Dynamic Complementarity in Elementary Schools: Experimental Estimates from Ecuador*

Pedro Carneiro† Yyannú Cruz-Aguayo‡ Rafael Hernández-Pachón§
Norbert Schady¶

November 6, 2025

See latest version here

Abstract

This paper examines experimentally the impact of classroom inputs on student achievement from kindergarten through 6th grade. We use data from a large cohort of elementary school students in Ecuador, who were randomly assigned to different classrooms at the start of each academic year. We estimate reduced form and structural models of the process of skill accumulation to show that learning at the end of a grade is close to an additive function of classroom quality experienced in all previous grades. There is no evidence of dynamic complementarity between classroom quality across different grades.

*Carneiro gratefully acknowledges the support of the ESRC through the ESRC Centre for the Microeconomic Analysis of Public Policy at IFS (ES/T014334/1), and the ERC through grant ERC-2015-CoG-682349. We thank Flavio Cunha, Jonah Rockoff, Jesse Rothstein, and participants in multiple seminars and conferences for their comments, Alejandra Campos, Nicola Dehnen, Nicolás Fuertes, Matías Martínez and Margarita Isaacs for outstanding research assistance, and the Government of Ecuador for collaboration at every step in this research project.

†University College London, Centre for Microdata Methods and Practice, Institute for Fiscal Studies
‡Inter-American Development Bank
§University of Zurich
¶World Bank

1 Introduction

The process of human capital accumulation over the life cycle is complex, involving many inputs, provided by different actors at multiple points in time. Some inputs may have larger effects at some ages than at others, and there may be important interactions between them.

Establishing causal effects of these inputs on outcomes is difficult. The choices made by parents, teachers, and other actors are likely to be correlated with characteristics of the child and the family which are unobservable to the analyst, and they may also respond to the investments made by each other (e.g., Pop-Eleches and Urquiola, 2013).

In this paper, we analyze the dynamic impacts of one important school input—classroom quality—on learning outcomes in elementary school. To do this, we use data from a multi-year experiment conducted in 204 schools in Ecuador, a middle-income country in South America. In these schools, a cohort of approximately 13,500 children was randomly assigned to kindergarten classrooms. Subsequently, these children were randomly re-assigned to different classrooms in 1st, 2nd, 3rd, 4th, 5th and 6th grades.[1] In every grade, compliance with random assignment was very high—98.9 percent on average.

Because of random assignment, each student experiences a unique random sequence of classroom inputs throughout elementary school. This setup is therefore well-suited to estimating the technology of skill formation, and investigating questions such as whether the impact of classroom inputs in one grade increases (or decreases) with the level of classroom inputs in other grades, i.e., dynamic complementarity (e.g., Cunha et al., 2010; Heckman and Mosso, 2014).

To illustrate our main results, we start with a simple representation of the data, using a classification that is crude but easy to visualize. Specifically, we divide classrooms into two categories within each school and grade: high and low quality, defined as above and below the average classroom quality in each school. We compare math and language achievement at the end of each grade for students experiencing different sequences of high- and low-quality classrooms up to that grade. We document that achievement is approximately linear in the number of high-quality classrooms experienced up to a grade, and that the particular sequence in which they occur is irrelevant. This suggests that dynamic interactions are unlikely to be important in our sample.

[1] We refer to our assignment as "random" as shorthand, although technically random assignment only occurred from 3rd grade onward. In the earlier grades, the assignment rules were as-good-as-random. Specifically, in kindergarten, all children in each school were ordered by their last name and first name, and were then assigned to teachers in alternating order; in 1st grade, they were ordered by their date of birth, from oldest to youngest, and were then assigned to teachers in alternating order; in 2nd grade, they were divided by gender, ordered by their first name and last name, and then assigned in alternating order; and in 3rd to 6th grades, they were divided by gender and then randomly assigned to one or another classroom. We provide a number of randomization checks in Appendix C.

We then formally test for interactions between classroom quality in different grades, in a more realistic setup where classroom quality can be continuous, implementing a procedure suggested by Kinsler (2016). We model a student's achievement at the end of each grade as a flexible function of current and past classroom indicators, from kindergarten up to that grade.

We compare the fit of two models. The first model specifies achievement in grade t as an additive function of indicators for classroom assignment in the current and previous grades. These $t+1$ indicators (one for each grade, from 0 to t), or classroom fixed effects, capture the average effect on learning at the end of grade t of assignment to each classroom in all grades up to t. This model assumes no interactions between classroom quality across grades.[2]

The second specification saturates the model described above by adding interactions between classroom indicators in each grade. This is equivalent to using one indicator for each sequence of classrooms (as opposed to one fixed effect per classroom and grade). For example, if there are two classrooms per school and grade, then the model without interactions (first specification) includes $2 \cdot (t+1)$ classroom indicators, whereas the model with interactions (second specification) includes 2^{t+1} sequence indicators.

We find that, for all grades, we cannot reject that the additive model fits the data as well as the model with interactions. In other words, we do not reject the hypothesis that the interactions are jointly equal to zero, i.e., that there are no dynamic complementarities between classroom quality in each grade. A nice feature of this procedure is its flexibility. It does not require specifying a production function or strong distributional assumptions on unobservables.

Finally, we estimate grade-specific production functions of learning (e.g., Cunha et al., 2010; Heckman and Zhou, 2026), where inputs are classroom quality in each (current and past) grade. From the procedure just described, at best we can recover the output of teams of teachers at the end of each grade. Specifying a production function imposes additional structure on the data, but it also allows us to recover a value for the quality of each classroom and to simulate the impacts of changing classroom quality in each grade on achievement. Consistent with the evidence discussed so far, we show that, for most grades, classroom inputs are highly substitutable over time.

When estimating this production function, the output measure is an aggregate of math and language test scores (measured in terms of grade equivalents), taken at the end of each grade, from kindergarten through 6th grade. The set of inputs we consider are classroom quality in each grade. Classroom quality is a natural input to look at since it subsumes everything that

[2]Another way to see this is as a variance decomposition, and our finding as saying that these interactions do not explain a lot of the variance in the outcome. This procedure requires the assumption that the effects of classrooms or sequences of classrooms are homogeneous across students, but that is a standard assumption in this literature.

happens in the classroom, whether it is due to the teacher, peers, or other aspects of the classroom environment.

Typically, the main challenge in estimating (education) production functions is that inputs are chosen endogenously, and may be correlated with unobserved shocks or other unobserved inputs that also affect outcomes. This issue is addressed by our research design, which provides exogenous variation in the sequence of classroom qualities experienced by different students. Nevertheless, our study faces two additional challenges. First, as in most of the recent literature on teacher quality (summarized, for example, in Hanushek and Rivkin, 2012 and Jackson et al., 2014), our input measures are unobserved, and therefore need to be estimated. In that literature, and in our earlier work with this cohort of students (Araujo et al., 2016), it is typical to rely on value added (VA) measures of teacher and classroom quality as inputs in the production of learning. We rely on a similar concept in our paper, and simultaneously estimate VA and the parameters of the production function.

Second, parental investments may respond to classroom inputs. If these responses are substantial, our estimates will conflate the effects of classroom quality on learning, with the effects of parental responses to classroom quality. Fortunately we have some (albeit limited) parental investment data which is helpful for understanding the potential importance of this issue in our setting. Using parental investment data measured at the end of kindergarten (the only period for which it is available in our dataset) Araujo et al. (2016) estimate little to no response of parental behaviors to classroom quality.

More generally, the distinction between policy effects and structural parameters has been emphasized in, for example, Heckman (2000a), Todd and Wolpin (2003) or Heckman (2020). Todd and Wolpin (2003) argue that, in the context of the education production function literature, experimental and quasi-experimental studies target primarily policy parameters (impacts of interventions) while non-experimental observational studies target structural estimates of the production function, and this distinction may partly explain the differences in the results across these two types of studies. Policy parameters include both the direct impacts of policy interventions on outcomes as well as the impacts of subsequent behavioral responses, and are of central interest for policy design. In the context of our paper, even if there were important parental responses to classroom inputs, the question of whether there is dynamic complementarity between school inputs in different time periods would still be of great policy interest to education authorities manipulating such inputs.

The main innovation of this paper is our testing of dynamic complementarity, i.e., the extent to which classroom inputs are substitutable across different grades. As others have noted, finding multiple shocks to human capital accumulation for the same individuals (a requirement to write this paper) is difficult, with one paper arguing that it may be akin to "asking for lightning to

strike twice" (Almond and Mazumder, 2013). Faced with these identification challenges, earlier research has taken one of two approaches. Some papers use panel data on inputs and outcomes (skills) at various points in time to estimate the parameters of a structural model of skill formation in which inputs are allowed to interact with one another. This is the approach taken by Cunha and Heckman (2007), Cunha et al. (2010) and Agostinelli and Wiswall (2016a) using panel data from the U.S., and Attanasio et al. (2020) using panel data from India.

Other papers are based on quasi-experimental variation in policies that affected human capital accumulation at two points in the life cycle. This approach is taken by Johnson and Jackson (2019), who study possible interactions between access to Head Start and court-ordered increases in school spending in the U.S., and by Goff et al. (2025), who test for dynamic complementarities between improvements in the home environment that resulted from the repeal of a ban on abortion and access to higher-quality schools in Romania. Closer in spirit to our work, Kinsler (2016) uses non-experimental data on children in 3rd and 4th grades in North Carolina to test for interactions in teacher quality.

The results of these earlier studies have been mixed.[3] We add to this literature by estimating the importance of dynamic complementarities using experimental data with many years of randomly assigned inputs: by design, children in our experiment were randomly assigned to seven exogenous, orthogonal shocks to classroom quality that affected skill formation, as measured by test scores.

In Campos et al. (2025) we study in detail a much more narrowly defined and observable measure of the teachers' input: the quality of teacher-student interactions. This quality measure is assessed by enumerator ratings of classroom videos using a well known instrument called the CLASS. These videos are only available for the first 5 years of elementary school, while in this paper we use the full panel, with all seven years of elementary school for the cohort we follow. The CLASS is likely to capture only a fraction of what classroom quality really is, whereas in this paper we rely on an omnibus measure of quality more akin to classroom value added. In Campos et al. (2025) we show that the CLASS predicts classroom value added. However, the effects are relatively small, accounting for between a quarter and a third of the within school variation in value added in our sample, but they are persistent, and they are similar across grades. This means that most of what is included in the classroom input is not considered by this measure.

Our focus on testing for dynamic complementarity leads to a much more comprehensive study of this than Campos et al. (2025). In addition to a transparent data visualization exercise and a previously developed flexible test for interactions between teachers (Kinsler, 2016), we implement a novel estimation strategy which extends standard (additive) value added models, and models

[3] Attanasio et al. (2020), Cunha and Heckman (2007), and Johnson and Jackson (2019) find evidence of dynamic complementarities; Kinsler (2016) and Goff et al. (2025) do not.

learning at the end of each grade as an output of the team of teachers a student is exposed to up to that grade. Subsequently, we uncouple the quality of the individual teachers belonging to each team, and show how they interact. Nevertheless, consistent with the findings in this paper, Campos et al. (2025) also document that teacher quality as measured by the CLASS is perfectly substitutable across grades.

The rest of the paper proceeds as follows. In Section 2 we describe the setting for our experiment and the data. Section 3 discusses methodology, and Section 4 presents results. We conclude in Section 5.

2 Setting and Data

We study student achievement in math and language in Ecuador, a middle-income country in South America. As is the case in most other Latin American countries, educational achievement of young children in Ecuador is low (Berlinski and Schady, 2015).

The data we use comes from an experiment in 204 schools. Each school has at least two classrooms per grade (most have exactly two). An incoming cohort of children was randomly assigned to kindergarten classrooms within schools in the 2012 school year.[4] These children were randomly re-assigned to 1st grade classrooms in 2013, 2nd grade classrooms in 2014, 3rd grade classrooms in 2015, 4th grade classrooms in 2016, 5th grade classrooms in 2017, and 6th grade classrooms in 2018. Compliance with random assignment rules was very high—98.9 percent on average. As a result, children who were in our sample of schools for the entirety of the elementary school cycle were exposed to seven exogenous, orthogonal shocks to classroom quality.

Random assignment means that we can deal effectively with concerns about any purposeful matching of students with teachers and peers that often arise in non-experimental settings. Throughout the paper we work with a balanced panel of 8,780 children for whom we have baseline data on preschool attendance, maternal education, and wealth; their receptive vocabulary at the beginning of kindergarten, as measured by the *Test de Vocabulario en Imágenes Peabody* (TVIP), the Spanish version of the widely used Peabody Picture Vocabulary Test (PPVT) (Dunn et al., 1986);[5] and math and language test results at the end of all seven grades. We provide further details on the assignment rules and compliance in Appendix C.

[4]These schools are a random sample of all public schools that had at least two kindergarten classrooms in the coastal region of the country. See Araujo et al. (2016) for details.

[5]Performance on this test at early ages has been shown to predict important outcomes in various settings, including Ecuador. Schady (2012) shows that children with low TVIP scores before they enter school are more likely to repeat grades and have lower scores on tests of math and reading in early elementary school in Ecuador; Schady et al. (2015) show that many children in Ecuador start school with substantial delays in receptive vocabulary, and that the difference in vocabulary between children of high and low socioeconomic status is constant throughout elementary school.

At the end of each grade, we applied age-appropriate math and language tests to children, which we aggregate into a single score. We aggregate correct responses using Item Response Theory (IRT) scores. Since there are common items in tests given in adjacent grades, we are able to construct grade equivalent scores, separately for math and language. This procedure is fairly standard and is described in Appendix B (it is also similar to what is proposed in Attanasio et al., 2020). The final score averages the individual math and language scores, with one-half the weight given to each. In contrast with several papers in education, which measure skills using standardized test scores, to estimate the production function of skill it is important that the test scores we use have a cardinal scale (e.g., Cunha et al., 2010; Agostinelli and Wiswall, 2016b; Freyberger, 2021). In our paper test scores are measured in grade equivalents, so a one unit increase in test scores is anchored to how much the median student in this sample learns in one year (alternatively, Cunha et al. (2010) anchor test scores on a cardinal outcome measured in adulthood, such as schooling or earnings).[6]

Table 1, Panel A, summarizes the characteristics of children and their families. Children were about five years old on the first day of kindergarten. Half of them are girls. At the time children enrolled in kindergarten, mothers were in their early thirties, and fathers were in their mid-thirties. Education levels are similar for both parents—just under nine years of school (completed middle school). The average child in our sample has a TVIP score that places her more than one standard deviation below the reference population that was used to norm the test, indicating that children begin formal schooling with significant delays.[7]

Table 1, Panel B, summarizes the characteristics of teachers in our sample, separately by grade. Across grades, on average teachers are in their mid-40s. Almost all teachers are female in kindergarten, and the proportion of male teachers increases by grade. Kindergarten teachers are less experienced than those in other grades, and they are also less likely to be tenured (rather than working on a contract basis). The average class size in the schools we study is about 38.

In Araujo et al. (2016) we discuss in detail the selection of schools in this study. We show that the characteristics of students and teachers in our sample are very similar to those of students and teachers in a nationally-representative sample of schools in Ecuador.

The most important feature of our data relative to other longitudinal studies in schools is that, in our context, students are randomly assigned to classrooms in every grade. In Appendix A of Carneiro et al. (2023), replicated in Appendix C of this paper, we present a test of random assignment developed by Jochmans (2023), analogous to a standard balance test but adapted

[6]Table B.1 shows percentiles of the distribution of grade equivalent scores at the end of each grade.

[7]The TVIP was standardized on a sample of Mexican and Puerto Rican children. The test developers publish norms that set the mean at 100 and the standard deviation at 15 at each age (Dunn et al., 1986).

to our context with multiple classrooms (as opposed to a treatment and a control group). As expected, we do not reject the hypothesis that students were randomly assigned to classrooms.

3 Empirical Strategy

3.1 Visualizing Our Data

Our goal is to examine how math and language achievement depend on the sequences of classrooms that students experience during elementary school. We classify each classroom (c) in school (s) and grade (t) according to its quality (Q_{cst}). To help visualize the essence of our data it is helpful to consider a simple example where classroom quality is discrete and takes only two values, high or good (G), and low or bad (B): $Q_{cst} = \{G_{cst}, B_{cst}\}$. In practice, classrooms G_{cst} are those in which quality is above the grade- and school-specific mean, while classrooms B_{cst} are below this mean. Therefore, under this definition, quality is defined relatively to other classrooms in the same school. In each school there is always at least one G and one B classroom in each grade (since every school in our sample has at least 2 classrooms per grade).

At the end of kindergarten, each student experienced either a G or a B classroom. At the end of 1st grade, a student could have been in one of four classroom sequences: 1) B in kindergarten and B in 1st grade, or BB; 2) B in kindergarten and G in 1st grade, or BG; 3) G in kindergarten and B in 1st grade, or GB; or 4) G in kindergarten and G in 1st grade, or GG. Each subsequent grade produces increasingly complex sequences of classroom assignment. By the end of 6th grade (the last year of elementary school) there are $2^7 = 128$ sequences of classroom quality any student could have experienced.

Figure 1 shows the full set of sequences we can consider up to 4th grade (it is easy to imagine what happens in subsequent grades, but the diagram becomes too crowded to show it in a single page). For example, students in the sequence $GBBGB$ were in a high-quality classroom in kindergarten and in 3rd grade, but they were in a low-quality classroom in all other grades. Because of random assignment to classrooms within schools, the baseline observable and unobservable characteristics of students in each sequence are identical in expectation.

Discretizing quality in this way greatly simplifies the description of our data, and it helps to visualize its basic features. We can group students in different cells, depending on the sequence of classrooms they experienced, denoted by $\tilde{Q}^t = \{Q_{cs0}, \ldots, Q_{cst}\}$. Average achievement ($Y$) at the end of grade t in each cell (where j is the student) is:

$$\mathbb{E}[Y_{cstj}|\tilde{Q}^t] = \mathbb{E}[Y_{cstj}|Q_{cs0}, \ldots, Q_{cst}]$$

It is then easy to represent graphically how learning depends on the sequence of G or B classrooms experienced by each student. This discretization of the data is only used in this section, for data description and visualization. Our main estimates in the remaining of the paper are based on a more standard framework where classroom inputs are continuous.

In order to determine the sequences faced by each student, we first need to classify classrooms according to their relative quality. Throughout the paper we assume that all students in a given classroom experience the same level of classroom input (this standard assumption rules out, for example, the possibility that a teacher provides different inputs for students at the top or bottom of the class, or for boys and girls). However, the impact of that input on the learning of each student depends on the sequence of classroom inputs experienced by her in previous grades.

Unfortunately, classroom quality is unobserved, so it is not obvious how to determine which classrooms are G or B. In the literature on teacher quality, the quality of a classroom or a teacher is typically measured as the average learning of students in that classroom, or value added (VA). The starting point in this literature (e.g., Araujo et al., 2016), is the following regression:

$$Y_{cstj} = X_{cst-1j}\gamma_t + \delta_{cst} + u_{cstj} \tag{1}$$

where Y_{cstj} is the achievement of student j in classroom c and school s at the end of grade t. X_{cst-1j} is a vector of controls which includes child age, child gender, and a fourth order polynomial in Y_{cst-1j} (as in Chetty et al., 2014). δ_{cst} is a classroom fixed effect, and u_{cstj} is the residual in the model.

Let $v_{cstj} = Y_{cstj} - X_{cst-1j}\gamma_t$. This v_{cstj} represents the amount of learning (or growth in achievement, since we control for Y_{cst-1j}) of student j in grade t. We can then compute VA as:

$$VA_{cst} = \frac{1}{N_{cst}} \sum_{k=1}^{N_{cst}} v_{cstk}$$

where N_{cst} is the number of students in classroom c, school s, and grade t. This means that VA is the average residual learning in the classroom during grade t, after accounting for achievement at the end of grade $t-1$ and other controls.

Since the random assignment of students to classrooms occurs within (and not across) schools, we should demean classroom VA by its school (and grade) mean. Let C_{st} be the number of classrooms, and N_{st} the number of students in school s and grade t. School average VA (at grade t) is given by:

$$\overline{VA}_{st} = \sum_{c=1}^{C_{st}} \frac{N_{cst}}{N_{st}} VA_{cst} \tag{2}$$

Finally, $\alpha_{cst} = VA_{cst} - \overline{VA}_{st}$ denotes the demeaned classroom effect.

One important drawback of the standard VA literature is that it assumes that learning is a linear function of classroom quality. The resulting VA estimates are only valid under this assumption. This framework does not allow, for example, classroom inputs in different grades to be complements, nor does it allow for diminishing returns to accumulated classroom quality over time. Estimating VA in a setting where this assumption is not valid requires a different procedure, which we implement below.

However, even if we allow for a non-additive model, notice that if there is random assignment of students to classrooms in each grade, then students in different classrooms (within the same school) will on average have experienced similar sequences of classroom qualities in previous grades. Therefore, if VA differs across classrooms at the end of a particular grade, with one classroom having a higher VA than the other, it is reasonable to infer that students in the classroom with a high VA received a higher level of classroom input than students in a classroom with low VA. This means that even though that the estimates of VA from an additive model are incorrect, the ranking of classroom VA within each school and grade is likely to be correct, and therefore we can use the standard VA model to classify classrooms in G and B categories. In this section we define G and B classrooms as follows:

$$\begin{aligned} \alpha_{cst} > 0 &\Rightarrow G_{cst} = 1 \\ \alpha_{cst} \leq 0 &\Rightarrow B_{cst} = 1 \end{aligned} \tag{3}$$

Below we discuss the magnitude of the differences in VA between G and B classrooms.

Using this G-B classification we group students into cells, depending on the sequence of G-B classroom assignments they experienced up to a given grade. Let $GB_{tj}^{m_t}$ be an indicator variable that takes value 1 if child j in grade t experienced sequence $m_t = \{G_{cs0j}, \ldots, G_{cstj}\}$, where $t = K, \ldots, 6$.

In order to estimate average learning per cell, we run the following regression:

$$Y_{cstj} = \sum_{m_t} \kappa^{m_t} GB_{tj}^{m_t} + X_{cs0j}\zeta_t + \vartheta_{st} + w_{cstj} \tag{4}$$

where ϑ_{st} is a school-by-grade fixed effect; X_{cs0j} includes age, gender, a wealth index, maternal education (both measured at baseline), and a fourth order polynomial in the baseline vocabulary score (the only assessment we conducted at baseline); Y_{cstj} is the test score (math and language aggregate) at the end of grade t; and w_{cstj} is a residual. Notice that for each child j, $GB_{tj}^{m_t}$ takes value 1 only for the sequence the child experienced, and 0 for all other sequences. There is a different regression for each grade t. It is essential to include school fixed effects in equation (4)

because the randomization of students to classrooms occurs only within schools. With this representation of the data, we can visualize how achievement depends on the sequence of classroom assignments and assess to what extent dynamic complementarities are likely to be important.[8]

3.2 Testing for Additive Classroom Effects

The data visualization exercise just described is intuitive and simple to implement, but requires us to discretize classroom quality, which is quite artificial. Going back to the standard VA model for classroom c in school s and grade t, described in equation (1), we see that it considers test scores for student j at the end of grade t (Y_{cstj}) as a linear function of classroom indicators (δ_{cst}), student level controls (X_{cstj}), and a residual (u_{cstj}). A typical control variable is the lagged test score (Y_{cst-1j}). Under a (conditional) random assignment assumption, δ_{cst} corresponds to the causal impact of being assigned to classroom c on test scores at the end of grade t. δ_{cst} includes the impact of teachers and other classroom shocks (to separate the two one needs data on multiple cohorts of students taught by the same teacher).

Since the assignment of students to classrooms is random in each grade, we do not have to assume that the assignment is random conditional on controls as is standard in the literature relying on observational data, nor do we necessarily need to include controls in the model. We nevertheless include controls to absorb variance in the outcome and increase the power of our tests. Furthermore, since the randomization occurs only within schools, we need to include school fixed effects in the model, which means that δ_{cst} can only capture within school variation in classroom and teacher quality.

To test formally for the presence of interactions between classroom or teacher quality in different grades in the production of learning, Kinsler (2016) proposes augmenting the model in equation (1) by including indicators for current and lagged classroom assignment. Kinsler (2016)

[8]One potential issue with this procedure is that data on the same individuals shows up on both sides of the regression, since the *G-B* indicators and sequences are constructed using all the observations in each classroom (this is not a problem in the remaining sections of the paper, in particular when we estimate the production function of learning, because classroom quality and the parameters of the production function are estimated simultaneously). The fact that the *G-B* indicators are all discrete and the sequences are quite complex is likely to attenuate this problem substantially. In fact, we show below and in Figure A.1 that if we estimate equation (4) using baseline scores as the outcome the estimates of κ^{m_t} are small and statistically indistinguishable from zero, suggesting that this problem is not a major threat to our results. Alternatively, a standard way to address this is to use leave-one-out means when constructing VA: $VA_{cstj} = \left(N_{cst}-1\right)^{-1} \sum_{k=1,k\neq j} v_{cstk}$. However, this produces a strong negative correlation between an individual's achievement and the leave-one-out VA measure corresponding to her, a problem sometimes labeled "exclusion bias" (e.g., Jochmans, 2023; Caeyers and Fafchamps, 2024). In an attempt to address this, when calculating the mean VA in the school for individual i, we leave out not only individual i but also individuals similar to i (with the same classroom achievement rank) from the other classrooms in the school. We show in Figure A.3, and discuss below, estimates of equation (4) based on leave-one-out VA measures. These estimates are very similar to those presented in the main body of the paper.

begins by taking the standard model which assumes additivity of classroom effects over time, and from which we get the following equation:

$$Y_{c_0 \ldots c_t stj} = X_{c_0 \ldots c_t stj} \gamma_t + \sum_{k=0}^{t} \delta_{c_k st} + u_{c_0 \ldots c_t stj} \tag{5}$$

This is a simple extension of equation (1), where the indices of the regression variables have been modified to include the entire history of classroom assignments up to grade t: $c_0 \ldots c_t$. In this model, the impact of classroom quality in different grades on learning at the end of grade t is additive across grades, ruling out any complementarities between classroom quality across different grades.

We then extend the model by saturating it with indicators for the whole sequence of classroom assignments. Not all of them can be added to model in (5) because they would be collinear with the main classroom effects. The model becomes:

$$Y_{c_0 \ldots c_t stj} = X_{c_0 \ldots c_t stj} \gamma_t + \sum_{k=0}^{t} \delta_{c_k st} + \varphi_{c_0 \ldots c_t st} + u_{c_0 \ldots c_t stj} \tag{6}$$

$\varphi_{c_0 \ldots c_t st}$ is the impact of the sequence of classroom assignments $(c_0 \ldots c_t)$ over and above the base impacts of classrooms in each grade $(\delta_{c_k st})$.[9]

Since the randomization of students to classrooms happens only within schools, one also needs to include school fixed effects in the model, and then do the appropriate normalizations with the remaining classroom fixed effects (and interactions). This means that we are only able to estimate the importance of dynamic complementarities across inputs within schools. These are likely to vary less than inputs across schools, but there is still substantial variation in classroom (and teacher) quality to explore within schools (e.g., Araujo et al., 2016).

Kinsler (2016) develops a procedure to test whether the interaction terms, $\varphi_{c_0 \ldots c_t st}$, belong in the model (i.e., a test of the hypothesis that they are all equal to zero). In principle one could simply use an F-test. In Kinsler (2016), however, the very large number of constraints to be tested always leads to very low p-values, which he argues are meaningless. Therefore, he proposes another procedure, which we implement here, and which starts by computing school-specific F-tests of whether the interaction terms are equal to zero, and then calculates the proportion of schools in which this F-test indicates a rejection of the null hypothesis that the model is additive (all interactions equal to zero). Finally, one can compare this proportion with what would be

[9] If classroom shocks were good measures of teacher VA, we could say that the most complete specification captures flexibly the impact of each particular team of teachers one is exposed to up to grade t, whereas the simpler specification imposes that these team effects are additive in the contributions of individual teachers.

expected if the null hypothesis was true (e.g., if the level of significance used in the test is 5%, under the null we would expect this hypothesis to be rejected in 5% of the schools).

3.3 Estimating a Production Function

The procedure just described provides a formal test of the importance of dynamic interactions between classroom inputs, but it does not give us quantitative assessment of their magnitude. Therefore, in this section we describe the estimation of a production function of achievement, which we then use to quantify the impacts of different sequences of inputs on student achievement.

We model learning at the end of grade t as a function of the sequence of classroom qualities experienced up to that grade. Theory does not inform us about which functional form to use, but whatever our choice is, it should be flexible enough to accomodate different substitution patterns between classroom quality in different grades. In this paper we approximate this function using a CES, although we could have several other approximations (e.g., translog):

$$Y_{c_0 \ldots c_t stj} = A_{c_0 \ldots c_t stj} \left(\sum_{k=0}^{t} \pi_{kt} \delta_{c_k sk}^{\rho_t} \right)^{\frac{\theta_t}{\rho_t}} u_{c_0 \ldots c_t stj} \tag{7}$$

As before, $Y_{c_0 \ldots c_t stj}$ is learning at the end of grade t and $\delta_{c_k sk}$ is classroom quality experienced by a student assigned to classroom c_k in grade k. The parameters of this CES function are all grade-specific (e.g., Cunha et al., 2010; Heckman and Zhou, 2026). ρ_t (where $-\infty < \rho_t < 1$) determines the degree of substitution between classroom inputs in different grades ($\sigma_t = \frac{1}{1-\rho_t}$ is the elasticity of substitution), θ_t determines the returns to scale, and π_{kt} give us the relative productivity of classroom quality in each grade (we use the following normalization: $\sum_{k=0}^{t} \pi_{kt} = 1$). $A_{c_0 \ldots c_t stj}$ is a productivity parameter that includes school fixed effects (ϑ_{st}) as well as individual level observables ($X_{c_0 \ldots c_t stj}$, consisting of test scores, maternal education, and an index of household wealth, all measured at baseline, i.e., at the beginning of kindergarten). We assume that $\ln A_{c_0 \ldots c_t stj} = \vartheta_{st} + X_{c_0 \ldots c_t stj} \gamma_t$. $u_{c_0 \ldots c_t stj}$ is an individual level i.i.d. shock.

As discussed above, classroom quality, $\delta_{c_k sk}$, is unobserved. Unlike most of the literature, we allow learning to be a non-separable function of classroom quality in different grades. This means that we cannot rely on additive VA models to recover $\delta_{c_k sk}$. Instead, $\delta_{c_k sk}$ needs to be estimated together with the parameters of the production function.

Typically, it would not be possible to estimate a production function with unobserved inputs. It is possible to do it here because we know that every student in the same classroom in a particular grade experiences the same level of that grade specific input, regardless of the sequence of inputs they are exposed to in other (previous and subsequent) grades. It is this group structure

of the data that allows us to recover simultaneously the classroom inputs students are exposed to (which shows up as a fixed effect in the residual learning regression) and the parameters of the production function.

Equation (7) defines a system of equations, one for each grade $t = 0, \ldots, 6$. In order to estimate it, we start by taking logs:

$$\ln Y_{c_0 \ldots c_t s t j} = \vartheta_{st} + X_{c_0 \ldots c_t s t j} \gamma_t + \frac{\theta_t}{\rho_t} \ln \left(\sum_{k=0}^{t} \pi_{kt} \delta_{c_k s k}^{\rho_t} \right) + \upsilon_{c_0 \ldots c_t s t j} \tag{8}$$

We define $\upsilon_{c_0 \ldots c_t s t j} = \ln u_{c_0 \ldots c_t s t j}$. In addition, we need to initialize the system. Notice that the implied equation for grade 0 (kindergarten) only has one classroom input, and therefore it simplifies to:

$$\ln Y_{c_0 s 0 j} = \vartheta_{s0} + X_{c_0 s 0 j} \gamma_0 + \theta_0 \ln \left(\pi_{c_0 s 0}^{\frac{1}{\rho_0}} \right) + \theta_0 \ln(\delta_{c_0 s 0}) + \upsilon_{c_0 s 0 j} \tag{9}$$

This is a standard VA equation for kindergarten, where $\ln Y_{c_0 s 0 j}$ is a linear function of classroom assignment indicators, which are estimated to be $\theta_0 \ln(\delta_{c_0 s 0})$. θ_0 is normalized to be equal to 1. The θ_t parameters in the remaining grades can then be freely estimated.

As discussed above, the assumption that classroom inputs are common to all students in a particular classroom means that the parameters of the system of equations outlined in (8) (one equation per grade) and the vector of classroom qualities are identified, and should be estimated simultaneously. In practice, it is computationally easier to proceed iteratively, one grade at a time, starting with the lower grades.

The estimation procedure, which approaches each grade in sequence, is described in detail in Appendix D. We start from equation (9), $t = 0$, from which we recover estimates of $\delta_{c_0 s 0}$ for each classroom (and estimate the remaining parameters of the model, which are not of substantive interest). From the equation for 1st grade (equation (8) for $t = 1$), we use $\delta_{c_0 s 0}$ from the $t = 0$ equation, and estimate all the parameters of the production function $(\theta_1, \rho_1, \pi_{c_0 s 1})$ together with $\delta_{c_1 s 1}$ (as well as the parameters on the controls). In grade t, we use $\{\delta_{c_0 s 0}, \ldots, \delta_{c_{t-1} s t-1}\}$ obtained from the previous grade equations, and estimate $(\theta_t, \rho_t, \pi_{c_0 s 0}, \ldots, \pi_{c_{t-1} s t})$ together with $\delta_{c_t s t}$.

Before we proceed to the empirical results, there are two additional issues to discuss concerning the estimation of this production function. The first one concerns its specification. This is one important aspect in which the production function in equation (7) is different from the specifications in other recent papers such as, for example, Cunha et al. (2010), Agostinelli and Wiswall (2016a), or Attanasio et al. (2020). In those papers the production function has a first

order Markov structure (as in the standard VA model of equation (1)), where skills in period t, say Y_{tj}, depend on skills in period $t-1$, Y_{t-1j}, and inputs in period t, δ_{tj}: $Y_{tj} = f_t(Y_{t-1j}, \delta_{tj})$, where $f_t(\cdot)$ is the period t production function. In other words, all interactions between inputs in period t (δ_{tj}) and inputs in prior periods ($\delta_0 j \dots \delta_{t-1j}$) operate through lagged skills (Y_{t-1j}).

The specification in equation (7) relaxes this assumption,[10] which is found to be too restrictive in Attanasio et al. (2020), but restricts the substitutability of inputs in different grades to be determined by a single parameter, so, for example, the elasticity of substitution between inputs in grades t and $t-1$ is the same as the elasticity of substitution between inputs in grades t and $t-k$, where $k \neq 1$.

Second, since the estimation procedure examines each grade sequentially, using as data previously estimated classroom inputs, one could be worried about the impact of estimation error (in past classroom inputs) on our estimates of the production function parameters (although this would not be an issue if we estimated the production functions for all grades simultaneously). Using a procedure developed by Evdokimov and Zeleneev (2025) we show in Appendix E that our estimates are robust to this problem.

4 Results

4.1 Sequences of High- and Low-Quality Classrooms

We begin by estimating classroom VA for each grade, based on equations 1 and 2. Table 2 shows estimates of the standard deviation of classroom VA for each grade (demeaned within each school), with and without correction (shrinking) for sampling error (using the standard shrinking procedure in the literature; see, e.g., Araujo et al., 2016). Taking the corrected estimates, one standard deviation increase in classroom quality corresponds to a 0.034 grade equivalent increase in kindergarten test scores, and a 0.178 increase in sixth grade test scores. Given the standard deviation of our grade equivalent scores (see Table B.1), these estimates correspond to about 10% of a standard deviation of test scores in each of these grades, similar to classroom VA estimates in the US (see, e.g., Chetty et al., 2014).

Using these VA estimates we construct the G and B indicators, using equation 3. Table 3 shows the mean and various quantiles of the distribution of the differences in VA between G and B classrooms. On average, G classrooms have a VA which is 0.081 grade equivalents higher than B classrooms in kindergarten, whereas this difference goes up to 0.361 by sixth grade.

[10] In addition, another advantage of proceeding this way is that all inputs are randomly assigned, where as lagged achievement is not.

We can now present the data using the simple visualization procedure described in Section 3. We discretize classroom quality in each grade to take only two values, G and B, and estimate the average learning for children in each sequence of (discretized) classroom qualities across grades, as in equation (4). The impact of each sequence is reported relative to the worst possible sequence (being in the B classroom in every grade). Therefore, there are 3 parameters to estimate at the end of 1st grade, 7 at the end of 2nd grade, 15 at the end of 3rd grade, 31 at the end of 4th grade, 63 at the end of 5th grade, and 127 at the end of 6th grade. These estimates are displayed graphically in Figure 2.

The bars in the six panels of Figure 2 have different colors, and are ordered from left to right, according to the number of G classrooms in the sequence. Take, for example, Panel A, which shows impacts at the end of 1st grade. The left-most bar in this panel shows that students who have a B classroom in kindergarten and a G classroom in 1st grade (a BG sequence) have test scores that are 0.134 grade equivalents (GE) higher than those who have a B classroom both in kindergarten and 1st grade (a BB sequence). Those in a GG sequence have test scores that are 0.208 GE higher than those in a BB sequence.

Taken together, the panels in Figure 2 show that a larger number of G classrooms in a sequence generally leads to higher achievement, as one would expect (the average height of the bars of a color corresponding to a particular number of G classrooms increases with the number of such classrooms; see also Figure 3). At the end of 6th grade, students who were randomly assigned to the best (G) classroom in the school for seven grades in a row (sequence $GGGGGGG$) have scores that are 1.34 GE higher than those who were randomly assigned to the worst (B) classroom in the school in every grade (sequence $BBBBBBB$). In the first row of Table 4 we test and reject the hypothesis that students in sequences with more G classrooms do not have different achievement than students in sequences with fewer G classrooms (i.e., we estimate that achievement significantly rises with the number of good classrooms in the sequence).

Keeping constant the number of G classrooms in the sequence, the bars are not all of equal height. This suggests that the specific grades in which G classrooms appear within each sequence (and not just the number of G classrooms) may be important (either because quality in the earlier grades has higher or lower productivity than quality in later grades, or because of dynamic complementarity). In particular, the bars are frequently taller in sequences in which G classrooms are more recent. For example, in Figure 2, Panel B, taking sequences with only one G classroom, the bar corresponding to BBG is taller than those corresponding to BGB or GBB. If we look at sequences with two G classrooms, the bar for GGB is shorter than those corresponding to BGG or GBG. This is consistent with the idea that there is depreciation (or "fade-out") of the effects of classroom inputs (i.e., more recent inputs have larger impacts), as documented in several papers on teacher effects estimated with U.S. data (for example, Chetty et al., 2014, Jacob et al., 2010).

Figure 2 also suggests that the timing of classroom quality could matter beyond depreciation. Even keeping the number of good classrooms in a sequence fixed, it is not always true that sequences with more recent G classrooms are the ones where student achievement is the highest. For example, if we take the impact of the sequences with three G classrooms on 3rd grade achievement (the bars corresponding to $BGGG$, $GBGG$, $GGBG$, and $GGGB$ in Panel C), the impact of $BGGG$ is lower than the impact of $GBGG$.[11]

The impacts of various sequences on learning can be aggregated in different ways. Figure 3 shows a particularly simple and instructive one for our purposes: in each panel, we average the height of bars of the same color from the corresponding panels of Figure 2. Take, for example, Panel C (3rd grade). The zero good classrooms case serves as the benchmark in every panel, with a height of zero by definition. The first bar in Figure 3, Panel C shows the average height of the four bars representing sequences with exactly one good classroom from Panel C of Figure 2. The second bar averages all bars corresponding to sequences with exactly two good classrooms. The third bar represents the average for three good classroom sequences. Finally, the last bar is identical in both figures because there is only one possible sequence with four good classrooms at the end of third grade.

In each panel of Figure 3 we overlay the bars with a line corresponding to a linear regression of the height of each bar on the number of good classrooms they represent (including 0 good classrooms, which has an implicit bar of height equal to zero).

It is striking that for Panels A, B, and C (corresponding to 1st to 3rd grades) the linear regression fit is close to perfect, indicating that achievement is a linear function of the number of good classrooms in the sequence. In the remaining panels, there are deviations from linearity but they are small. Furthermore, we should also note that the sample size for each bar becomes smaller for later grades, because the overall sample size has to be split across a larger number of bars. This means that sampling error for each bar is larger for later than for earlier grades. We cannot reject for any grade that achievement is a linear function of the number of good classrooms in the sequence. The p-values for this test are shown in the second row of Table 4.

[11] As an additional check of the validity of our procedure we re-estimate equation (4) using baseline test scores (TVIP) as the outcome. Since students were randomly assigned to sequences, we should not observe any impact of being assigned to a sequence on baseline test scores, which are measured right at the start of elementary school. In Figure A.1 we replicate Figure 2 for the case where TVIP scores are used as the outcome, the scale of the graphs being the same as in Figure 2. Impacts on TVIP scores of being assigned to different sequences are very small. We do not reject that they are equal to zero in grade 1 through 5 for which the p-values of the test that the coefficients on the sequences are jointly equal to zero are 0.31, 0.58, 0.73, 0.87 and 0.43 respectively. Surprisingly we reject this hypothesis in grade 6 (p-value = 0.01), although we can see that the bars in Figure A.1 are both negative and positive and generally small in magnitude, so this may be due to the fact that we are testing a large number (127) of hypotheses simultaneously. In Figure A.2 we average bars of the same color from Figure A.1, to show that the number of good classrooms one is exposed to is not correlated with baseline scores.

This (perhaps surprising) linearity suggests that there are no strong dynamic complementarities in our data. If these were important, we would expect achievement to be a convex function of the number of good classrooms in each sequence, because the marginal impact of an additional high-quality classroom would increase with the number of high-quality classrooms experienced elsewhere in the sequence.[12]

4.2 Testing for Complementarity in the Impacts of Classroom Assignment on Learning

We now turn to the test of complementarity proposed in Kinsler (2016), and which consists of comparing the fit of the models in equations (5) and (6). In particular, we test whether the classroom interaction terms in equation (6) are jointly equal to zero.

We conduct this test at the end of each grade, from 1st to 6th grade (since it only makes sense to do it when there are at least two grades). The number of classroom effects and interactions in the model increases with the grade we consider, since we interact classroom effects for two grades (corresponding to *teams* of two teachers) at the end of 1st grade, but we interact classroom effects for seven grades (corresponding to *teams* of seven teachers) at the end of 6th grade. We start by using only controls (test scores, maternal education, and household wealth) measured at baseline, which means that the set of indicators for each sequence captures the total impact of the sequence of classroom quality up to grade t on achievement at the end of that grade. The results are shown in Panel A of Table 5. Panel B, which is similar to A, shows the case where controls (in particular, lagged test scores) are measured in $t-1$, which means that the indicators for classroom sequences capture the impact of the sequence of classroom quality up to grade t on learning (or VA) occurring only in that grade.

In the first column of Table 5 we report the proportion schools for which we reject the null hypothesis of no interactions between classroom effects on student achievement, using a significance level of 10%. Each row corresponds to a different grade. Across all grades, the proportion of schools for which we reject the null is approximately 10%. This is exactly what we would expect if the null hypothesis is true.

In the remaining two columns we document that the proportion of schools for which we reject the null of no interactions at the 5% level is approximately 5%. If the significance level used is 1%, then the proportion of schools for which the hypothesis is rejected is approximately 1% across grades. Again, this is what we would expect under the null hypothesis of no interactions

[12] Figure A.4, constructed from Figure A.3, shows that results are essentially the same when using leave-one-out measures of classroom and school VA to construct the *G-B* sequences.

between classroom effects.[13] In Panel B we show that the results are very similar if we control for test scores in grade $t-1$ ($Y_{c_0 \ldots c_t s t-1 j}$), as opposed to baseline test scores ($Y_{c_0 \ldots c_t s 0 j}$).

In sum, we cannot reject that the model is additive in classroom quality or, in other words, that there are no strong dynamic complementarities between classroom inputs in different grades.

4.3 Estimates of the Production Function

Finally, we show the estimates of a (CES) production function, which allow us to simulate the impacts on learning of being exposed to different counterfactual sequences of classroom input (or quality). As discussed above, we allow for grade-specific production functions. The estimated parameters (θ_t, ρ_t, $\pi_{c_k s t}$) as well as the estimated classroom inputs ($\delta_{c_k s k}$) are reported in Tables A.1 and A.2.[14]

In Table 6 we show grade specific estimates of the average marginal product of increasing classroom quality in one unit, and the corresponding standard error. As expected, all estimates are positive and statistically different from zero. The impact of increasing classroom quality by 0.1 units (which, depending on the grade, corresponds to moving from the 25th percentile of classroom quality to somewhere between the 50th and 75th percentile, see Table A.2) is on average 0.09 grade equivalents in grade 1, declining to 0.03 grade equivalents in grade 6.

To understand what these estimates imply for the production of achievement we simulate average predicted scores for various combinations of classroom inputs. The different panels in Figure 4 show achievement at the end of each grade as a function of classroom quality in that grade, keeping fixed classroom quality in previous grades.[15] Each panel has five lines. Three of these are thick lines, and they differ between themselves because they fix previous classroom quality at different values: the thick solid line (labeled "P50") fixes previous classroom quality at the median value in each grade, the thick dotted line ("P25") fixes these values at the 25th percentile, and the thick dashed line ("P75") fixes these values at the 75th percentile. For example, in Panel A, the P25 line shows how achievement at the end of 1st grade depends on 1st grade

[13] We should also note that even if we had just performed a standard joint F-test to the overall sample, as opposed to doing it school by school, the p-values at the end of grades 1 through 6 would be 0.2899, 0.9433, 0.7653, 0.7092, 0.5810 and 0.2183, respectively, which also means that we do not reject the null that there are no dynamic interactions between classroom inputs. This is in spite of the fact that (as in Kinsler, 2016) we are testing a large number of restrictions.

[14] In Table A.1 we also show the implied estimates of the elasticity of substitution ($\frac{1}{1-\rho_t}$) between classroom inputs in different grades. These are restricted to be constant by the functional form we use, which is a strong assumption. As mentioned above, we take this functional form as a useful approximation to an unknown function. We find it more helpful to understand the implied properties of our estimated production function by simulating it (as we show below), rather than focusing on these particular parameter estimates.

[15] To minimize the influence of extreme values over which it may be difficult to estimate the production function, we limit the support of classroom quality in the figures, so we consider only values of the input between the 10th and 90th percentiles of its distribution in each grade.

classroom quality when kindergarten classroom quality is fixed at the 25th percentile of the distribution. As another example, the P75 line in Panel E shows how achievement at the end of 5th grade depends on fifth grade classroom quality when classroom quality in each of the previous grades is fixed at the 75th percentile of the within-grade distribution of classroom quality.

All three lines are upward sloping in all panels, indicating that the marginal product of classroom quality is always positive (regardless of the value of previous inputs). For the same reason, in every panel the P75 line is everywhere above the P50 line, which in turn is above the P25 line (because the marginal product of lagged inputs is also positive). For 1st grade it is difficult to distinguish the three lines, because the more recent input is considerably more important than the lagged input. These lines become more clearly apart as we look at production functions at higher grades, probably because more inputs have been accumulated in previous grades.

The remaining two lines in each panel are a thin dashed line (labeled "P75-Substitutes") and a thin dotted line ("P25-Substitutes"), which are superimposed respectively on the thick dashed and thick dotted lines just described. They are meant to represent how the production function would look like if classroom qualities were perfect substitutes over time. To draw them, we first calculate the difference between the average achievement across the thick dashed (dotted) and the thick solid lines, and then we add this difference to the thick solid line. In other words, we ask what the production function would look like if exposure to different classroom qualities in previous grades only shifted the production function (as a function of the current quality keeping previous qualities fixed) in a parallel way, without affecting its slope.

It is remarkable how small the differences are between the thick and thin lines across all panels. In other words, inputs in different grades are close to perfect substitutes.[16]

Figure 5 replicates Figure 4, but instead of evaluating the relationship between achievement and classroom inputs in grade t at the 25th, 50th, and 75th percentiles of classroom inputs in all previous grades, we do this at percentiles 10, 50 and 90. The main reason for this exercise is that dynamic complementarities may be visible only if we consider large differences in lagged inputs.

Our results indicate that, to a small extent, this may be true, especially when we consider the last grades of elementary school (perhaps because, for example, the differences between the P10 and P90 lines correspond to 6 years of accumulated skills at the end of 6th grade, but only 2 years

[16] At first sight this is not very consistent with the estimated elasticities of substitution in Table A.1, which are not very high. Nevertheless, because of the scale of the inputs and the value of other parameters it is still possible that our estimates result on the figures just discussed. This is why we believe it is better to present simulations of the production function than to focus on the estimated parameters, which are very specific to the functional form we use. These estimates are also consistent with what we observe in the data visualization exercise and the non-parametric test of dynamic complementarity presented above.

of accumulated skills at the end of 2nd grade). That said, these figures show small departures from the case of perfect substitutes.[17]

To summarize, when assessing to what extent dynamic complementarity is an important feature of our data one should take the entire evidence presented in the paper together. We started by showing that learning is an approximately linear function of the number of good classrooms a child was assigned to. Next, we documented that there is no strong evidence that interactions between classroom quality in different grades are important to explain student achievement. Finally, we showed that, for almost all grades, the elasticity of substitution between classroom inputs in different periods is quite high. Taken together, these results suggest that dynamic complementarity between classroom inputs in different grades is unlikely to be an important feature of our data.

4.4 Parental Responses to Classroom Quality

The main challenge estimating production functions is that several inputs may be unobserved. Notably unobserved in our case are time varying parental inputs (whereas time invariant inputs may be thought to be controlled for by baseline controls, namely baseline test scores). Therefore, even if there is plausible exogenous variation in inputs as in our case, one cannot always rule out that other unobserved inputs also respond to this variation.

If parental investments respond to classroom inputs, our estimates will conflate the effects of classroom quality on learning with the effects of parental responses to classroom quality. This is a notoriously difficult subject to study because it requires exogenous variation in school inputs linked to rich data on parental investments. There is however a small literature that attempts to estimate such responses, and generally finds them to be compensatory, i.e., increases in school quality crowd out parental investments (see, e.g., Datar and Mason, 2008; Houtenville and Smith Conway, 2008; Das et al., 2014; Pop-Eleches and Urquiola, 2023; Fredriksson et al., 2016; Greaves et al., 2023; Carneiro et al., 2025).

These results would be consistent with strong incentives to substitute, because of diminishing returns to human capital, which would only be counteracted if the production function exhibits strong complementarity between parental skills and classroom inputs. This could of course happen, making the sign of potential parental responses ambiguous. Fortunately we have some (albeit limited) parental investment data which is helpful for understanding the potential importance of this issue in our particular setting. Using parental investment data measured at the end of kindergarten (the only period for which it is available in our dataset) Araujo et al.

[17] In Figures A.5 to A.8, we show that these patterns remain true even after we account for measurement error in the estimation of classroom quality. The details of the measurement error correction are explained in Appendix E.

(2016) and Campos et al. (2025) estimate little to no response of parental behaviors to classroom quality. These results suggest that, in our setting, our estimates are not conflating structural features of the production function of skill with parental responses.

Even if we had found, as in most of the literature, that parental investments decrease when school investments increase, it is not clear why this would lead us to a finding of no dynamic complementarity in every grade. The conflation of structural and behavioral responses would be very complex in that case, and a finding of no dynamic complementarity for every grade would require the structural and behavioral responses to cancel each other out in each and every grade, a remarkable coincidence.

That said, as discussed in the introduction, the conflation of structural and behavioral effects give us the total effect of changing policy inputs on outputs. The distinction between policy effects and structural parameters has been emphasized in, for example, Heckman (2000a), Todd and Wolpin (2003) or Heckman (2020). Todd and Wolpin (2003) argue that, in the context of the education production function literature, experimental and quasi-experimental studies target primarily policy parameters (impacts of interventions) while non-experimental observational studies target structural estimates of the production function, and this distinction may partly explain the differences in the results across these two types of studies. Policy parameters include both the direct impacts of policy interventions on outcomes as well as the impacts of subsequent behavioral responses, and are of central interest for policy design. In the context of our paper, even if there were important parental responses to classroom inputs, the question of whether there is dynamic complementarity between school inputs in different time periods would still be of great policy interest to education authorities manipulating such inputs.

5 Conclusion

This paper estimates how classroom quality in different grades in elementary school affects achievement. It focuses on whether there are dynamic interactions between classroom inputs across grades.

We use data from a unique experiment in elementary schools in Ecuador, where, within each school, students were randomly assigned to classrooms in every grade, between kindergarten and 6th grade. This ensures that each student was exposed to a sequence of seven exogenous, orthogonal shocks to skill formation in elementary school.

Using a variety of approaches, we do not uncover evidence of dynamic complementarities in classroom quality across grades. The productivity of classroom quality in one grade does not

depend on the quality experienced by children in earlier grades. Rather, the production function of education is remarkably additive in classroom quality across different grades.

We show this by documenting that: 1) in a model with only two categories of classroom quality, good and bad, learning is a linear function of the number of good classrooms experienced up to a grade; 2) in a flexible model where learning is a function of classroom assignments in each grade, there is no evidence of strong interactions between classroom effects across grades; 3) when estimating a CES production function, the implied elasticity of substitution between classroom inputs across grades is generally large, indicating that classroom inputs in different grades are highly substitutable.

This is perhaps a surprising result, at least at first sight. As Heckman has emphasized in several papers, the idea that *skill begets skill* (for example, Heckman, 2000b) is an intuitive description of the learning process. If that were the case, a good 1st grade classroom would help students learn the 1st grade material well, give them solid building blocks for 2nd grade learning and, therefore, allow them to benefit more from a high-quality learning environment in 2nd grade. Our results suggest, however, that the production process in schools in the setting we study does not occur in this way.

There would be clear benefits to future research to understand the absence of dynamic complementarity, and whether this result holds in other settings. It could be that teachers effectively tailor their instruction to specific children, so that each child benefits equally from instruction in a given grade (regardless of the quality of the classroom she was exposed to in earlier grades). Although this would be consistent with the results we observe, it seems unlikely to us given overall low teacher quality in Ecuador, and the large number of children in each classroom—between 35 and 40, on average.[18]

Alternatively, teachers may not tailor instruction to individual students, but they may focus on material that is particularly relevant for lagging students.[19] That is, if the input is defined as relevant material covered in class, teachers may provide more of that input to students who had low-quality classrooms in earlier grade(s). In this scenario, there could be dynamic complementarities in the learning process of *individual* children, but teachers in essence offset these complementarities. Or, perhaps, it is parents who offset dynamic complementarities, by making larger investments in children who had worse classrooms in the past—although in our earlier work on kindergarten we found no evidence of such offsetting behaviors by parents (see Araujo et al., 2016).

[18]We note, also, that in almost none of the classrooms in our data is there a teacher's aide, so the number of children in a classroom are, effectively, the number of children per teacher.

[19]Duflo et al. (2011) argue that the opposite occurs in elementary schools in Kenya. Rather, in their model teachers focus on the highest-performing children in a classroom because they seek to maximize the number of students who pass an exam that determines entrance into high school.

More generally, we stress that this is a literature where inputs are hard to define, observe, and measure. The parameters of the production function may not be invariant to the choice of inputs that are estimated. Our paper shows, however, that if the input that is analyzed is classroom quality, measured by a broad aggregate such as classroom VA, the production function of skills in elementary school is additive in classroom inputs, at least in the setting that we study.

References

Agostinelli, F. and M. Wiswall (2016a). Estimating the technology of children's skill formation. *National Bureau of Economic Research Working Paper Series* (22442).

Agostinelli, F. and M. Wiswall (2016b). Identification of dynamic latent factor models: The implications of re-normalization in a model of child development. Technical Report 22441, National Bureau of Economic Research.

Almond, D. and B. Mazumder (2013). Fetal origins and parental responses. *Annual Review of Economics 5*(1), 37–56.

Araujo, M. C., P. Carneiro, Y. Cruz-Aguayo, and N. Schady (2016). Teacher quality and learning outcomes in kindergarten. *Quarterly Journal of Economics 131*(3), 1415–1453.

Attanasio, O., R. Bernal, M. Giannola, and M. Nores (2020). Child development in the early years: Parental investments and the changing dynamics of different dimensions. *National Bureau of Economic Research Working Paper Series* (27812).

Attanasio, O., S. Cattan, E. Fitzsimons, C. Meghir, and M. Rubio-Codina (2020). Estimating the production function for human capital: Results from a randomized control trial in colombia. *American Economic Review 110*(1), 48–85.

Attanasio, O., C. Meghir, and E. Nix (2020). Human capital development and parental investment in india. *Review of Economic Studies 87*(6), 2511–2541.

Berlinski, S. and N. Schady (2015). *The Early Years: Child Well-Being and the Role of Public Policy*. New York: Palgrave Macmillan.

Caeyers, B. and M. Fafchamps (2024). Exclusion bias and the estimation of peer effects. *Journal of Human Resources*. Published online November 7, 2024 before print.

Campos, A., P. Carneiro, Y. Cruz-Aguayo, C. Etcheverry, and N. Schady (2025). Interactions: Do teacher behaviors predict achievement, executive function, and non-cognitive outcomes in elementary school? Working Paper.

Carneiro, P., Y. Cruz-Aguayo, F. Salvati, and N. Schady (2023). The effect of classroom rank on learning throughout elementary school: Experimental evidence from ecuador. *Journal of Labor Economics 43*(2).

Carneiro, P., P. Glewwe, A. Guha, and S. Krutikova (2025). Teacher value added and classroom practices in vietnam. Working Paper.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014). Measuring the impact of teachers i: Evaluating bias in value-added estimates. *American Economic Review 104*(9), 2593–2632.

Cunha, F. and J. J. Heckman (2007). The technology of skill formation. *American Economic Review, Papers and Proceedings 97*(2), 31–47.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica 78*(3), 883–931.

Das, J., S. Dercon, J. Habyarimana, P. Krishnan, K. Muralidharan, and V. Sundararaman (2014). School inputs, household substitution, and test scores. *American Economic Journal: Applied Economics 5*(2), 28–57.

Datar, A. and B. Mason (2008). Do reductions in class size "crowd out" parental investment in education? *Economics of Education Review 27*(6), 712–722.

Duflo, E., P. Dupas, and M. Kremer (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *American Economic Review 101*(5), 1739–1774.

Dunn, L. M., D. M. Lugo, E. Padilla, and L. M. Dunn (1986). *Test de Vocabulario en Imágenes Peabody (TVIP)*. Circle Pines, MN: American Guidance Service.

Evdokimov, K. S. and A. Zeleneev (2025). Simple estimation of semiparametric models with measurement errors.

Fredriksson, P., B. Ockert, and H. Oosterbeek (2016). Parental responses to public investments in children: Evidence from a maximum class size rule. *Journal of Human Resources 51*(4), 832–868.

Freyberger, J. (2021). Normalizations and misspecification in skill formation models. Technical report, ArXiv e-prints. Revised July 2022.

Goff, L., O. Malamud, C. Pop-Eleches, and M. Urquiola (2025). Interactions between family and school environments. *Journal of Human Resources 60*(3), 907–949.

Greaves, E., I. Hussain, B. Rabe, and I. Rasul (2023). Parental responses to information about school quality: Evidence from linked survey and administrative data. *Economic Journal 133*(654), 2334–2402.

Hanushek, E. A. and S. G. Rivkin (2012). The distribution of teacher quality and implications for policy. *Annual Review of Economics 4*(1), 131–157.

Heckman, J. (2020). Epilogue: Randomization and social policy evaluation revisited. In F. Bedecarrats, I. Guerin, and F. Roubaud (Eds.), *Randomized Control Trials in the Field of Development: A Critical Perspective*. Oxford University Press.

Heckman, J. and S. Mosso (2014). The economics of human development and social mobility. *Annual Review of Economics 6*(1), 689–733.

Heckman, J. and J. Zhou (2026). A study of the microdynamics of early childhood learning. *Journal of Political Economy*. Forthcoming.

Heckman, J. J. (2000a). Causal parameters and policy analysis in economics. *Quarterly Journal of Economics 115*(1), 45–97.

Heckman, J. J. (2000b). Policies to foster human capital. *Research in Economics 54*(1), 3–56.

Houtenville, A. and K. Smith Conway (2008). Parental effort, school resources and student achievement. *Journal of Human Resources 43*(2), 437–453.

Jackson, C. K., J. E. Rockoff, and D. Staiger (2014). Teacher effects and teacher-related policies. *Annual Review of Economics 6*(1), 801–825.

Jacob, B. A., L. Lefgren, and D. P. Sims (2010). The persistence of teacher-induced learning gains. *Journal of Human Resources 45*(4), 915–943.

Jochmans, K. (2023). Testing random assignment to peer groups. *Journal of Applied Econometrics 38*(3), 321–333.

Johnson, R. C. and C. K. Jackson (2019). Reducing inequality through dynamic complementarity: Evidence from head start and public school spending. *American Economic Journal: Economic Policy 11*(4), 310–349.

Kinsler, J. (2016). Teacher complementarities in test score production: Evidence from primary school. *Journal of Labor Economics 34*(1), 29–61.

Pop-Eleches, C. and M. Urquiola (2023). Going to a better school: Effects and behavioral responses. *Economics Journal 133*(August), 2334–2402.

Pop-Eleches, C. and M. Urquiola (2013). Going to a better school: Effects and behavioral responses. *American Economic Review 103*(4), 1289–1324.

Sacerdote, B. (2001). Peer effects with random assignment: Results for dartmouth roommates. *The Quarterly Journal of Economics 116*(2), 681–704.

Schady, N. (2012). El desarrollo infantil temprano en américa latina y el caribe: Acceso, resultados y evidencia longitudinal de ecuador. In M. Cabrol and M. Székely (Eds.), *Educación para la Transformación*. Washington, D.C.: Inter-American Development Bank.

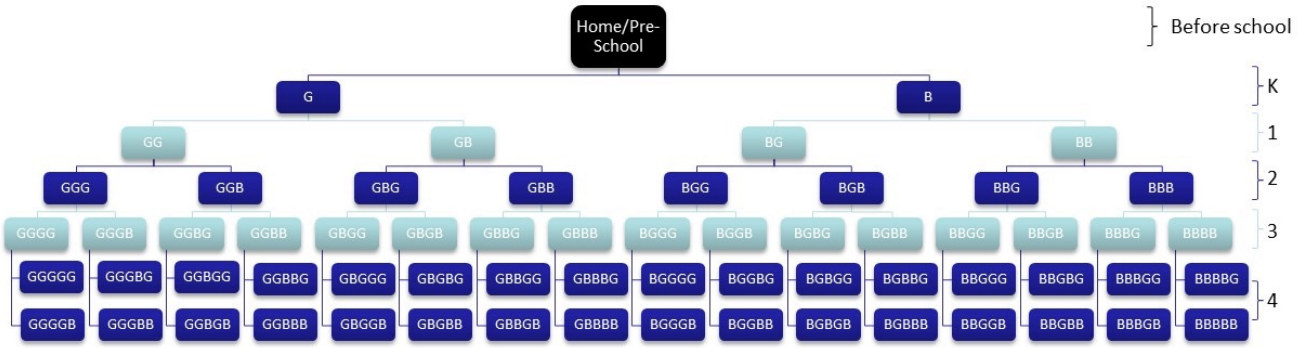
Schady, N., J. R. Behrman, M. C. Araujo, R. Azuero, R. Bernal, D. Bravo, F. Lopez-Boo, K. Macours, D. Marshall, C. Paxson, and R. Vakis (2015). Wealth gradients in early childhood

cognitive development in five latin american countries. *Journal of Human Resources 50*(2), 446–463.

Todd, P. and K. Wolpin (2003). On the specification and estimation of the production function for cognitive achievement. *Economic Journal 113*(485), F3–F33.

Figures and Tables

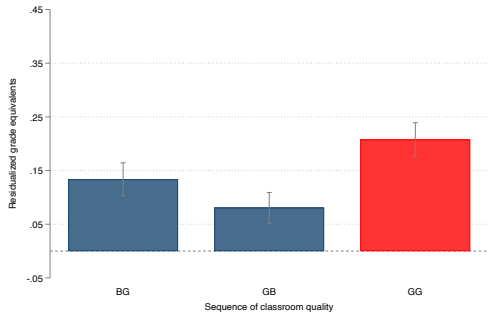
Figure 1: Sequences



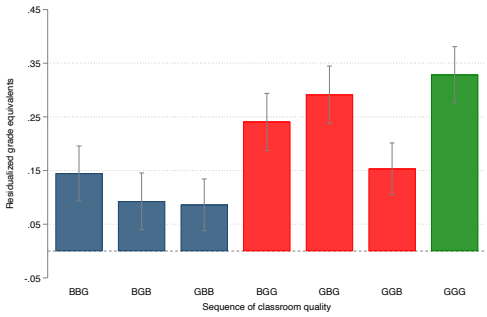
Notes: This figure shows the sequences of classroom quality which are possible between kindergarten and fourth grade. G denotes a classroom with value added above the average in the school, where B denotes a classroom with below school average value added.

Figure 2: Impact of Sequences of Classroom Quality on Achievement

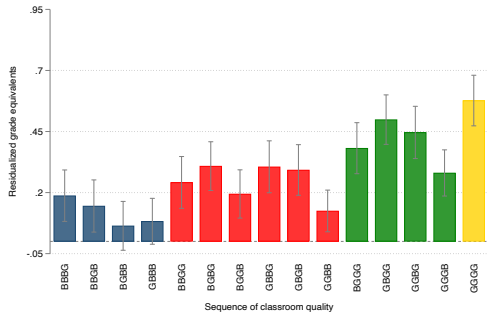
(a) First Grade



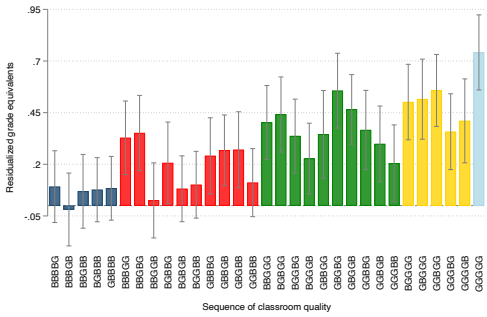
(b) Second Grade



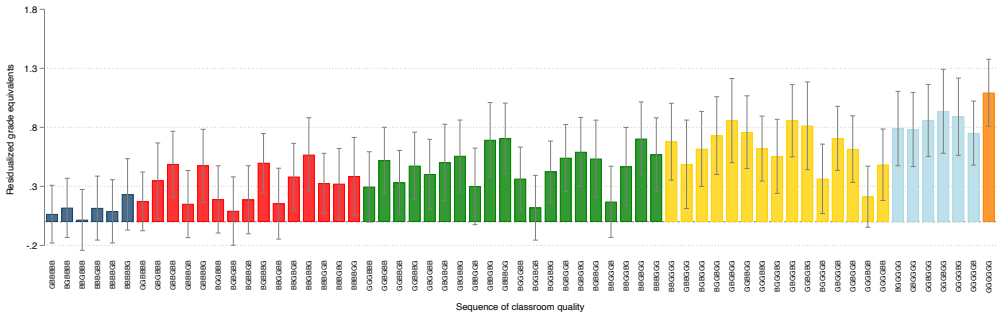
(c) Third Grade



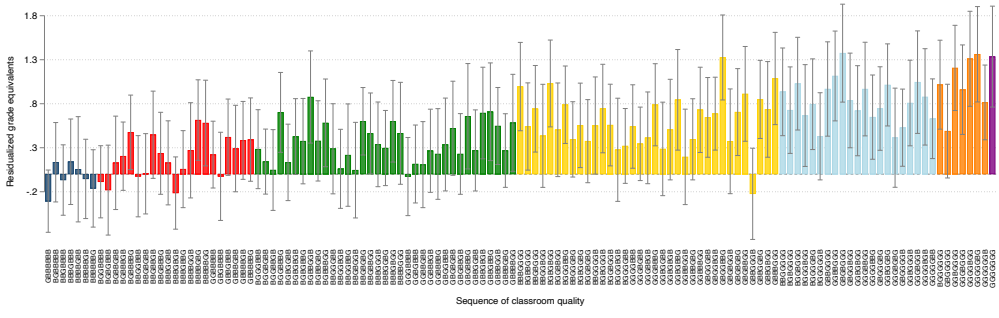
(d) Fourth Grade



(e) Fifth Grade



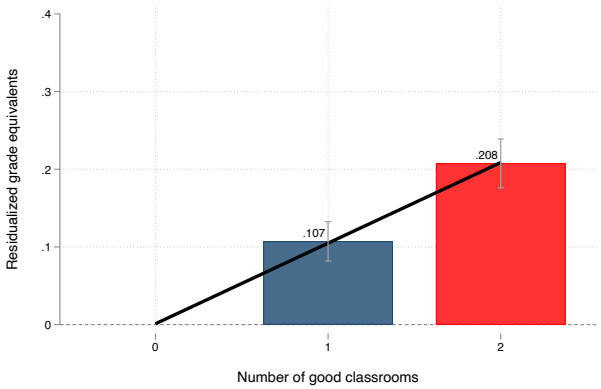
(f) Sixth Grade



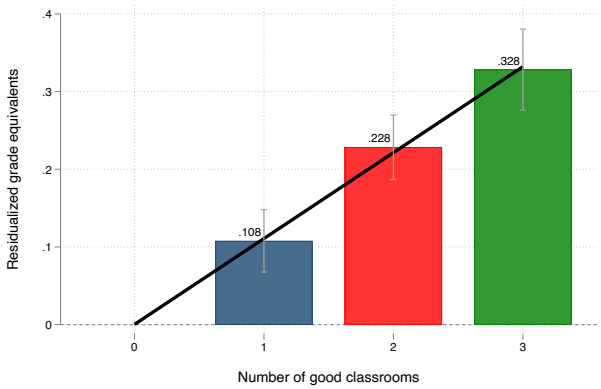
Notes: Each panel in this figure shows the average residual learning at the end of each grade for students in different sequences of classroom quality. B and G indicate below or above school average classroom quality, respectively (e.g., GBBGG in panel D means above average classroom quality in Kindergarten, 3rd and 4th grades; below average in the remaining grades). Residual learning is achievement in math and language at the end of a grade, after controlling for age, gender, baseline TVIP, maternal education, preschool attendance, wealth, and school fixed effects.

Figure 3: Impact of the Number of Good Classrooms across Grades on Achievement

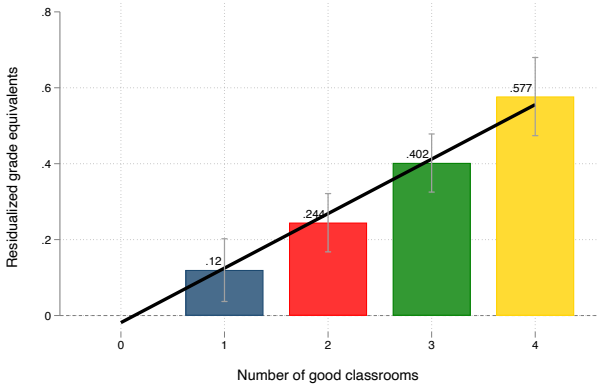
(a) First Grade



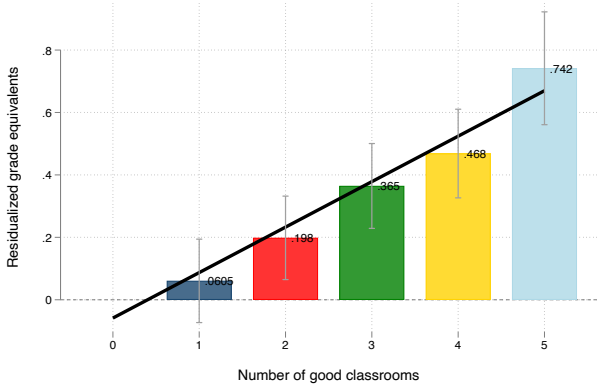
(b) Second Grade



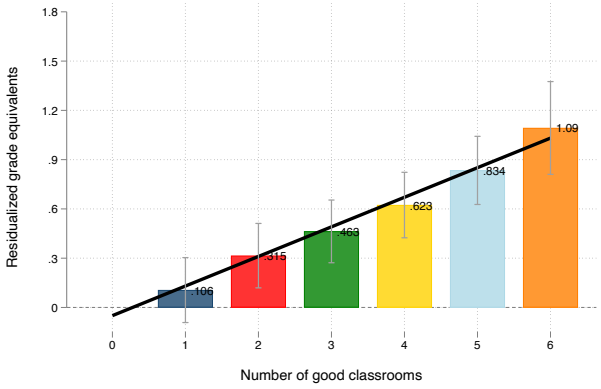
(c) Third Grade



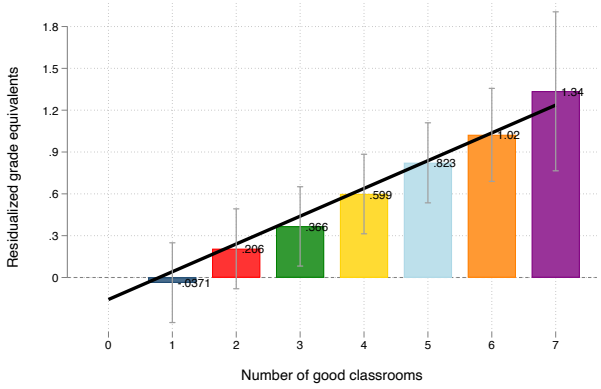
(d) Fourth Grade



(e) Fifth Grade



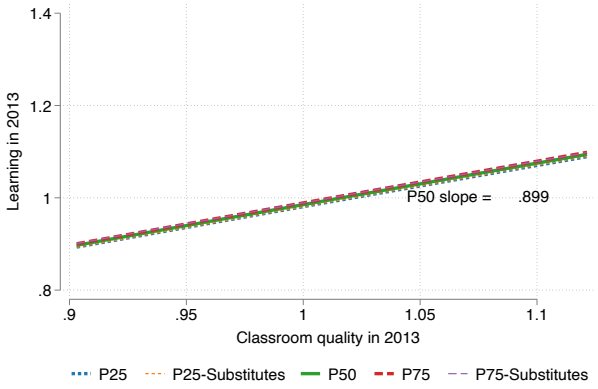
(f) Sixth Grade



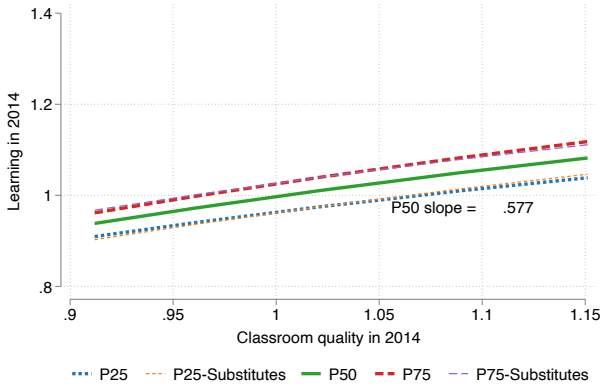
Notes: Each panel in this figure shows the average residual learning at the end of each grade for students in sequences with different numbers of good classrooms, relative to students with zero good classrooms up to the grade where achievement is measured. Residual learning is achievement in math and language at the end of a grade, after controlling for age, gender, baseline TVIP, maternal education, preschool attendance, wealth, and school fixed effects.

Figure 4: Impact of Classroom Quality on Achievement at Different Values (Percentiles 25 and 75) of Past Classroom Quality

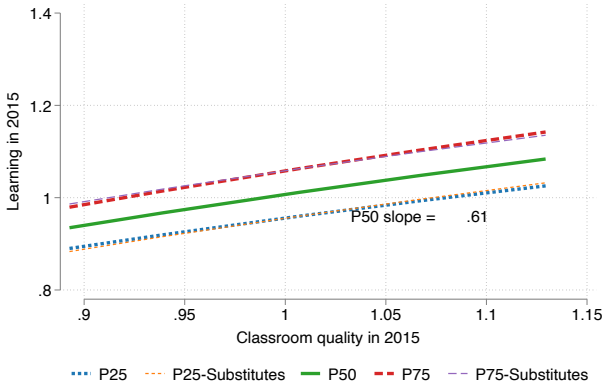
(a) Impact of Quality in First Grade on Learning in First Grade



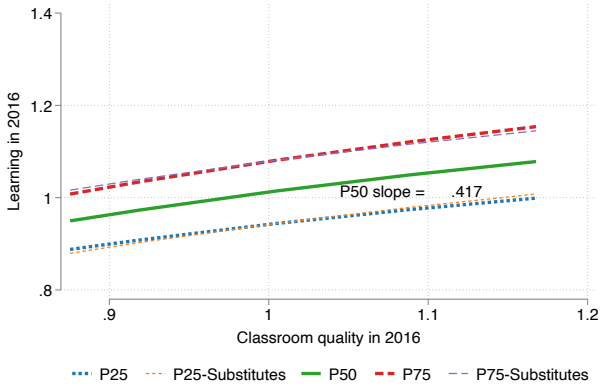
(b) Impact of Quality in Second Grade on Learning in Second Grade



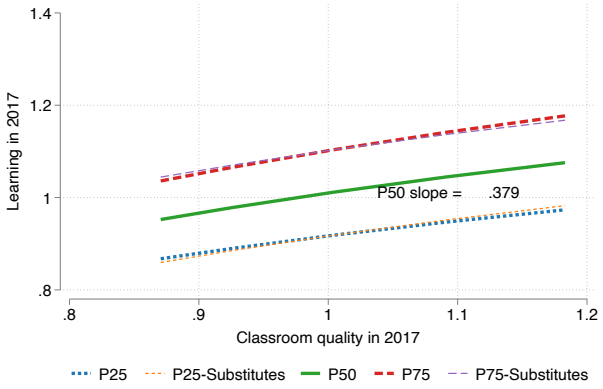
(c) Impact of Quality in Third Grade on Learning in Third Grade



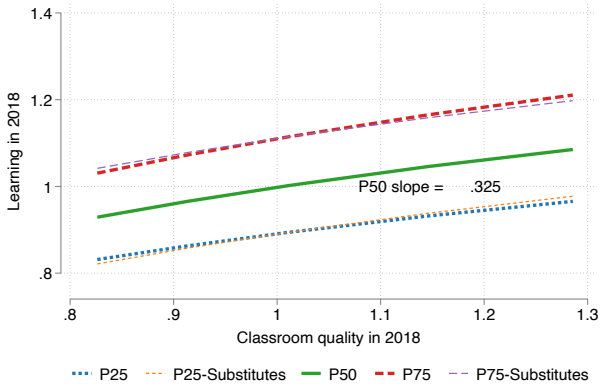
(d) Impact of Quality in Fourth Grade on Learning in Fourth Grade



(e) Impact of Quality in Fifth Grade on Learning in Fifth Grade



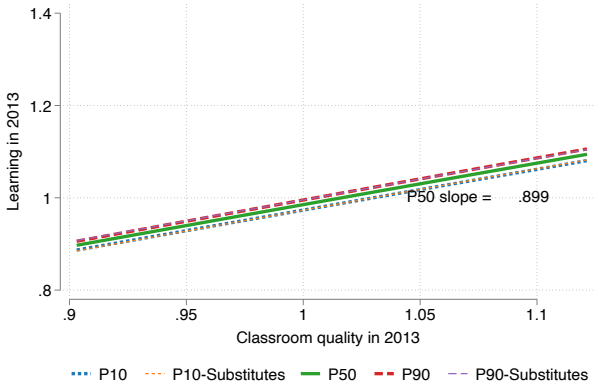
(f) Impact of Quality in Sixth Grade on Learning in Sixth Grade



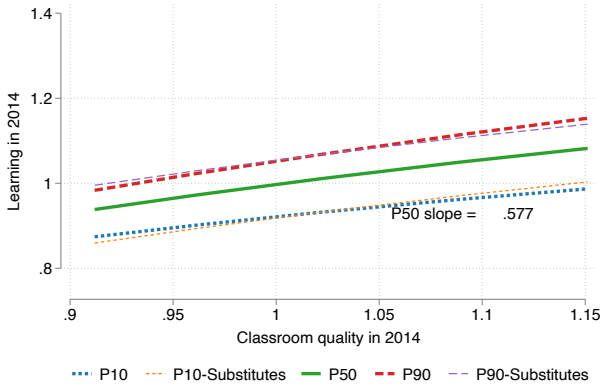
Notes: Each panel in this figure shows predicted average residual learning at the end of each grade (1st, 2nd, 3rd, 4th, 5th, 6th) for students experiencing different levels of classroom quality in that grade, keeping classroom quality in each of the previous grades fixed at the 25th (solid dotted line), 50th (solid line) and 75th percentiles (solid dashed line) of the distribution of classroom quality in those grades. Predictions are generated by the estimated CES production function for each grade, evaluated at the 10th, 25th, 50th, 75th and 90th percentiles of the distribution of inputs. Residual learning is achievement in math and language at the end of a grade, after controlling for age, gender, baseline TVIP, maternal education, preschool attendance, and wealth, as well as school fixed effects.

Figure 5: Impact of Classroom Quality on Achievement at Different Values (Percentiles 10 and 90) of Past Classroom Quality

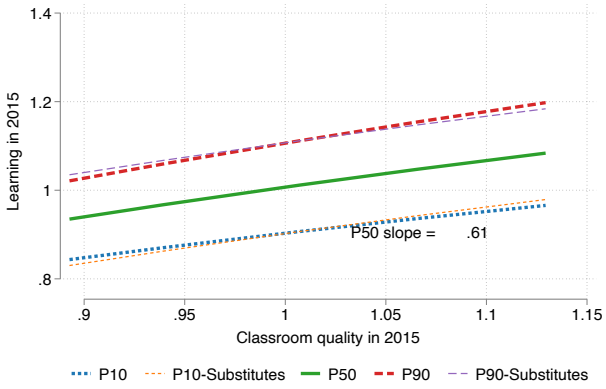
(a) Impact of Quality in First Grade on Learning in First Grade



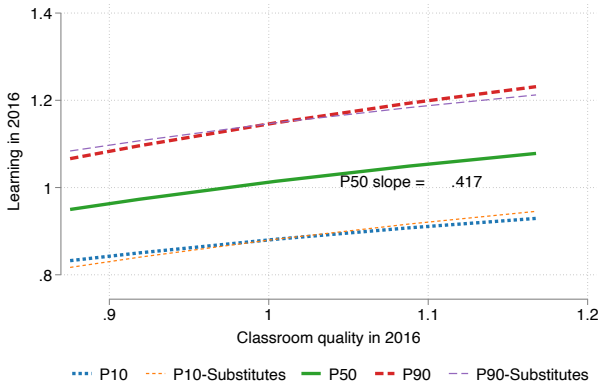
(b) Impact of Quality in Second Grade on Learning in Second Grade



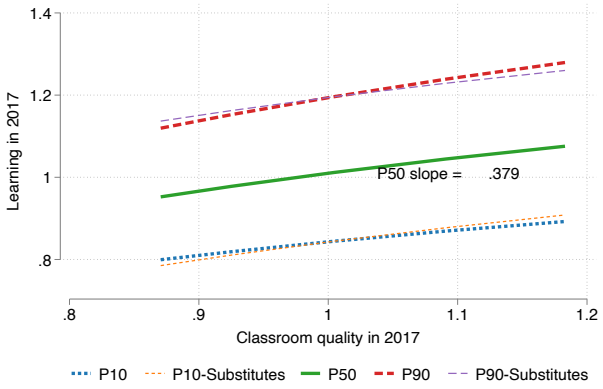
(c) Impact of Quality in Third Grade on Learning in Third Grade



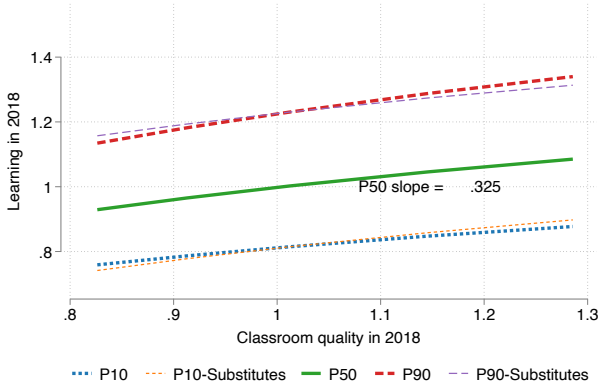
(d) Impact of Quality in Fourth Grade on Learning in Fourth Grade



(e) Impact of Quality in Fifth Grade on Learning in Fifth Grade



(f) Impact of Quality in Sixth Grade on Learning in Sixth Grade



Notes: Each panel in this figure shows predicted average residual learning at the end of each grade (1st, 2nd, 3rd, 4th, 5th, 6th) for students experiencing different levels of classroom quality in that grade, keeping classroom quality in each of the previous grades fixed at the 25th (solid dotted line), 50th (solid line) and 75th percentiles (solid dashed line) of the distribution of classroom quality in those grades. Predictions are generated by the estimated CES production function for each grade, evaluated at the 10th, 25th, 50th, 75th and 90th percentiles of the distribution of inputs. Residual learning is achievement in math and language at the end of a grade, after controlling for age, gender, baseline TVIP, maternal education, preschool attendance, and wealth, as well as school fixed effects.

Table 1: Descriptive Statistics

	Panel A: Children				
	Kindergarten				
Age (months)	67.475				
	(4.103)				
Proportion female	0.501				
	(0.500)				
TVIP	83.872				
	(16.634)				
Mother's age	30.608				
	(6.638)				
Father's age	34.798				
	(8.014)				
Mother's years of schooling	8.784				
	(3.724)				
Father's years of schooling	8.417				
	(3.694)				
	Panel B: Teachers				
	Kindergarten	First grade	Second grade	Third grade	Fourth grade
Age	42.232	45.060	46.130	43.936	43.948
	(9.577)	(10.689)	(9.955)	(10.656)	(9.544)
Proportion female	0.989	0.938	0.871	0.783	0.781
	(0.105)	(0.242)	(0.336)	(0.413)	(0.414)
Experience	14.914	18.986	20.323	17.983	17.181
	(8.884)	(10.448)	(10.837)	(11.042)	(10.168)
Proportion tenure	0.640	0.717	0.883	0.831	0.833
	(0.481)	(0.451)	(0.321)	(0.375)	(0.373)
Years of schooling	17.140	17.455	17.540	17.483	18.013
	(1.932)	(2.061)	(2.530)	(2.284)	(2.431)
CLASS score	3.407	3.289	3.337	3.281	3.394
	(0.283)	(0.232)	(0.242)	(0.240)	(0.185)
Class size	34.543	37.809	39.459	37.183	38.681
	(8.155)	(7.679)	(7.740)	(6.895)	(6.647)

Notes: Panel A shows means and standard deviations of student and family characteristics in our sample at the start of Kindergarten. Panel B shows means and standard deviations of teacher characteristics and class size in each grade. The TVIP is the Test de Vocabulario en Imágenes Peabody, the Spanish version of the Peabody Picture Vocabulary Test (PPVT). The test is standardized using the tables provided by the test developers which set the mean at 100 and the standard deviation at 15 at each age.

Table 2: Within School Standard Deviations of Classroom Effects

	Uncorrected	Corrected
Kindergarten	0.052	0.034
First grade	0.097	0.063
Second grade	0.100	0.072
Third grade	0.127	0.098
Fourth grade	0.159	0.123
Fifth grade	0.206	0.164
Sixth grade	0.229	0.178

Notes: This table shows the within-school standard deviations of classroom effects for each grade. It shows uncorrected estimates (column 1) and estimates corrected for sampling error (column 2). We control for a quartic polynomial in baseline test scores, maternal education, preschool attendance, and wealth, as well as child gender and age.

Table 3: Average Differences in Value-Added between Good and Bad Classrooms

Statistic	Kindergarten	First grade	Second grade	Third grade	Fourth grade	Fifth grade	Sixth grade
Mean	0.081	0.152	0.155	0.202	0.241	0.308	0.361
p10	0.014	0.030	0.020	0.042	0.040	0.054	0.051
p25	0.034	0.064	0.062	0.097	0.094	0.131	0.146
p50	0.067	0.131	0.126	0.185	0.205	0.255	0.301
p75	0.106	0.216	0.217	0.284	0.363	0.447	0.489
p90	0.165	0.291	0.308	0.388	0.465	0.599	0.723

Notes: This table shows the average value-added differences between good and bad classrooms in our sample of schools for each grade. We report the mean and various percentiles of the distribution of these differences across schools.

Table 4: More is Better and Linearity Tests

	Grade					
	First grade	Second grade	Third grade	Fourth grade	Fifth grade	Sixth grade
More is better	0.000	0.000	0.000	0.000	0.000	0.000
Linearity	0.967	0.878	0.828	0.574	0.895	0.957

Notes: This table shows p-values for the 'More is Better' and 'Linearity' tests for each grade. The 'More is Better' test examines whether a higher number of goods classrooms leads to higher achievement. It corresponds to testing whether all the bars in Figure 3 have the same height. The 'Linearity' test examines whether the relationship between the number of good classrooms and achievement is linear. It corresponds to testing whether the bars in Figure 3 are equal to the predicted points in the plotted regression line.

Table 5: Testing Interactions Between Classroom Effects in Different Grades

	10%	5%	1%
Panel A: Control for Baseline Scores			
First Grade	0.090	0.034	0.009
Second Grade	0.090	0.031	0.004
Third Grade	0.102	0.054	0.014
Fourth Grade	0.092	0.046	0.000
Fifth Grade	0.057	0.032	0.019
Sixth Grade	0.039	0.039	0.039
Panel B: Control for t-1 Scores			
First Grade	0.086	0.056	0.011
Second Grade	0.096	0.050	0.009
Third Grade	0.102	0.065	0.014
Fourth Grade	0.096	0.045	0.014
Fifth Grade	0.166	0.071	0.024
Sixth Grade	0.054	0.018	0.018

Notes: This table shows the proportion of schools at the end of each grade for which the p-value of the joint test that there are no interactions between classroom effects in different grades is less than 10% (column 1), 5% (column 2), or 1% (column 3). For each grade, we regress achievement in math and language at the end of that grade on current and past classroom assignments and their interactions. We control for a quartic polynomial in baseline test scores (Panel A) or lagged test scores (Panel B), maternal education, preschool attendance, and wealth, as well as child gender and age.

Table 6: Estimates of the Marginal Product of Classroom Quality in Different Grades

	Grade					
	First grade	Second grade	Third grade	Fourth grade	Fifth grade	Sixth grade
Marginal Product of	0.899	0.578	0.610	0.425	0.380	0.331
Classroom Quality	(0.025)	(0.030)	(0.022)	(0.019)	(0.017)	(0.013)

Notes: This table shows estimates of the marginal product of classroom quality in different grades at the median classroom quality in the current and all previous grades. Each column corresponds to a different production function, where the output is achievement at the end of grades 1 through 6, and inputs are all current and lagged levels of classroom quality.

"Dynamic Complementarity in Elementary Schools: Experimental Estimates from Ecuador"

Online Appendix

Pedro Carneiro
Yyannú Cruz-Aguayo
Rafael Hernández-Pachón
Norbert Schady

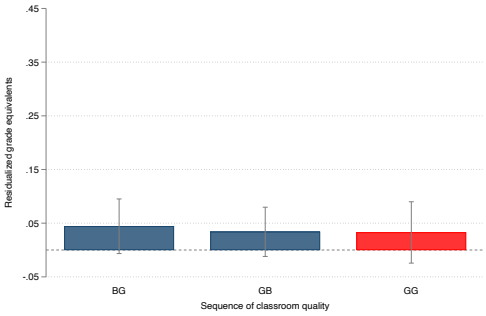
Table of Contents

A Figures and Tables

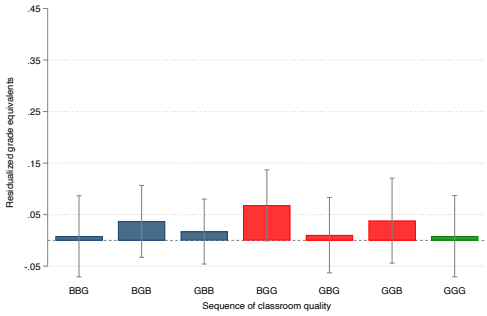
(This page intentionally left blank)

Figure A.1: Impact (Placebo) of Sequences of Classroom Quality on Baseline TVIP

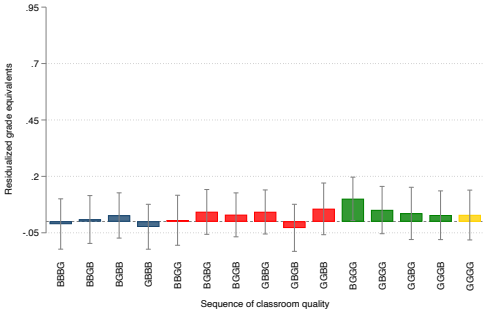
(a) First Grade



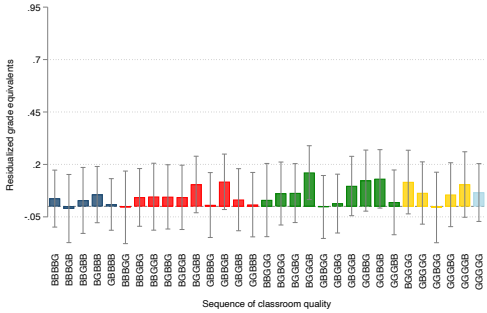
(b) Second Grade



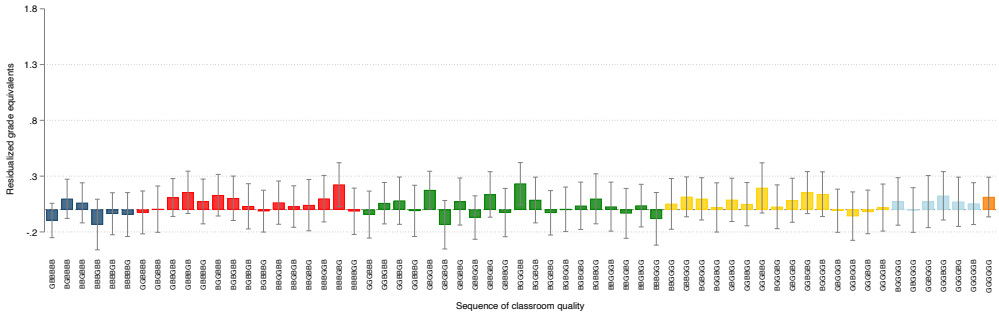
(c) Third Grade



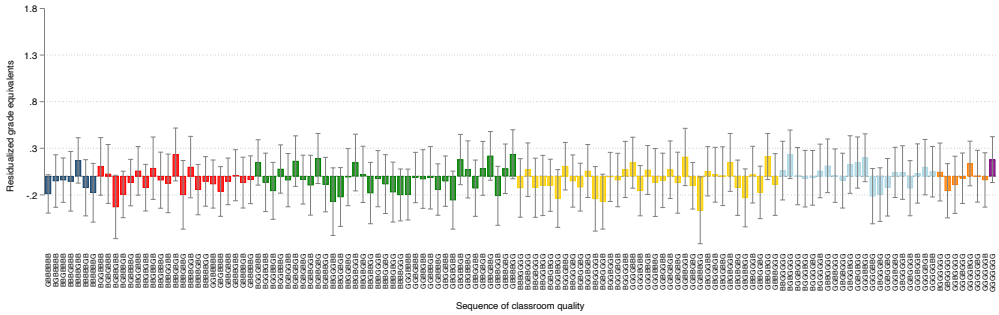
(d) Fourth Grade



(e) Fifth Grade



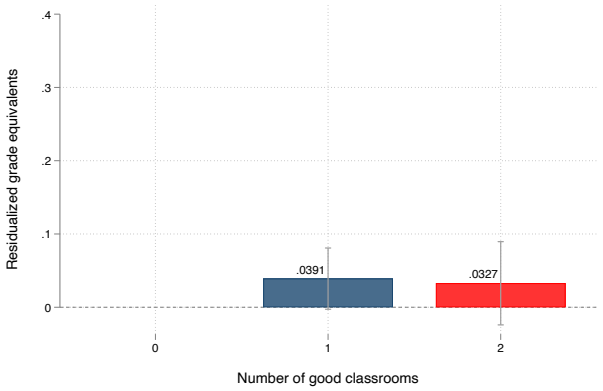
(f) Sixth Grade



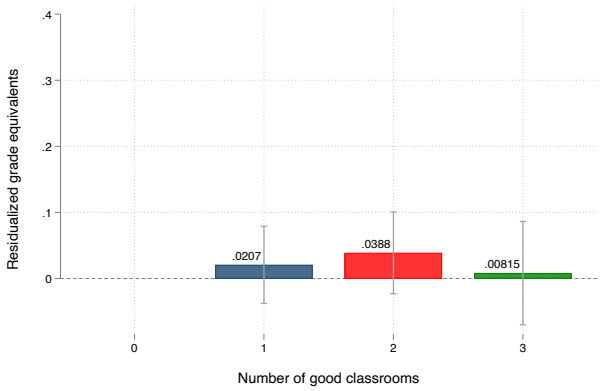
Notes: Each panel in this figure shows the average baseline TVIP (in grade equivalents) for students in different sequences of classroom quality. B and G indicate below or above school average classroom quality, respectively (e.g., GBBGG in panel D means above average classroom quality in Kindergarten, 3rd and 4th grades; below average in the remaining grades). All regressions control for age, gender, maternal education, preschool attendance, wealth, and school fixed effects.

Figure A.2: Impact (Placebo) of the Number of Good Classrooms across Grades on Baseline TVIP

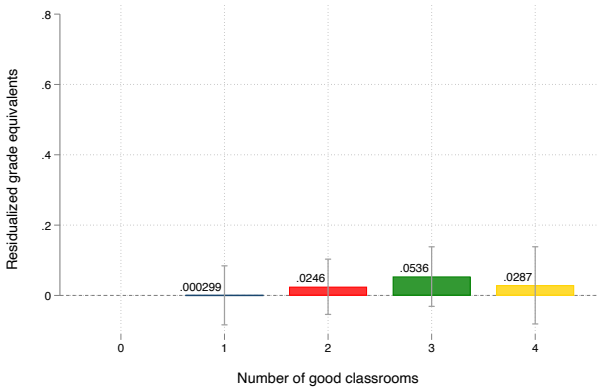
(a) First Grade



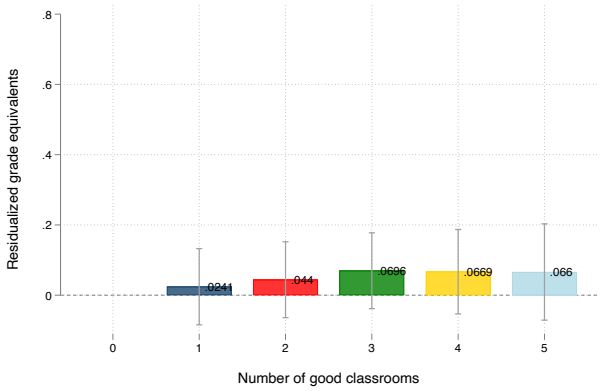
(b) Second Grade



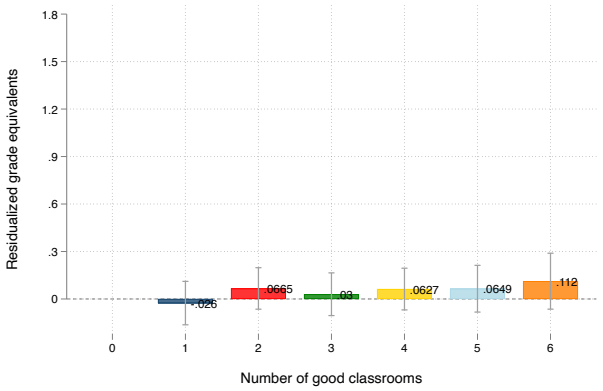
(c) Third Grade



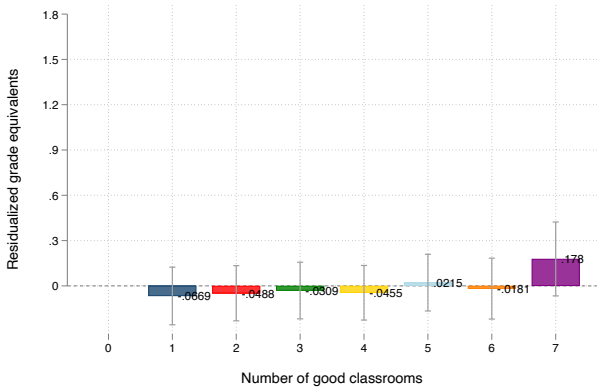
(d) Fourth Grade



(e) Fifth Grade



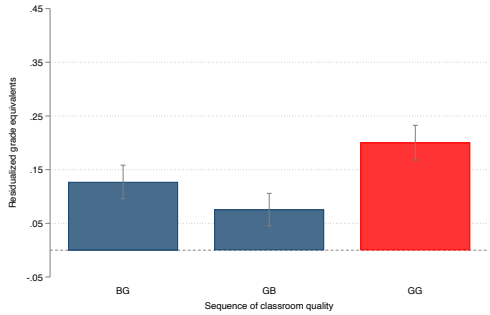
(f) Sixth Grade



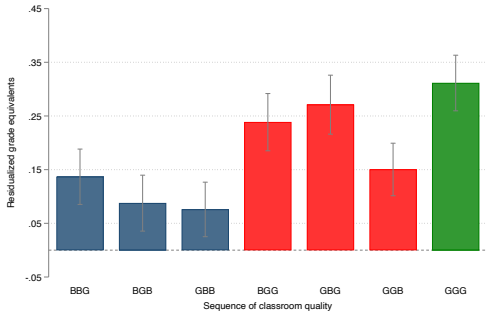
Notes: Each panel in this figure shows the average baseline TVIP (in grade equivalents) for students in sequences with different numbers of good classrooms, relative to students with zero good classrooms up to the grade where achievement is measured. Residual learning is achievement in math and language at the end of a grade, after controlling for age, gender, baseline TVIP, maternal education, preschool attendance, wealth, and school fixed effects.

Figure A.3: Impact of Sequences of Classroom Quality on Achievement, Leave-One-Out Estimates

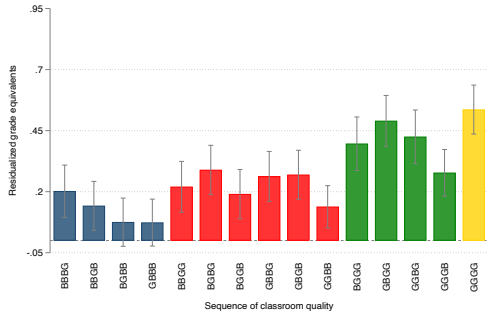
(a) First Grade



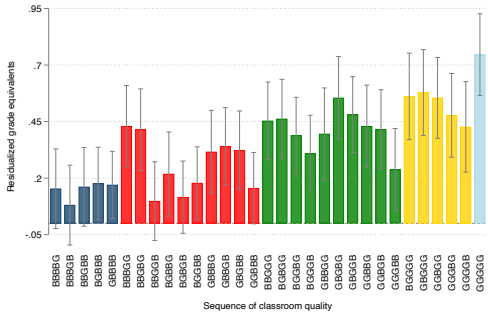
(b) Second Grade



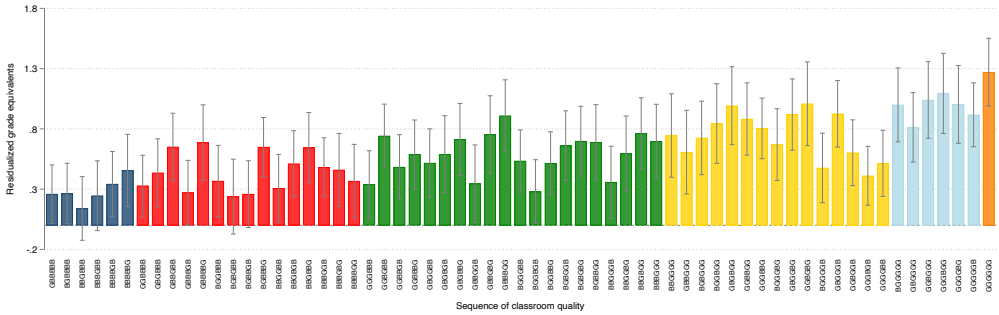
(c) Third Grade



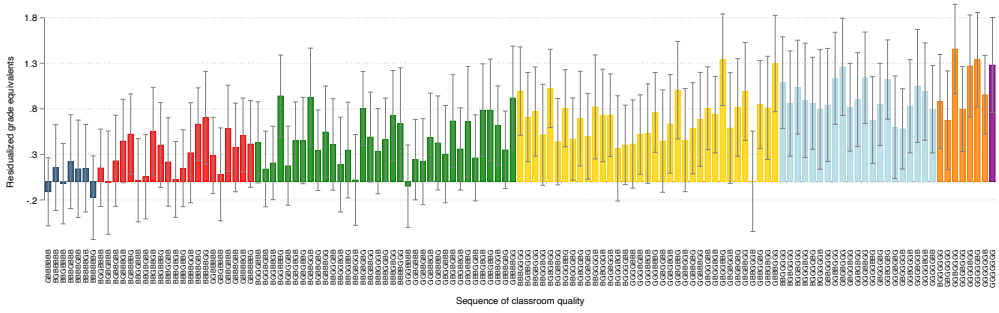
(d) Fourth Grade



(e) Fifth Grade



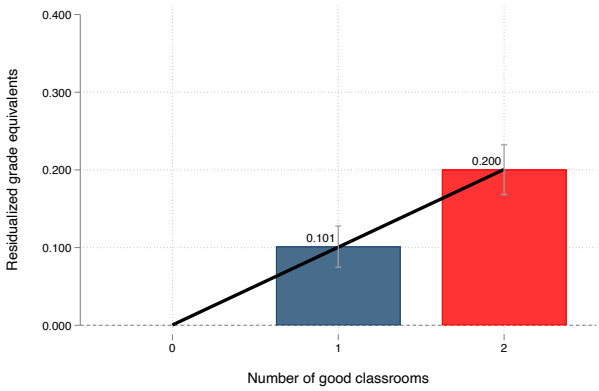
(f) Sixth Grade



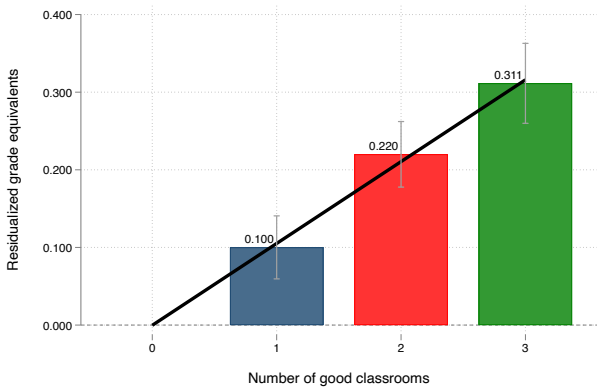
Notes: Each panel in this figure shows the average residual learning at the end of each grade for students in different sequences of classroom quality. B and G indicate below or above school average classroom quality, respectively (e.g., GBBGG in panel D means above average classroom quality in Kindergarten, 3rd and 4th grades; below average in the remaining grades). Residual learning is achievement in math and language at the end of a grade, after controlling for age, gender, baseline TVIP, maternal education, preschool attendance, wealth, and school fixed effects. Leave-one-out estimates remove student i and students with the same classroom achievement rank from other classrooms when computing classroom and school value added.

Figure A.4: Impact of the Number of Good Classrooms across Grades on Achievement, Leave-One-Out Estimates

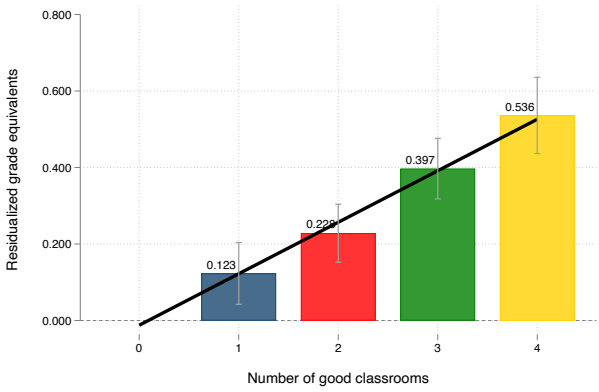
(a) First Grade



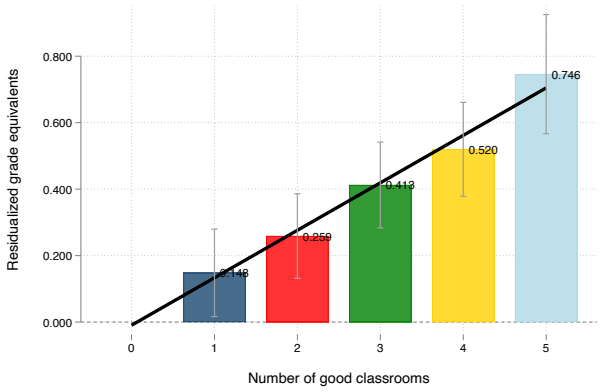
(b) Second Grade



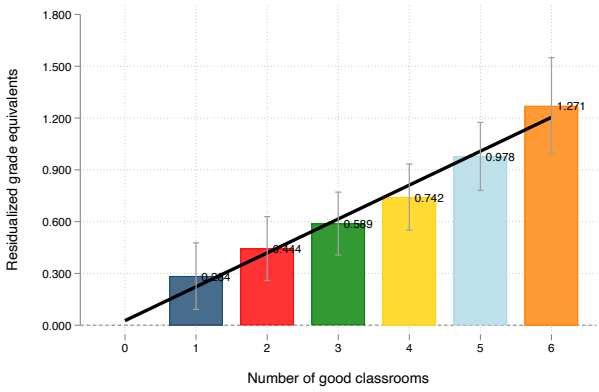
(c) Third Grade



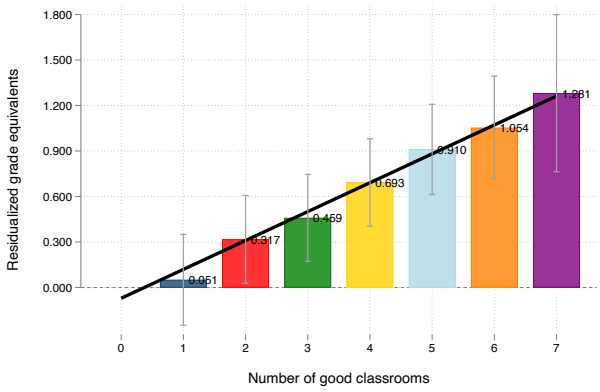
(d) Fourth Grade



(e) Fifth Grade



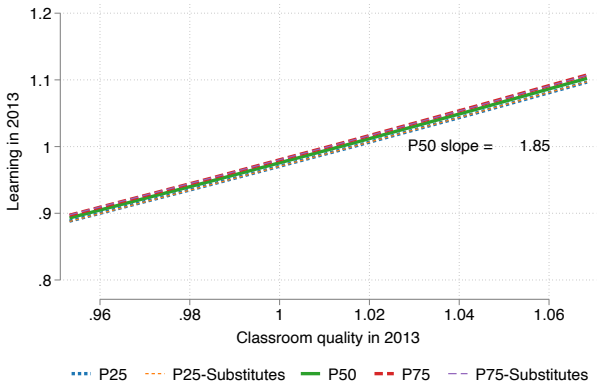
(f) Sixth Grade



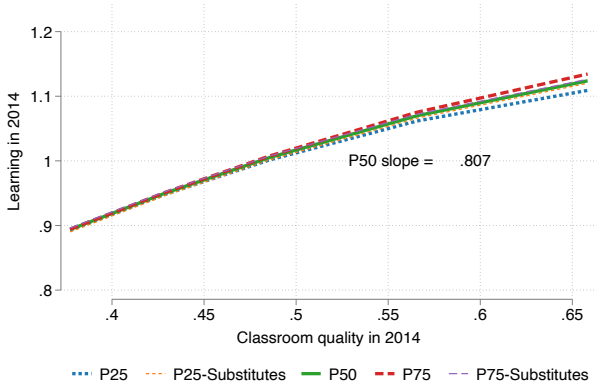
Notes: Each panel in this figure shows the average residual learning at the end of each grade for students in sequences with different numbers of good classrooms, relative to students with zero good classrooms up to the grade where achievement is measured. Residual learning is achievement in math and language at the end of a grade, after controlling for age, gender, baseline TVIP, maternal education, preschool attendance, wealth, and school fixed effects. Leave-one-out estimates remove student i and students with the same classroom achievement rank from other classrooms when computing classroom and school value added.

Figure A.5: Impact of Classroom Quality on Achievement at Different Values (Percentiles 25 and 75) of Past Classroom Quality. Measurement Error Corrected (Optimistic Case)

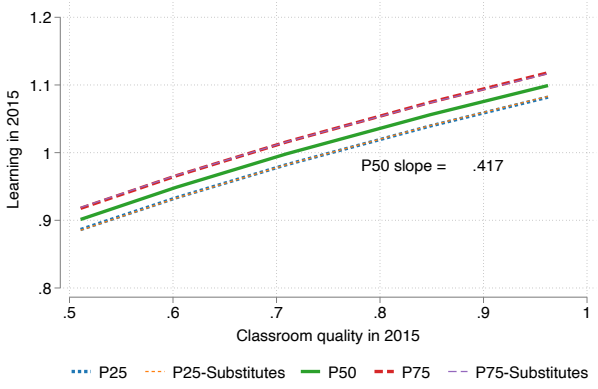
(a) Impact of Quality in First Grade on Learning in First Grade



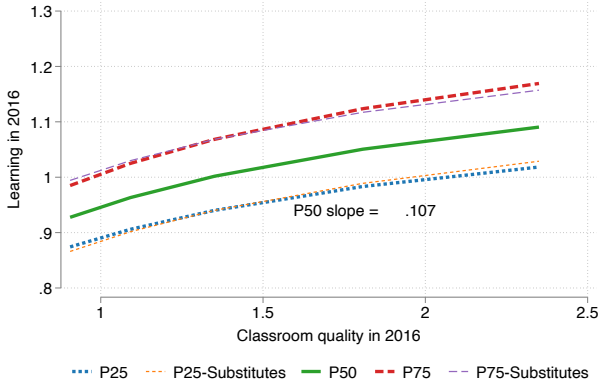
(b) Impact of Quality in Second Grade on Learning in Second Grade



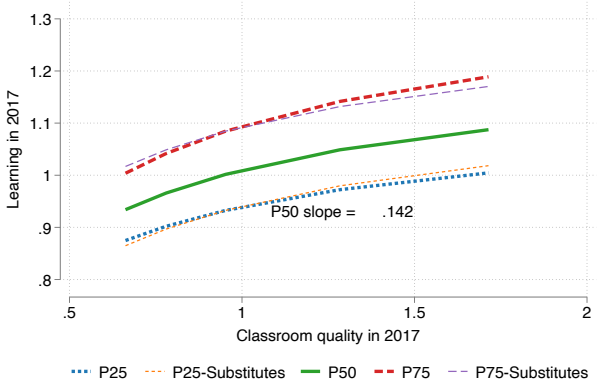
(c) Impact of Quality in Third Grade on Learning in Third Grade



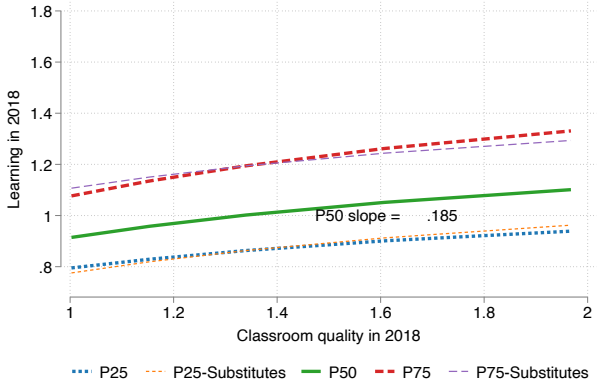
(d) Impact of Quality in Fourth Grade on Learning in Fourth Grade



(e) Impact of Quality in Fifth Grade on Learning in Fifth Grade



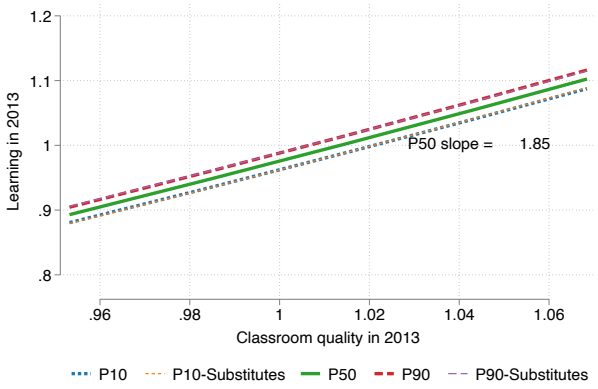
(f) Impact of Quality in Sixth Grade on Learning in Sixth Grade



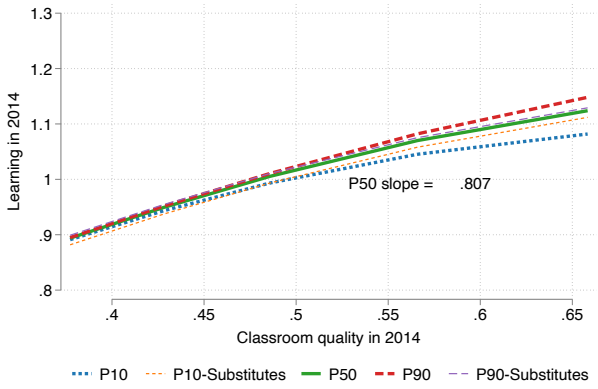
Notes: Each panel in this figure shows predicted average residual learning at the end of each grade (1st, 2nd, 3rd, 4th, 5th, 6th) for students experiencing different levels of classroom quality in that grade, keeping classroom quality in each of the previous grades fixed at the 25th (solid dotted line), 50th (solid line) and 75th percentiles (solid dashed line) of the distribution of classroom quality in those grades. Predictions are generated by the estimated CES production function for each grade, evaluated at the 10th, 25th, 50th, 75th and 90th percentiles of the distribution of inputs. All estimates are measurement error corrected, as explained in Appendix E. The variance of measurement error of previous classroom inputs is calibrated to be 0.0023 (optimistic case). Residual learning is achievement in math and language at the end of a grade, after controlling for age, gender, baseline TVIP, maternal education, preschool attendance, and wealth, as well as school fixed effects.

Figure A.6: Impact of Classroom Quality on Achievement at Different Values (Percentiles 10 and 90) of Past Classroom Quality. Measurement Error Corrected (Optimistic Case)

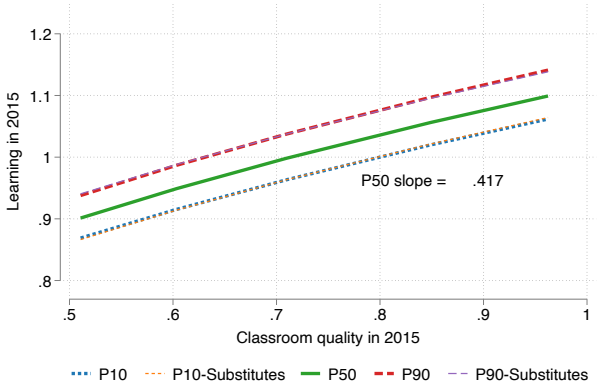
(a) Impact of Quality in First Grade on Learning in First Grade



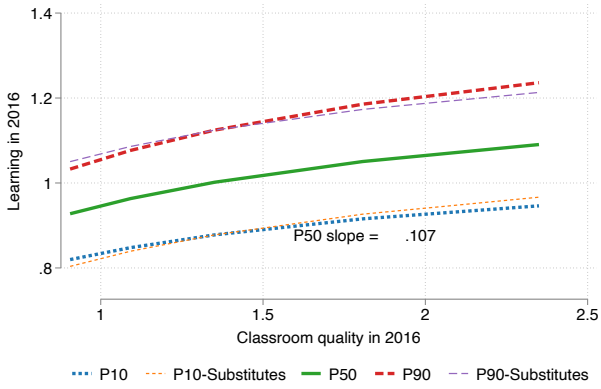
(b) Impact of Quality in Second Grade on Learning in Second Grade



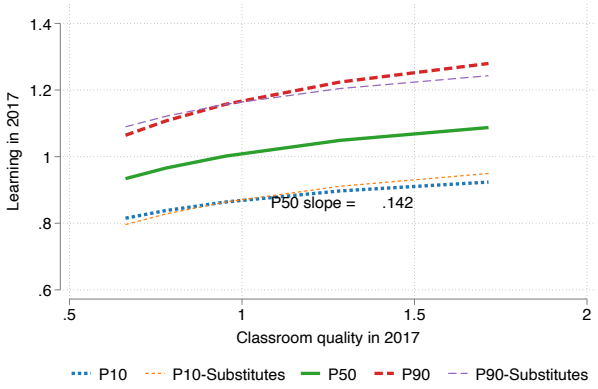
(c) Impact of Quality in Third Grade on Learning in Third Grade



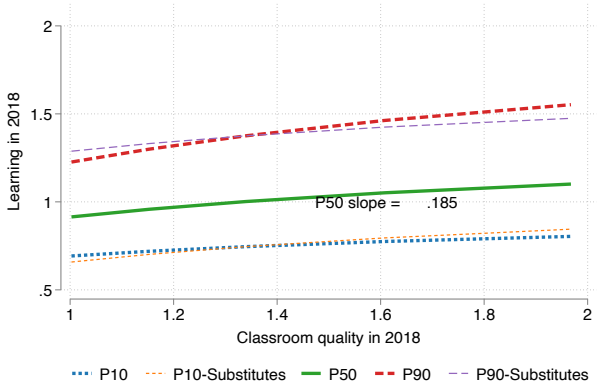
(d) Impact of Quality in Fourth Grade on Learning in Fourth Grade



(e) Impact of Quality in Fifth Grade on Learning in Fifth Grade



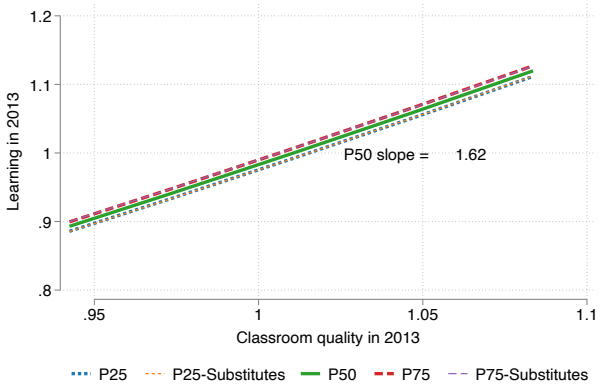
(f) Impact of Quality in Sixth Grade on Learning in Sixth Grade



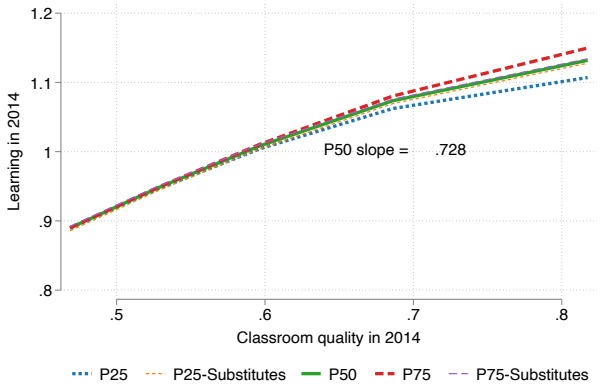
Notes: Each panel in this figure shows predicted average residual learning at the end of each grade (1st, 2nd, 3rd, 4th, 5th, 6th) for students experiencing different levels of classroom quality in that grade, keeping classroom quality in each of the previous grades fixed at the 25th (solid dotted line), 50th (solid line) and 75th percentiles (solid dashed line) of the distribution of classroom quality in those grades. Predictions are generated by the estimated CES production function for each grade, evaluated at the 10th, 25th, 50th, 75th and 90th percentiles of the distribution of inputs. All estimates are measurement error corrected, as explained in Appendix E. The variance of measurement error of previous classroom inputs is calibrated to be 0.0023 (optimistic case). Residual learning is achievement in math and language at the end of a grade, after controlling for age, gender, baseline TVIP, maternal education, preschool attendance, and wealth, as well as school fixed effects.

Figure A.7: Impact of Classroom Quality on Achievement at Different Values (Percentiles 25 and 75) of Past Classroom Quality. Measurement Error Corrected (Pessimistic Case)

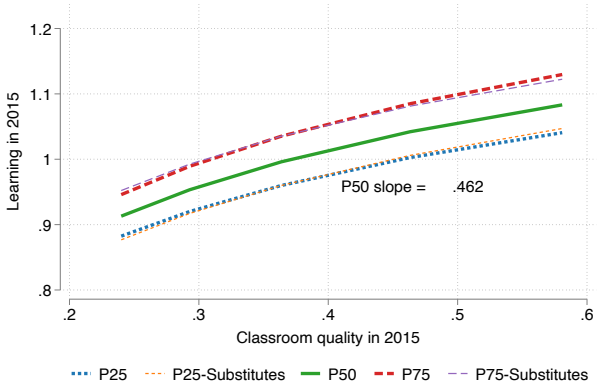
(a) Impact of Quality in First Grade on Learning in First Grade



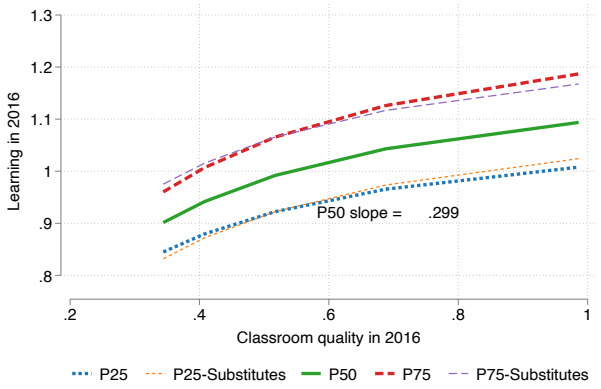
(b) Impact of Quality in Second Grade on Learning in Second Grade



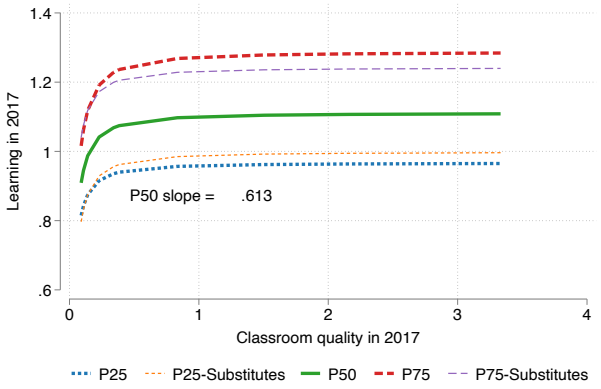
(c) Impact of Quality in Third Grade on Learning in Third Grade



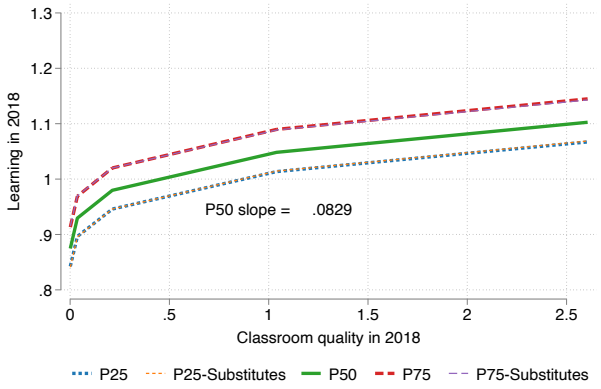
(d) Impact of Quality in Fourth Grade on Learning in Fourth Grade



(e) Impact of Quality in Fifth Grade on Learning in Fifth Grade



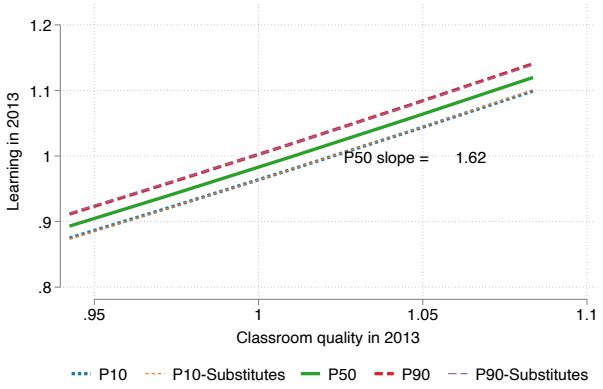
(f) Impact of Quality in Sixth Grade on Learning in Sixth Grade



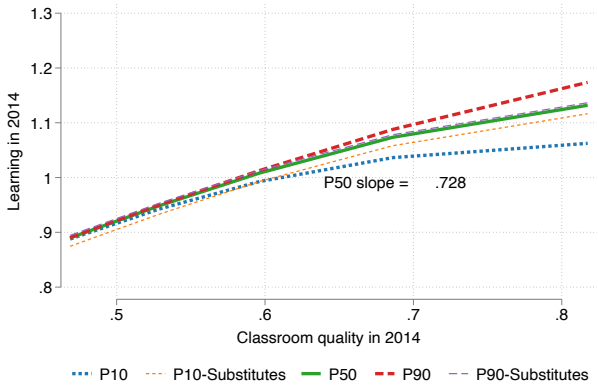
Notes: Each panel in this figure shows predicted average residual learning at the end of each grade (1st, 2nd, 3rd, 4th, 5th, 6th) for students experiencing different levels of classroom quality in that grade, keeping classroom quality in each of the previous grades fixed at the 25th (solid dotted line), 50th (solid line) and 75th percentiles (solid dashed line) of the distribution of classroom quality in those grades. Predictions are generated by the estimated CES production function for each grade, evaluated at the 10th, 25th, 50th, 75th and 90th percentiles of the distribution of inputs. All estimates are measurement error corrected, as explained in Appendix E. The variance of measurement error of previous classroom inputs is calibrated to be 0.0067 (pessimistic case). Residual learning is achievement in math and language at the end of a grade, after controlling for age, gender, baseline TVIP, maternal education, preschool attendance, and wealth, as well as school fixed effects.

Figure A.8: Impact of Classroom Quality on Achievement at Different Values (Percentiles 10 and 90) of Past Classroom Quality. Measurement Error Corrected (Pessimistic Case)

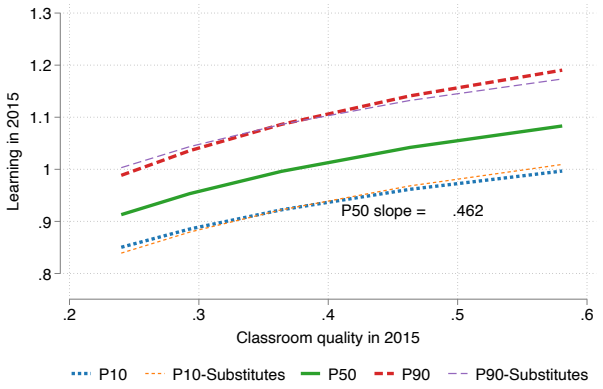
(a) Impact of Quality in First Grade on Learning in First Grade



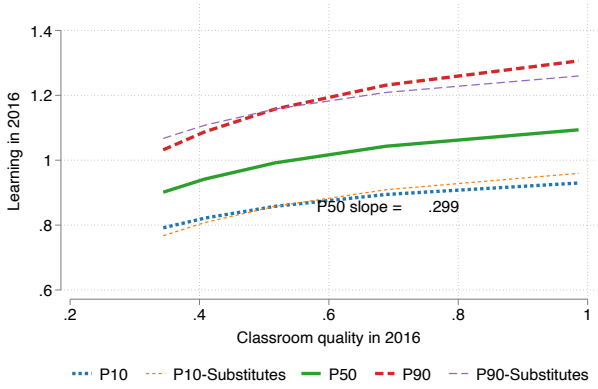
(b) Impact of Quality in Second Grade on Learning in Second Grade



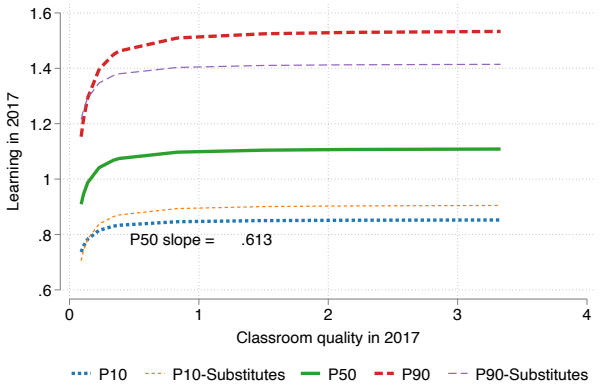
(c) Impact of Quality in Third Grade on Learning in Third Grade



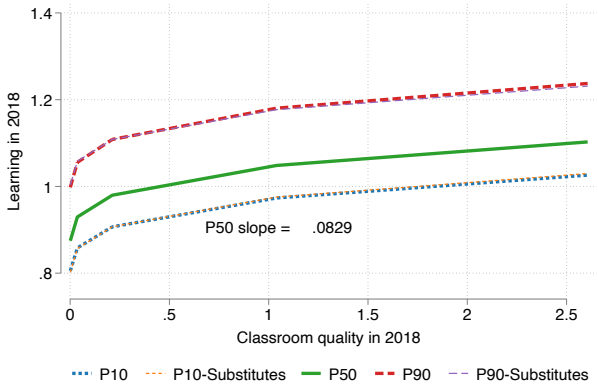
(d) Impact of Quality in Fourth Grade on Learning in Fourth Grade



(e) Impact of Quality in Fifth Grade on Learning in Fifth Grade



(f) Impact of Quality in Sixth Grade on Learning in Sixth Grade



Notes: Each panel in this figure shows predicted average residual learning at the end of each grade (1st, 2nd, 3rd, 4th, 5th, 6th) for students experiencing different levels of classroom quality in that grade, keeping classroom quality in each of the previous grades fixed at the 25th (solid dotted line), 50th (solid line) and 75th percentiles (solid dashed line) of the distribution of classroom quality in those grades. Predictions are generated by the estimated CES production function for each grade, evaluated at the 10th, 25th, 50th, 75th and 90th percentiles of the distribution of inputs. All estimates are measurement error corrected, as explained in Appendix E. The variance of measurement error of previous classroom inputs is calibrated to be 0.0067 (pessimistic case). Residual learning is achievement in math and language at the end of a grade, after controlling for age, gender, baseline TVIP, maternal education, preschool attendance, and wealth, as well as school fixed effects.

Table A.1: Estimates of the Parameters of the Production Function (Uncorrected)

	Grade					
	First grade	Second grade	Third grade	Fourth grade	Fifth grade	Sixth grade
ρ	-0.378	-2.376	-1.220	-2.205	-1.830	-1.117
	(2.400)	(1.341)	(0.791)	(0.724)	(0.689)	(0.303)
θ	0.989	1.202	1.513	1.574	1.830	1.956
	(0.029)	(0.033)	(0.038)	(0.044)	(0.049)	(0.055)
π_0	0.073	0.061	0.044	0.041	0.029	0.031
	(0.015)	(0.015)	(0.010)	(0.010)	(0.008)	(0.009)
π_1	0.927	0.426	0.264	0.181	0.188	0.131
	(0.015)	(0.023)	(0.017)	(0.015)	(0.014)	(0.014)
π_2		0.512	0.282	0.229	0.182	0.156
		(0.030)	(0.017)	(0.014)	(0.012)	(0.011)
π_3			0.410	0.264	0.210	0.181
			(0.013)	(0.013)	(0.011)	(0.011)
π_4				0.285	0.171	0.151
				(0.012)	(0.010)	(0.009)
π_5					0.220	0.164
					(0.009)	(0.009)
π_6						0.187
						(0.008)
Elasticity of substitution	0.726	0.296	0.450	0.312	0.353	0.472
	(1.264)	(0.118)	(0.161)	(0.070)	(0.086)	(0.068)

Notes: This table shows estimates of the parameters of the production function (and the implied elasticity of substitution between inputs in different grades) for each grade. The production function is specified in Equation (8). Standard errors in parentheses.

Table A.2: Distribution of Classroom Inputs in Each Grade (Uncorrected)

Percentile	Kindergarten	First grade	Second grade	Third grade	Fourth grade	Fifth grade	Sixth grade
10	0.852	0.900	0.899	0.886	0.868	0.865	0.826
25	0.941	0.957	0.958	0.936	0.916	0.921	0.912
50	1.005	1.019	1.020	1.001	1.003	1.004	1.012
75	1.062	1.075	1.091	1.064	1.096	1.097	1.153
90	1.163	1.122	1.166	1.128	1.173	1.193	1.295

Notes: This table shows estimates of the distribution of classroom inputs in each grade estimated from the production function. The production function is specified in Equation (8).

Table A.3: Estimates of the Parameters of the Production Function (ME Corrected - Optimistic)

	Grade					
	First grade	Second grade	Third grade	Fourth grade	Fifth grade	Sixth grade
ρ	0.236	-5.900	0.151	-0.637	-0.900	-0.951
θ	1.932	0.498	0.513	0.704	0.682	1.269
π_0	0.045	0.676	0.146	0.091	0.159	0.031
π_1	0.955	0.235	0.071	0.181	0.082	0.140
π_2		0.089	0.170	0.086	0.077	0.057
π_3			0.613	0.321	0.172	0.153
π_4				0.321	0.253	0.138
π_5					0.259	0.182
π_6						0.298
Elasticity of substitution	1.309	0.145	1.178	0.611	0.526	0.512

Notes: This table shows estimates of the parameters of the production function (and the implied elasticity of substitution between inputs in different grades) for each grade. The production function is specified in Equation (8). These estimates are corrected for measurement error using the optimistic calibration for γ. The details about the measurement error correction can be found in Appendix E.

Table A.4: Distribution of Classroom Inputs in Each Grade (ME Corrected - Optimistic)

Percentile	Kindergarten	First grade	Second grade	Third grade	Fourth grade	Fifth grade	Sixth grade
10	0.852	0.951	0.374	0.502	0.906	0.642	0.995
25	0.941	0.982	0.422	0.594	1.089	0.774	1.152
50	1.005	1.013	0.483	0.701	1.354	0.956	1.355
75	1.062	1.044	0.557	0.846	1.826	1.306	1.634
90	1.163	1.068	0.646	0.959	2.410	1.767	2.173

Notes: This table shows estimates of the distribution of classroom inputs in each grade estimated from the production function. The estimates are corrected for measurement error using the optimistic calibration for γ. The production function is specified in Equation (8).

Table A.5: Estimates of the Parameters of the Production Function (ME Corrected - Pessimistic)

	Grade					
	First grade	Second grade	Third grade	Fourth grade	Fifth grade	Sixth grade
ρ	0.829	-8.578	-0.541	-1.091	-1.313	0.234
θ	1.749	0.542	0.561	0.860	1.009	0.395
π_0	0.073	0.873	0.227	0.245	0.232	0.465
π_1	0.927	0.035	0.082	0.175	0.160	0.125
π_2		0.092	0.424	0.386	0.430	0.012
π_3			0.267	0.025	0.014	0.199
π_4				0.169	0.144	0.045
π_5					0.020	0.036
π_6						0.118
Elasticity of substitution	5.856	0.104	0.649	0.478	0.432	1.305

Notes: This table shows estimates of the parameters of the production function (and the implied elasticity of substitution between inputs in different grades) for each grade. The production function is specified in Equation (8). These estimates are corrected for measurement error using the pessimistic calibration for γ. The details about the measurement error correction can be found in Appendix E.

Table A.6: Distribution of Classroom Inputs in Each Grade (ME Corrected - Pessimistic)

Percentile	Kindergarten	First grade	Second grade	Third grade	Fourth grade	Fifth grade	Sixth grade
10	0.852	0.936	0.464	0.239	0.345	0.090	0.000
25	0.941	0.972	0.520	0.287	0.410	0.107	0.025
50	1.005	1.006	0.591	0.357	0.522	0.141	0.206
75	1.062	1.045	0.678	0.461	0.712	0.242	1.043
90	1.163	1.078	0.803	0.580	1.053	3.773	2.661

Notes: This table shows estimates of the distribution of classroom inputs in each grade estimated from the production function. The estimates are corrected for measurement error using the pessimistic calibration for γ. The production function is specified in Equation (8).

B Grade Equivalent Scores

We are able to link test scores across grades because there are common items that are administered across several grades, both for the math and language assessments. We use standard linking procedures, where we begin by estimating an unrestricted IRT model (for each subject) for end of kindergarten assessments. Then we estimate an IRT model for end of grade 1 assessment restricting the coefficients on the common items to be the same as in the kindergarten model. We proceed sequentially in a similar way until we reach end of sixth grade assessments, restricting the coefficients on common items to the coefficients on the same items estimated in previous grades. This procedure is also similar to what is used in Attanasio et al. (2020)

We pool together all assessments given in one subject in a given grade. Let D_{ijst} be an indicator that takes value 1 if student i in grade t provided a correct answer to item j in subject s (math, language, or executive function). Let θ_{ist} be the latent ability measure, and J be the total number of items in a given subject and grade (it can change with both subject and grade). The measurement system for subject s in grade t looks like:

$$D_{ijst} = \mathbf{1}\{\, a_{jst} + b_{jst}\,\theta_{ist} + \varepsilon_{ijst} > 0\}, \quad j = 1, \ldots, J.$$

With logit errors we get the standard two-parameter IRT model:

$$\Pr\big(D_{ijst} = 1 \mid \theta_{ist}\big) = \frac{\exp\big(a_{jst} + b_{jst}\,\theta_{ist}\big)}{1 + \exp\big(a_{jst} + b_{jst}\,\theta_{ist}\big)}, \quad j = 1, \ldots, J.$$

We start with end of kindergarten (K) assessments for each subject. We pool all the items (in the same subject) in the same measurement system, and we normalize $a(=0)$ and $b(=1)$ for one of the items, as usual, before we estimate the remaining parameters of the IRT model.

$$\Pr\big(D_{ijsK} = 1 \mid \theta_{isK}\big) = \frac{\exp\big(a_{jsK} + b_{jsK}\,\theta_{isK}\big)}{1 + \exp\big(a_{jsK} + b_{jsK}\,\theta_{isK}\big)}, \quad j = 1, \ldots, J.$$

From this procedure we obtain estimates $(\hat{a}_{1sK}, \ldots, \hat{a}_{jsK}, \hat{b}_{1sK}, \ldots, \hat{b}_{jsK})$. Then we can potentially construct the best estimate of θ_{isK} for each individual in K.

We then go to end of first grade. The IRT system is:

$$\Pr\big(D_{i1s1} = 1 \mid \theta_{is1}\big) = \frac{\exp\big(a_{1s1} + b_{1s1}\,\theta_{is1}\big)}{1 + \exp\big(a_{1s1} + b_{1s1}\,\theta_{is1}\big)}$$

$$\ldots$$

$$\Pr\big(D_{iJs1} = 1 \mid \theta_{is1}\big) = \frac{\exp\big(a_{Js1} + b_{Js1}\,\theta_{is1}\big)}{1 + \exp\big(a_{Js1} + b_{Js1}\,\theta_{is1}\big)}$$

J can of course vary by grade. For simplicity we ignore variation in J across grades. We identify the items that are common in K and 1, and when estimating the IRT system for grade 1, we restrict the a_{js1} and b_{js1} parameters to be the same as those estimated for K. Therefore, for a common item j_c, the restriction is $a_{j_c s1} = \hat{a}_{j_c sK}$ and $b_{j_c s1} = \hat{b}_{j_c sK}$.

With this procedure, we also construct our best prediction of θ_{is1} (empirical bayes mean) for each student in grade 1. We also end up with $(\hat{a}_{1s1}, \ldots, \hat{a}_{Js1}, \hat{b}_{1s1}, \ldots, \hat{b}_{Js1})$, although not all of these are estimated, since we constrain some of them to the K values. One implicit assumption in our procedure is that the performance of an individual on a common item in K and 1 depends

only on the value of θ at that age, and not on what other items/assessments are given at the same time. If performance on an item depends also on which other items or assessments are given (because, for example, the individual gets tired if assessments are very large or very hard, or gets better at answering an item if there are many other similar items in the assessment, or for some other reason), then our procedure is not valid.

Going on to second grade, the IRT system is the same:

$$\Pr\big(D_{i1s2}=1 \mid \theta_{is2}\big)=\frac{\exp\big(a_{1s2}+b_{1s2}\,\theta_{is2}\big)}{1+\exp\big(a_{1s2}+b_{1s2}\,\theta_{is2}\big)}$$
$$\dots$$
$$\Pr\big(D_{iJs2}=1 \mid \theta_{is2}\big)=\frac{\exp\big(a_{Js2}+b_{Js2}\,\theta_{is2}\big)}{1+\exp\big(a_{Js2}+b_{Js2}\,\theta_{is2}\big)}$$

We identify the common items administered in grade 1 and 2 assessments, get the estimates for these items from the grade 1 system (some of them may even be common to K), and we constraint the corresponding grade 2 parameters to be the same as the estimated grade 1 parameters in this set of items.

We repeat this procedure until grade 6. We obtain estimates of $(\theta_{isK},\dots,\theta_{is6})$ which can be arbitrarily correlated (within student, across grades), since they are estimated from completely separate systems (if they were estimated jointly, as in Attanasio et al. (2020), we would need to worry about having a flexible specification for their joint distribution, as they point out in their paper, since something like a normal would restrict substitutability parameters in the production function).

From the first step we obtain estimates of $(\theta_{isK},\dots,\theta_{is6})$ for each individual. These estimates have a common location and scale across grades, so they can be used, for example, to look at growth curves. The second step of our procedure is to convert $(\theta_{isK},\dots,\theta_{is6})$ into grade equivalent scores, which we denote by $(\varphi_{isK},\dots,\varphi_{is6})$.

A standard way to estimate grade equivalents is to try to fit median scores in each grade. We can do this within our sample. We start by computing

$$M_K=\text{median}(\theta_{isK})$$
$$\dots$$
$$M_6=\text{median}(\theta_{is6}).$$

This gives us 7 points in the grade equivalence function. Let median end of K scores correspond to 1 grade of learning, median end of grade 1 scores correspond to 2 grades of learning, and so on. Then the 7 points in the grade equivalence function we have are: $(1, M_K)$, $(2, M_1)$, $(3, M_2)$, $(4, M_3)$, $(5, M_4)$, $(6, M_5)$, $(7, M_6)$. In other words, if individual i in grade t has a score of $\theta_{ist} = M_4$, then we say this individual has the equivalent of 4 grades of learning.

However, so far we only have 7 points in the function, while $(\theta_{isK}, \ldots, \theta_{is6})$ are continuous variables. We need to fill in the remaining points in the function by fitting a function $g_s(\cdot)$ to this data:

$$\varphi_{ist} = g_s(\theta_{ist}).$$

The function $g_s(\cdot)$, which needs to be estimated, converts scores θ_{ist} into grade equivalents φ_{ist}. It turns out that an exponential function provides a good fit for $g_s(\cdot)$ (using the 7 data points for the medians).

Actually, there are 7 points for math, but 8 points for language. The baseline TVIP gives us an additional point at baseline for language. However, for simplicity, in the discussion below we keep mentioning 7 points throughout.

We need to address one additional issue. Since we are only fitting 7 points, $(1, M_K)$, $(2, M_1)$, $(3, M_2)$, $(4, M_3)$, $(5, M_4)$, $(6, M_5)$, $(7, M_6)$, although we can be more or less confident about our grade equivalent scores within the support of this data, we are likely to be less confident outside of it. In particular, grade equivalent scores for the very low end of kindergarten scores, or the very high end of 6th grade scores, depend on how reliable our estimated function is outside the range of the data. This is a problem of finding reasonable extrapolation outside the range of the data. We experiment with a few alternatives.

One additional practical issue we encountered is that both the factor model coefficients and the predicted factor scores (mean of the posterior distribution of the factor for each individual, given their response patterns) depend on how many items there are in each test. So, if we have, for example, 90% of items coming from one test and 10% of items coming from another test, the second test is not going to weigh too much in the determination of the factor and of the coefficients.

The tests we give have an unbalanced number of items. They do not correspond to the relative importance of each test. For example, there are some years where we have 70 items for the TVIP, 7 or 8 times more than we have for all the other tests. Therefore, we need to rebalance the data by reweighting the items in each test depending on how many items were given overall.

Table B.1 shows percentiles of the distribution of grade equivalents resulting from this procedure:

Table B.1: Distribution of Grade Equivalent Scores at the End of Each Grade

	Kindergarten	First grade	Second grade	Third grade	Fourth grade	Fifth grade	Sixth grade
σ	0.347	0.698	0.851	1.100	1.303	1.619	1.948
p10	0.735	1.269	2.047	2.710	3.426	4.214	4.819
p25	0.958	1.484	2.593	3.343	4.100	5.064	5.834
p50	1.088	1.839	3.139	4.024	4.897	6.086	7.037
p75	1.217	2.263	3.701	4.739	5.787	7.218	8.420
p90	1.400	2.738	4.242	5.465	6.694	8.326	9.745

Notes: This table shows the standard deviation and the 10th, 25th, 50th, 75th, and 90th percentiles of grade equivalent score distributions at the end of each grade.

C Test of Random Assignment

An important assumption underlying our empirical strategy is that children's classroom rank at the beginning of a given grade is random, due to random assignment of children to classrooms within schools in every year.[20] Random assignment is closely monitored, and compliance is very high, 98.9 percent on average. In this appendix, we present tests of random assignment using a methodology developed in Jochmans (2023).

First, we briefly discuss the procedure outlined in Jochmans (2023). Consider our setting, in which we observe data on S schools, and each school has $n_1, \ldots, n_S$ students. Within each school, children are assigned to a classroom—and therefore their peer group—every year. Let $x_{s,i}$ be an observable characteristic of child i in school s. If assignment to peer groups is random, $x_{s,i}$ will be uncorrelated with $x_{s,j}$ for all j belonging to the set of i's classroom peers. Let $\bar{x}_{s,j}$ be the average value of characteristic x among student i's peers. The procedure tests whether the correlation in a within-school regression of $x_{s,i}$ on $\bar{x}_{s,i}$ is statistically significantly different from zero (a methodology first proposed in Sacerdote (2001)), introducing a bias correction for the inclusion of group fixed effects (in our case, schools). It is important to control for school fixed effects, as randomization happens within schools, but there may be selection into a school based on individual characteristics. Jochmans (2023) shows that a fixed-effects regression of $x_{s,i}$ on $\bar{x}_{s,i}$ will yield biased estimates due to inconsistency of the within-group estimator. The proposed corrected estimator is given by

[20] We use the word "random" as shorthand but, as discussed at length in Araujo et al. (2016) and Campos et al. (2025), strictly speaking random assignment only occurred in 3rd through 6th grade. In the other grades, the assignment rules were as-good-as-random. Specifically, the assignment rules we implemented were as follows: In kindergarten, all children in each school were ordered by their last name and first name, and were then assigned to teachers in alternating order; in 1st grade, they were ordered by their date of birth, from oldest to youngest, and were then assigned to teachers in alternating order; in 2nd grade, they were divided by gender, ordered by their first name and last name, and then assigned in alternating order; in 3rd through 6th grades, they were divided by gender and then randomly assigned to one or another classroom.

$$\hat{\rho} = \frac{\sum_{s=1}^{S} \sum_{i=1}^{n_s} \tilde{x}_{s,i} \left(\bar{x}_{s,j} - \frac{x_{s,i}}{n_s - 1} \right)}{\sqrt{\sum_{s=1}^{S} \left(\sum_{i=1}^{n_s} \tilde{x}_{s,i} \left(\bar{x}_{s,j} - \frac{x_{s,i}}{n_s - 1} \right) \right)^2}}$$

where $\tilde{x}_{s,i}$ is the deviation of $x_{s,i}$ from its within-school mean. The null hypothesis is thus absence of correlation between i's characteristics and those of her peers. To test the random assignment in our setting, we implement this procedure by testing for the presence of correlation between child i's scores measured at the end of grade $t-1$ and the average end-of-grade scores in $t-1$ of the classroom peers assigned to her in a given grade t. We do so for each grade. We implement the test for all children in the sample, and restricting the sample to those children who have both end of grade $t-1$ scores as well as end of grade t scores (as these will be the children that end up being included in the estimation of our models). The results are shown in Tables C.1 and C.2, respectively. Our results show that we cannot reject the null hypothesis that there is no correlation between child i's achievement and that of her classroom peers. This result is true for all grades and both samples. Hence, we conclude that random assignment was successful in our setting.

Table C.1: Testing for random assignment of children to classrooms, full sample

	Kindergarten	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6
Test statistic	1.359	-0.383	0.905	0.300	-0.445	-0.222	0.980
P-value	0.174	0.702	0.366	0.764	0.657	0.825	0.327

Notes: In this table, we report results for tests of random assignment of children to classrooms within schools using a methodology proposed by Jochmans (2023). The null hypothesis is absence of correlation between a child's ability measured at the end of the previous grade and the average ability of classroom peers assigned to her at the beginning of a given grade, conditional on school. The sample includes all children.

Table C.2: Testing for random assignment of children to classrooms, restricted sample

	Kindergarten	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6
Test statistic	1.392	-0.005	1.425	0.413	-0.043	0.001	1.037
P-value	0.164	0.996	0.154	0.680	0.966	0.999	0.300

Notes: In this table, we report results for tests of random assignment of children to classrooms within schools using a methodology proposed by Jochmans (2023). The null hypothesis is absence of correlation between a child's ability measured at the end of the previous grade and the average ability of classroom peers assigned to her at the beginning of a given grade, conditional on school. The sample is restricted to children who have available both beginning- and end-of-grade scores for a given grade.

D Procedure for Estimating the Production Function

Basic Model

Equation (7) defines a system of equations, one for each grade $t = 0 \dots 6$. In order to estimate it, we start by taking logs:

$$\ln Y_{c_0 \dots c_t s t j} = \mu_{st} + X_{c_0 \dots c_t s t j} \gamma_t + \frac{\theta_t}{\rho_t} \ln \left(\sum_{k=0}^{t} \pi_{c_k s t} \delta_{c_k s k}^{\rho_t} \right) + v_{c_0 \dots c_t s j} \tag{8}$$

We define $v_{c_0 \dots c_t s j} = \ln(u_{c_0 \dots c_t s j})$. In addition, we need to initialize the system. Notice that the implied equation for grade 0 (kindergarten) only has one classroom input, and therefore it simplifies to:

$$\ln Y_{c_0 s 0 j} = \mu_{s0} + X_{c_0 s 0 j} \gamma_0 + \theta_0 \ln \left(\pi_{c_0 s 0}^{\frac{1}{\rho_0}} \right) + \theta_0 \ln(\delta_{c_0 s 0}) + v_{c_0 s 0 j} \tag{9}$$

This is a standard VA equation for kindergarten, where $\ln Y_{c_0 s 0 j}$ is a linear function of classroom assignment indicators, which are estimated to be $\theta_0 \ln(\delta_{c_0 s 0})$. θ_0 is normalized to be equal to 1. This normalization does not affect our estimates of the elasticity of substitution across inputs in different grades since it affects classroom inputs in kindergarten proportionally. The return to scale parameters in the remaining grades can then be freely estimated.

Identification

As mentioned in the paper, the assumption that classroom inputs are common to all students in a particular classroom means that the parameters of the system of Equations (8) and (9) (one per grade) and the vector of classroom qualities are identified, and should be estimated simultaneously. In other words, each student in a given classroom faces the same classroom input, which affects learning in that grade and all subsequent grades (in future grades, the same classroom input could have different productivity).

As an illustrative example, suppose we have data from a single school with three classrooms in each grade: A, B and C. Assume also there are no other X controls we need to consider. We already saw that for grade 0 (kindergarten) we need one normalization which we will discuss below. For now, assume we have an estimate of $\delta_{c_0 s 0}$ for each classroom, $c = A, B, C$. Start from the grade 1 production function. Define $Y_{c_0 c_1 s 1} = E(\ln Y_{c_0 c_1 s 1 j} | c_0, c_1) = \frac{\theta_1}{\rho_1} \ln[\pi_{1s1} \delta_{As0}^{\rho_1} + (1 - \pi_{1s1}) \delta_{As1}^{\rho_1}]$. Then:

$$
\begin{aligned}
Y_{AAs1} &= E(\ln Y_{c_0c_1sj}|c_0 = A, c_1 = A) = \frac{\theta_1}{\rho_1}\ln[\pi_{1s1}\delta_{As0}^{\rho_1} + (1-\pi_{1s1})\delta_{As1}^{\rho_1}] \\
Y_{ABs1} &= E(\ln Y_{c_0c_1sj}|c_0 = A, c_1 = B) = \frac{\theta_1}{\rho_1}\ln[\pi_{1s1}\delta_{As0}^{\rho_1} + (1-\pi_{1s1})\delta_{Bs1}^{\rho_1}] \\
Y_{ACs1} &= E(\ln Y_{c_0c_1sj}|c_0 = A, c_1 = C) = \frac{\theta_1}{\rho_1}\ln[\pi_{1s1}\delta_{As0}^{\rho_1} + (1-\pi_{1s1})\delta_{Cs1}^{\rho_1}] \\
&\vdots \\
Y_{BAs1} &= E(\ln Y_{c_0c_1sj}|c_0 = B, c_1 = A) = \frac{\theta_1}{\rho_1}\ln[\pi_{1s1}\delta_{Bs0}^{\rho_1} + (1-\pi_{1s1})\delta_{As1}^{\rho_1}] \\
&\vdots \\
Y_{CCs1} &= E(\ln Y_{c_0c_1sj}|c_0 = C, c_1 = C) = \frac{\theta_1}{\rho_1}\ln[\pi_{1s1}\delta_{Cs0}^{\rho_1} + (1-\pi_{1s1})\delta_{Cs1}^{\rho_1}]
\end{aligned}
$$

Taking ratios, since there are 9 kindergarten and first grade combinations, there are 8 unique ratios that are not linearly dependent:

$$
\begin{aligned}
\frac{Y_{AAs1}}{Y_{ABs1}} &= \frac{\ln[\pi_{1s1}\delta_{As0}^{\rho_1} + (1-\pi_{1s1})\delta_{As1}^{\rho_1}]}{\ln[\pi_{1s1}\delta_{As0}^{\rho_1} + (1-\pi_{1s1})\delta_{Bs1}^{\rho_1}]} \\
\frac{Y_{AAs1}}{Y_{ACs1}} &= \frac{\ln[\pi_{1s1}\delta_{As0}^{\rho_1} + (1-\pi_{1s1})\delta_{As1}^{\rho_1}]}{\ln[\pi_{1s1}\delta_{As0}^{\rho_1} + (1-\pi_{1s1})\delta_{Cs1}^{\rho_1}]} \\
\frac{Y_{AAs1}}{Y_{BAs1}} &= \frac{\ln[\pi_{1s1}\delta_{As0}^{\rho_1} + (1-\pi_{1s1})\delta_{As1}^{\rho_1}]}{\ln[\pi_{1s1}\delta_{Bs0}^{\rho_1} + (1-\pi_{1s1})\delta_{As1}^{\rho_1}]} \\
\frac{Y_{AAs1}}{Y_{BBs1}} &= \frac{\ln[\pi_{1s1}\delta_{As0}^{\rho_1} + (1-\pi_{1s1})\delta_{As1}^{\rho_1}]}{\ln[\pi_{1s1}\delta_{Bs0}^{\rho_1} + (1-\pi_{1s1})\delta_{Bs1}^{\rho_1}]} \\
\frac{Y_{AAs1}}{Y_{BCs1}} &= \frac{\ln[\pi_{1s1}\delta_{As0}^{\rho_1} + (1-\pi_{1s1})\delta_{As1}^{\rho_1}]}{\ln[\pi_{1s1}\delta_{Bs0}^{\rho_1} + (1-\pi_{1s1})\delta_{Cs1}^{\rho_1}]} \\
\frac{Y_{AAs1}}{Y_{CAs1}} &= \frac{\ln[\pi_{1s1}\delta_{As0}^{\rho_1} + (1-\pi_{1s1})\delta_{As1}^{\rho_1}]}{\ln[\pi_{1s1}\delta_{Cs0}^{\rho_1} + (1-\pi_{1s1})\delta_{As1}^{\rho_1}]} \\
\frac{Y_{AAs1}}{Y_{CBs1}} &= \frac{\ln[\pi_{1s1}\delta_{As0}^{\rho_1} + (1-\pi_{1s1})\delta_{As1}^{\rho_1}]}{\ln[\pi_{1s1}\delta_{Cs0}^{\rho_1} + (1-\pi_{1s1})\delta_{Bs1}^{\rho_1}]} \\
\frac{Y_{AAs1}}{Y_{CCs1}} &= \frac{\ln[\pi_{1s1}\delta_{As0}^{\rho_1} + (1-\pi_{1s1})\delta_{As1}^{\rho_1}]}{\ln[\pi_{1s1}\delta_{Cs0}^{\rho_1} + (1-\pi_{1s1})\delta_{Cs1}^{\rho_1}]}
\end{aligned}
$$

From here we have enough moments to recover $\pi_{1s1}, \rho_1, \delta_{As1}, \delta_{Bs1}, \delta_{Cs1}$. Even though $\delta_{As1}, \delta_{Bs1}, \delta_{Cs1}$ are classroom fixed effects, embedded in a nonlinear model, they can be estimated from a large number of students per classroom. Finally, the levels' equations allow us to recover θ_1.

In the case of schools with only two classrooms per grade (which is true of most but not all schools in our sample), we only have 3 linearly independent ratios per school. Therefore we cannot identify the model from a single school, but we have enough moments to recover all the parameters if we use at least two different schools (since we need to estimate π_{1s1}, ρ_1 plus two classroom effects per school, a total of six parameters which can be recovered from six ratios across two schools; in addition to θ_1, which can then be recovered from level equations). Since we have more than two schools, we can estimate all the parameters of the model, even if schools only had two classrooms per grade.

To estimate the model at the end of grade 2 we have one more parameter to recover (π_{2s1}). This means that, if we have a single school with three classrooms, we need at least 6 linearly independent ratios like the ones above to recover all the parameters. If schools only have two classrooms, we need to have data from three schools. Each additional grade adds one more parameter to the model. Regardless, we have enough schools and classrooms to identify the entire model, even at the end of sixth grade.

Notice that we need a normalization to recover the kindergarten classroom input: θ_0 is normalized to be equal to 1. This is a fairly innocuous normalization. Nevertheless, our main results are presented in the form of counterfactual simulations of different sequences of classroom inputs, which are not influenced by this normalization.

Estimation

Rather than estimating the entire model for all grades simultaneously, we use a computationally tractable iterative approach that proceeds one grade at a time, beginning with the lowest grades. We start from Equation (9), $t = 0$, from which we recover estimates of $\delta_{c_0 s0}$ for each classroom, along with the other model parameters, which are not of substantial interest. Moving to first grade, we use Equation (8) for $t = 1$, incorporating the previously estimated $\delta_{c_0 s0}$ values as known quantities. We then estimate all production function parameters $(\theta_1, \rho_1, \pi_{c_0 s1})$ jointly with the new classroom effects $\delta_{c_1 s1}$, as well as the parameters on the control variables. We continue with this approach for each subsequent grade t: we use the full set of previously estimated classroom effects $\{\delta_{c_0 s0} \ldots \delta_{c_{t-1} s t-1}\}$ as inputs, and estimate the current grade's parameters $(\theta_t, \rho_t, \pi_{c_0 s t}, \ldots, \pi_{c_{t-1} s t})$ together with $\delta_{c_t s t}$.

For each grade t, we employ a four-step estimation procedure designed to handle the computational complexity arising from the large number of classroom indicators in the nonlinear CES production function specification:

1. **Residualized outcome:** In the first step we estimate γ_t using grade-specific regressions of log test scores on classroom fixed effects and standard controls: $\ln Y_{c_0 \ldots c_t s t j} =$

$\vartheta_{cst} + X_{c_0 \ldots c_t stj}\gamma_t + v_{c_0 \ldots c_t stj}$.[21] While we could alternatively use indicators for the complete sequence of classroom assignments instead of ϑ_{cst}, this substantially increases the parameter count without meaningfully changing our results. From this regression, we obtain γ_t estimates and construct the residualized outcome variable $\ln \tilde{Y}_{c_0 \ldots c_t stj} = \ln Y_{c_0 \ldots c_t stj} - X_{c_0 \ldots c_t stj}\gamma_t$. We then use this residualized outcome to estimate the production function:

$$\ln \tilde{Y}_{c_0 \ldots c_t stj} = \mu_{st} + \frac{\theta_t}{\rho_t} \ln \left(\sum_{k=0}^{t} \pi_{c_k st} \delta_{c_k sk}^{\rho_t} \right) + v_{c_0 \ldots c_t stj} \tag{10}$$

2. **Initial classroom effect estimation:** We begin by generating initial values for the current grade's classroom effects $\delta_{c_t st}$, treating the previously estimated classroom qualities $\{\delta_{c_0 s0} \ldots \delta_{c_{t-1} st-1}\}$ as data in this step. For initial values, we use estimates of school-demeaned classroom effects from linear value-added models for grade t. We then fix the CES parameters and school fixed effects at arbitrary initial values (typically zeros) and estimate new values of $\delta_{c_t st}$ using GMM.[22]

3. **CES parameter estimation and convergence:** Using the $\delta_{c_t st}$ values from step two, we estimate new CES parameters and school fixed effects via GMM.[23] We use these updated CES estimates and the current $\delta_{c_t st}$ values as starting points to compute a new round of classroom effects. We iterate between steps 2 and 3 until either the change in $\delta_{c_t st}$ values or the weighted change in CES parameters falls below our convergence threshold.

4. **Global optimization:** To ensure we find the global optimum rather than a local minimum, we restart the algorithm from step 2 using 500 different random starting values for the CES production function parameters $(\theta_t, \rho_t, \pi_{c_0 st}, \ldots, \pi_{c_{t-1} st})$. For each set of starting values, we fix these parameters and estimate corresponding $\delta_{c_t st}$ values that are consistent with those parameters. We continue between steps 2 and 3 until convergence. Among all 500 estimation runs, we select the parameter estimates that minimize the standard GMM objective function.

Table A.1 reports the estimated production function parameters, while Table A.2 presents the percentiles of the estimated distribution of classroom inputs for each grade.

[21] We control for age and gender, as well as the following baseline controls: a quartic polynomial in TVIP, maternal education, preschool attendance, and wealth. Since classroom assignment is orthogonal to these controls, controlling for classroom fixed effects instead of school fixed effects is equivalent.

[22] In this step, we use the following moment function: $g(x; \theta, \rho, \pi) = \hat{v}\, h(x)$, where $h(x)$ is the set of classroom dummies and $\hat{v}$ is the residual from Equation (10).

[23] In this step, we use the following moment function: $g(x; \theta, \rho, \pi) = \hat{v}\, h(x)$, where $h(x)$ includes the set of school dummies, $\{\delta_{c_0 s0} \ldots \delta_{c_t st}\}$, and an interaction term of all the previously and currently estimated classroom qualities, and $\hat{v}$ is the residual from Equation (10).

E Measurement Error Correction

Evdokimov and Zeleneev (2025) develop a measurement error correction for GMM applications where the goal is to estimate the moment condition $\mathbb{E}[g(X_i^*, S_i, \theta)]$, but X_i^* is unobserved and the econometrician only observes the mismeasured variable $X_i = X_i^* + \epsilon_i$, where ϵ_i is the measurement error. They propose substituting the original moment with a modified moment condition $\mathbb{E}[\psi(X_i, Si, \theta, \gamma)]$, where $\psi(X_i, Si, \theta, \gamma)$ is given by:

$$\psi(X_i, Si, \theta, \gamma) = g(X_i, S_i, \theta) - \gamma_2 g_x^{(2)}(X_i, S_i, \theta)\,,$$

where $g_x^{(2)}$ is the second derivative of function g with respect to X and $\gamma_2 = \mathbb{E}[\epsilon_i^2]/2 = V(\epsilon_i)/2$ (where $V(.)$ indicates variance), since $\mathbb{E}[\epsilon_i]$ is assumed to be zero.

The authors then proceed by presenting an instrumental variables method to estimate the model. In our setting we do not have an instrumental variable, but at each stage of the sequential (grade by grade) procedure described above, we can nevertheless obtain an estimate of γ (from the estimates of classroom quality obtained in previous grades). This is because we can obtain estimates of the quality of each classroom with corresponding standard errors (as in standard VA models). So although we will not use exactly the estimation method proposed in Evdokimov and Zeleneev (2025), we will rely very heavily on their idea and on their proposed adjustments to the moment conditions of our model.

For example, the mismeasured classroom quality variable in $t = 1$ (1st grade) is $X_i = \delta_{cs0}$, classroom quality in kindergarten, and we can estimate γ_2 using our VA equation for kindergarten:

$$\ln Y_{sc_0 0j} = \mu_{s0} + X_{sc_0 0j}\gamma_0 + \theta_0 \ln(\pi_{c_0 s0}^{\frac{1}{\rho}}) + \theta_0 \ln(\delta_{cs0}) + \upsilon_{sc_0 0j}\,,$$

where $X_{sc_0 0j}$ is a vector of controls that include child age, child age squared, gender, a household wealth index, mother's education, a dummy for preschool attendance before kindergarten, and a fourth order polynomial in $\ln Y_{sc_0-1j}$ (baseline TVIP in grade equivalents and logs).

We then can estimate $\ln Y_{sc_0 0j}$ as a function of classroom assignment indicators, which are estimated to be $\theta_0 \ln(\delta_{cs0})$. θ_0 is normalized to be equal to 1.

We define the demeaned classroom effect as $\gamma_{cs0} = \ln(\delta_{cs0}) - \ln(\delta_{s0}) = \ln(\delta_{cs0}) - \frac{\sum_{d=1}^{C_s} N_{ds} \ln(\delta_{ds0})}{\sum_{d=1}^{C_s} N_{ds}}$. The variance of $\hat{\gamma_{cs0}}$ conditional on γ_{cs0}, or the variance of measurement error in kindergarten classroom quality, can be written as[24]:

[24] See the full derivation in Araujo et al. (2016).

$$V(\hat{\gamma_{cs0}}|\gamma_{cs0}) = E\left[\frac{\left(\sum_{d=1}^{C_s} N_{ds}\right) - N_{cs}}{N_{cs}\sum_{d=1}^{C_s} N_{ds}}\sigma^2\right]$$

Since $\hat{\gamma_{cs0}}$ is in logs, but our mismeasured variable X_i is in levels, we use the delta method and rewrite the expression above as:

$$V(\exp(\hat{\gamma_{cs0}})|\gamma_{cs0}) = (\exp(\hat{\gamma_{cs0}}))^2\,\mathbb{E}\left[\frac{\left(\sum_{d=1}^{C_s} N_{ds}\right) - N_{cs}}{N_{cs}\sum_{d=1}^{C_s} N_{ds}}\sigma^2\right]$$

To generate a unique value for our γ_2 parameter, we take the average across classrooms:

$$\gamma_2 = E\left(V(\exp(\hat{\gamma_{cs0}})|\gamma_{cs0})\right) = E\left((\exp(\hat{\gamma_{cs0}}))^2 E\left[\frac{\left(\sum_{d=1}^{C_s} N_{ds}\right) - N_{cs}}{N_{cs}\sum_{d=1}^{C_s} N_{ds}}\sigma^2\right]\right) = 0.0155$$

This estimate corresponds to the average standard error of the estimate of γ_{cs0}, which can be obtained directly from the estimation of the production function for Kindergarten. Finally, we use the moment adjustment proposed in Evdokimov and Zeleneev (2025) in step 3 of the procedure described in section D in the Appendix to estimate the production function for $t = 1$.

We can now move to grade $t = 2$ (second grade). We estimate equation 10 by GMM with the GMM measurement error adjustment, where the grade 0 and grade 1 inputs are measured with error, and the measurement error variance was obtained from the estimation of the production functions for those grades. And we can proceed analogously for the subsequent grades, slightly modifying step 3 of the procedure described in section D in the Appendix, as described in the previous paragraph.

Although this is not required, in practice it is much less cumbersome to assume that the measurement error adjustment is similar across grades. There is a range of values one could use based on the estimated measurement error in each grade. Therefore, we decided to examine the sensitivity our main conclusions to variation in the measurement error variance, within ranges of values consistent with the data.

In particular, we consider an optimistic scenario, with relatively low measurement error ($\gamma_2 = 0.0023$), and a pessimistic case, with relatively high measurement error ($\gamma_2 = 0.0067$). The estimated parameters of the production function for the optimistic case are in Table A.3, the distribution of estimated classroom quality is in table A.4, and the estimated production

functions for each grade are represented in Figures A.5 and A.6. The estimated parameters of the production function for the pessimistic case are in Table A.5, the distribution of estimated classroom quality is in table A.6, and the estimated production functions for each grade are represented in Figures A.7 and A.8.

There are of course some changes in the production function estimates, as one would expect. However, the main result, that there is no evidence of strong dynamic complementarity in classroom inputs, is robust to the measurement error correction implemented here.