median (Q50) and the last quintile (Q80). The dotted lines represent the quantile curves (the first quintile Q20, the median Q50 and the last quintile Q80, respectively) obtained by means of estimated regressions for all students in the general stream: for example, 20% of the points in the sample are below the Q20 dotted line. The solid lines represent the results of these estimates by adding the fixed effects specific to the high school under consideration: for example, 20% of the students from high school A fall below the Q20 solid line.
Sources: MENJ-DEPP, FAERE files.

quintile are significantly higher, which means that at least 80% of students at this high school have performed better than expected; this shows that the high school achieved above-average results without this being to the detriment of certain students. Conversely, high school B succeeded in making its top 20% of students perform slightly better than expected, but the weakest 20% did significantly worse than expected. Unlike the previous example, not only does this high school have poorer results at the median level, it also tends to magnify the performance gaps when compared with expectations.

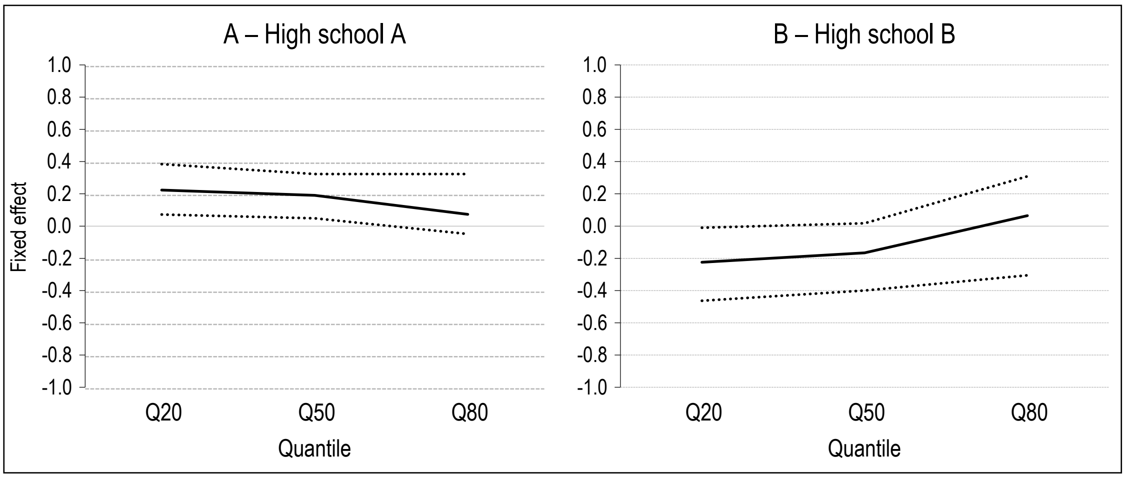
These stylised facts are summarised in Figure IV, which shows the estimated fixed effects for the first quintile, the median and the last quintile in the two high schools. For high school A, all of the coefficients are positive, although the coefficient corresponding to the last quintile is not significant. In high school B, only the coefficient corresponding to the last quintile is positive (but not significant), while the others are negative. The effects show a downward trend for high school A, which also means that the gaps are smaller than expected at this high school, whereas they show an upward trend in high school B, which means that the gaps there are larger than expected.

This analysis can be performed in a more systematic manner: more precisely, it is possible to compare the specific effect for each high school as estimated by the model at the level of the last and first quintiles of the distribution of conditional grades. It is therefore possible to test whether the difference is significantly positive, which would indicate that the high school in question tends to increase performance inequalities among its students, or, conversely, significantly negative, which would indicate that it tends to reduce it. However, these tests must take account of the fact that the repeated use of statistical tests (across all high schools) may lead to an overly frequent acceptance of significantly non-zero differences (see Givord & Suarez Castillo, 2019 for an in-depth discussion). While, for the majority of high schools, the gap does not statistically differ from zero, 8.2% of general high schools and 6% of technological high schools tend to significantly increase performance gaps among their students, while, conversely, 8.5% of general high schools and 7.6% of technological high schools tend to reduce them.

However, reducing the dispersion of the results achieved by students is not an objective in itself. It is not desirable if it will trigger a race to the bottom, i.e. where less unequal results are achieved among students to the detriment of requirements. The case of high school A illustrates that it is possible to observe both improved performance and reduced inequality. To assess whether this phenomenon can be observed more generally, it is possible to compare the effect of each high school at the median level (which corresponds to an approximation of its average value-added) with the gap between the own effects measured at the first and last quintile, respectively, which corresponds to a measure of the effect of the high school on the dispersion of grades.

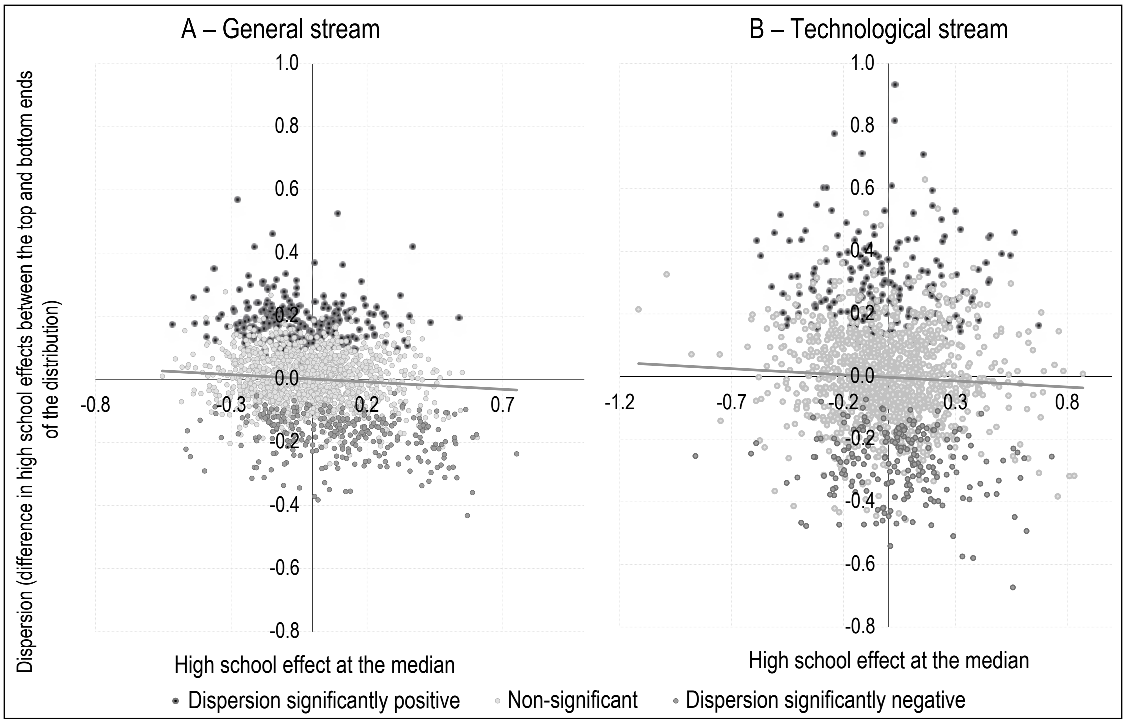
This relationship is illustrated separately for the general and technological streams in Figure V. Each point in this figure represents a high school. The x-axis represents the estimated effect of the high school on the median: a positive value means that the high school tends to improve

Figure IV – **Estimates of the fixed effects in two high schools at three levels of the grade distribution**



Coefficients of high school effects by quantile regressions. The dotted lines correspond to the 95% confidence intervals.
Sources: MENJ-DEPP, FAERE files.

Figure V – **Effects of high schools on the dispersion and the median (general and technological streams)**



Notes: Fixed effects obtained by quantile regressions at the first and last quintiles, as well as at the median.
Reading note: Each dot represents a high school with the estimated fixed effect at the median shown on the x-axis and the difference between the fixed effects of this high school at the first and last quintiles, on the y-axis.
Sources: MENJ-DEPP, FAERE files.

the performance of at least half of its students, while, conversely, a negative value means that it tends to worsen the performance of the majority of students. The y-axis represents the difference between the estimated coefficients for the last and first quintiles. A positive value is associated with more dispersed results than expected at this high school, which means that the high school tends to increase performance inequalities at a given initial composition, and a negative value signifies that the high school tends to reduce performance inequalities among these students.

The first lesson to be learned from this illustration is that high school A is not an isolated case. Across the high schools observed, there was a negative correlation between an increase in performance and an increase in performance inequalities. Numerous high schools are therefore able to help their students to succeed without sacrificing the weakest. However, the slope is steeper in the general stream than in the technological stream. Moreover, this relationship between efficiency and equality is far from deterministic. While "egalitarian" high schools, i.e. those that succeed in reducing the gaps in performance between their students, are more often equally successful, in the sense that they are able to increase average performance, the majority of students in other high schools perform worse than expected. Likewise, while "inegalitarian" high schools (i.e. where the dispersion of results among the students is greater than expected) are more likely to perform worse than average, some of them are also ranked among those that succeed in improving the performance of their students.

* *
*

The evaluation of schools has become a central issue in the public debate. As discussed in the literature review, this issue is further complicated by the fact that the quality of a school is inevitably multidimensional and cannot be judged on a single indicator: this is why the IVALs produced by DEPP look at a number of dimensions (not just success in the baccalaureate, but also retention rates, which are understood as the ability of high schools to support their students throughout their schooling). This article further enriches this description by illustrating the extent to which the indicators that focus purely on the average are as good at reflecting the ability of a school to help all of its students to progress as they are at reflecting a situation in which the focus is on just some of the students.

The results suggest that, while for the majority of high schools it is not possible to statistically highlight heterogeneous effects (the gaps observed are of the same statistical order as those expected), around one sixth of them tend to either amplify or reduce the gaps between the results obtained by their students. Contrary to the opinion sometimes expressed, "inclusive" high schools, which succeed in narrowing the performance gaps among all of their students, do not achieve this by levelling down all of the results. Indeed, these high schools appear to be over-represented in the group of schools that succeed in obtaining better results than expected at the median level. Several remarks must be made with regard to the interpretation of these results.

The first is that, by their very nature, the high school effects are estimated on the basis of limited numbers of observations and are therefore imprecise. It is then difficult to separate exceptional circumstances (such as a few very bright students or an accident that occurred within the school and disrupted schooling, etc.) from those that are fundamental to the school (school projects, school climate, cohesion of the teaching team, etc.). There is a risk that deviations from the mean that are simply statistical accidents could be over-interpreted. To verify the robustness of these findings and to make the estimates obtained less volatile, it would be interesting to compare the estimates obtained for the same high school from one year to the next, or to estimate these effects on the basis of multiple years where these are available (for this study, we were only able to use the data for a single year), as suggested by Bitler *et al.* (2019).

Another difficulty in assessing these effects stems from the fact that they are based on the assumption that all students sitting the baccalaureate have completed the entirety of their high school education at the same school. However, this assumption is not always borne out: some students may move during their school years, or may switch to a different high school to follow a course not offered at the school in which they completed their first year of high school (10th grade). Such changes of school are not just down to the students – some high schools may choose not to accept students whose chances of passing the exam are too low, for example by refusing to enrol them on a course offered or by refusing to allow them to repeat a year. Such strategic behaviour by schools can skew the performance indicators linked to the baccalaureate results. Excluding the students with the poorest results can lead to an overestimation of the value-added of high schools, and can also reduce the dispersion of the results – and therefore make them appear more egalitarian than they are (see Givord & Suarez Castillo, 2019 for a more in-depth discussion). As previously discussed, these effects may become even more important as the evaluation of the schools becomes an issue for stakeholders.[10]

Addressing this issue fully would require student levels to be measured more frequently, in particular to assess the progress that students are making from one year to the next. It is also important to look at other indicators concerning the study paths followed by students: this is made possible by the indicators produced by the DEPP, together with those relating to the baccalaureate pass rate, which provide information regarding the rates of students accessing the baccalaureate from 10th to 12th grades (or years 11, 12 and 13) and therefore potentially regarding these selection mechanisms during the course of their schooling. This question serves as a reminder, as discussed above, that a high school performance cannot be assessed based on a single dimension and that it is essential that multiple dimensions be combined. Beyond the performance in the baccalaureate examination, one option would be to look into the climate at the school and the well-being of its students, or their subsequent integration into higher education and the labour market.

A final key question relates to the ultimate use of indicators to measure school effects. While these measures can serve as guiding tools for various stakeholders within the limits set out in the introduction, their use by families, particularly when it comes to choosing a school, must still be questioned. In fact, studies carried out in New York City show that, in situations where people are choosing schools, and even when information on the value-added of the schools

[10] This point can be linked to the fact that the various experiments involving performance bonuses for teachers do not always provide conclusive results in terms of student progress. A review of the economic literature on this subject can be found in Imberman (2015). While some experiments have demonstrated the effectiveness of performance bonuses in certain developing countries, particularly in India (Muralidharan & Sundaraman, 2011) and in Tanzania (Mbiti et al.*, 2019), with more ambiguous results in Kenya (Glewwe* et al.*, 2010), the various experiments conducted in the United States in particular return findings that do not allow a consensus to be reached with regard to their effectiveness (Dee & Wyckoff, 2015; Fryer, 2013; Springer* et al.*, 2016). The various reasons put forward to explain the minimal or even negative consequences on student progress include the assertion that the bonuses are too small to have any real impact, the fact that financial incentives have no direct effect on teacher motivation or that they compel teachers to focus solely on the subjects and formats of the standardised tests on which the assessment is based. These findings suggest that the effects of performance-based incentive policies are highly sensitive to the specific nature of their implementation (Goodman & Turner, 2013), and in particular that teachers should be evaluated on a number of different criteria rather than relying purely on quantitative measures.*

is available, families do not seem to take this into account when making their choice, prioritising instead the schools that educate the best students (Abdulkadiroğlu *et al.*, 2020). It would be interesting to investigate this point in the case of France, where a significant communication effort around the measurement of high school effects has existed for a long time. □

BIBLIOGRAPHY

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