Dynamic Complementarity in Elementary Schools: Experimental Estimates from Ecuador

Pedro Carneiro[1]
University College London, Centre for Microdata Methods and Practice, Institute for Fiscal Studies, FAIR-NHH

Yyannu Cruz-Aguayo
Inter-American Development Bank

Rafael Hernandez-Pachon
University of Zurich

Norbert Schady
World Bank

Abstract

This paper examines experimentally the impact of classroom inputs in the production function of 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.

[1] Carneiro gratefully acknowledges the support of the ESRC for CEMMAP (ES/P008909/1) 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.

1. Introduction

The process of human capital accumulation in childhood is complex, involving many inputs provided by different actors, including parents, teachers, and peers. Some inputs may have larger effects at some ages than at others, and there may be important interactions between them. Establishing causal effects is difficult, however, because inputs are likely to depend on characteristics of the child which are unobservable to the researcher.

In this paper, we analyze the dynamic impacts of one important input—classroom quality—on learning outcomes in elementary school. Specifically, we focus on the extent to which being randomly exposed to a higher-quality classroom in one grade increases the returns to being in a higher-quality classroom in a subsequent grade. Cunha and Heckman (2007) refer to this process as "dynamic complementarities" in the production of human capital.

There is a small literature that tries to uncover evidence of dynamic complementarities. The fundamental challenge in this research is finding more than one exogenous source of variation that affects human capital accumulation for the *same* individuals at two points in time. One recent paper argues that this may be akin to "asking for lightning to strike twice" (Almond and Mazumder 2013).

Earlier research has taken one of two approaches to address this identification challenge. Some papers have relied on quasi-experimental variation in policies that affected human capital accumulation at two points in the life cycle. This approach is taken, for example, 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.

A different strand of the literature uses 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 (2016) using panel data from the U.S., and Attanasio et al. (2020) using panel data from India.

The results of these studies have been mixed.[2] In this paper, we add to this literature by using data from a unique experiment we 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 (so "lighting" strikes not only twice, but up to 7 times).[3] Compliance

[2] Attanasio et al. (2020), Cunha and Heckman (2007), and Johnson and Jackson (2017) find evidence of dynamic complementarities, whereas other authors such as Kinsler (2016) and Malamud et al. (2016) do not.

[3] We refer to our assignment as "random" as shorthand, although technically random assignment occurred after 3rd grade. In the earlier grades, the assignment rules were as-good-as-random. Specifically, the assignment rules we

with random assignment was essentially perfect—98.9 percent on average. With this data, we are able to combine experimental variation in school inputs at different ages of children, with structural models of skill accumulation to provide new estimates of the substitutability of school inputs at different ages.

Children were tested in math and language at the end of each grade. From these assessments we construct a measure of grade-specific achievement for each child, scaled in terms of grade equivalents. The test score data also allow us to construct classroom value added, our measure of classroom quality, as in our earlier work (Araujo et al. 2016), as well as in research using U.S. data (Chetty et al. 2011; 2014).

The fact that each child in our data was exposed to seven exogenous, orthogonal shocks to classroom quality is unique. It allows us to test for dynamic complementarities with minimal assumptions and using different approaches. 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. 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, suggesting that dynamic interactions are unlikely to be important in our sample. If there were dynamic complementarities in classroom quality, we would expect achievement to be a convex function of the number of high-quality classrooms.

Next, we formally test for interactions between classroom quality in different grades, in a more realistic setup where classroom quality is continuous. Specifically, we implement a procedure suggested by Kinsler (2016), who uses non-experimental data on children in 3rd and 4th grades in North Carolina to test for dynamic interactions in teacher quality. We model a student's achievement at the end of each grade as a flexible function of current and past classroom assignments, 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 that there are no interactions between classroom quality across grades. The second specification saturates the model described above by adding interactions between classroom fixed effects in each grade. This is equivalent

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; 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 A.

to using one indicator for each sequence of classrooms (as opposed to one fixed effect per classroom and grade).[4]

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 that the interactions are jointly equal to zero, or 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.[5]

Finally, we estimate grade-specific constant elasticity of substitution (CES) production functions of learning (as in Cunha et al. 2010), where the input is classroom quality in each (current and past) grade. This puts more structure on the data, but it allows us to estimate elasticities of substitution between classroom quality in different grades, and to simulate the impacts of changing classroom quality in each grade on achievement. Consistent with the evidence from the other two approaches discussed above, we show that classroom quality is highly substitutable over time.

Classroom value added is a natural input to look at since it includes everything that happens in the classroom. However, "classroom quality" also subsumes any responses by parents to variation in the quality of teacher, peers, or other aspects of the classroom environment. Measures of classroom quality that *include* these endogenous responses may be most useful for policy purposes, since policy makers cannot control what endogenous responses will be. In other words, a policy maker may more interested in learning whether two education policy tools are complements or substitutes after parents react to them, than in understanding whether they are complements or substitutes keeping parental reactions fixed.

That said, when discussing dynamic complementarities between two specific inputs (e.g., classroom quality in first and second grade), Cunha and Heckman (2007) and others typically have in mind a setting where other (e.g., parental) inputs are kept fixed. Parental responses could in principle reinforce or offset differences in classroom quality, as in Pop-Eleches and Urquiola (2013) or Malamud et al (2021). In that sense, parents can undo or reinforce dynamic complementarities in school inputs, that would manifest themselves differently if parents did not react. In the words of Malamud et al (2021), "well-identified evidence is necessary to assess dynamic complementarities, but as often, reduced form results do not necessarily reveal the possibly countervailing mechanisms that underlie them".

[4] For example, if there are two classrooms per school and grade, then the model without interactions (first specification) includes $2*(t+1)$ classroom indicators, whereas the model with interactions (second specification) includes 2^{t+1} sequence indicators.

[5] 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. See for example …

Parental reactions are likely to vary with the setting and the public inputs being offered. Fortunately, we have access to detailed data on parental investments, although it is only available for a single year: the end of the first year of elementary school (kindergarten). Using this dataset, Araujo et al. (2016) show that parental investments (in time or goods) do not react when their children are randomly assigned to high vs low quality classrooms (these are precisely estimated zero impacts).[6]

Interestingly, this lack of reaction is not due parents having a misperception of the quality of the classroom attended by their children. Using survey data from parents, Araujo et al (2016) show that parental perceptions of classroom quality are strongly correlated with objective measures of quality for the classrooms attended by their children (and to which they were randomly assigned).

The lack of similar data for future grades prevents us from generalizing these parental responses to other grades. From what we can observe from the detailed data we have for first grade, our best guess is that parental reactions to school inputs may be relatively unimportant in our setting. If that is the case, we can interpret our results as characterizing important features of the production function of learning in schools, keeping home inputs fixed.

In sum, in the setting that we study, we find that the productivity of classroom quality in one grade does not depend on the quality of the classroom experienced by that child in earlier grades. Rather, the production function of education is remarkably additive in classroom quality across different grades.

We conclude with a discussion of why this might be the case in our data, and of the policy implications of our finding. Specifically, we note that, had we found evidence of strong dynamic complementarities in classroom quality, a policy-maker assigning a limited number of high-quality teachers to classrooms would face a tradeoff between efficiency (the largest aggregate effect on learning would occur if the same children received a succession of high-quality teachers) and equity (it would be most equitable to spread high-quality teachers so that each child gets exposed to at least some very effective teachers).[7] The absence of dynamic complementarities essentially eliminates this tradeoff for the policymaker.

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.

[6] Araujo et al (2016) show that most of their measures of parental investments are correlated with end of kindergarten test scores (math and language), even after controlling for pre-kindergarten vocabulary scores.

[7] In principle, this applies to high- or low-quality peers as well. If there are dynamic complementarities, the returns to being assigned to higher-quality peers would be larger for children who had been assigned to higher-quality peers in the past.

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.[8] These children were randomly reassigned to 1st grade classrooms in 2013, to 2nd grade classrooms in 2014, to 3rd grade classrooms in 2015, to 4th grade classrooms in 2016, to 5th grade classrooms in 2017, and to 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);[9] 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 A.

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 A (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

[8] 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.

[9] Performance on this test at early ages has been shown to predict important outcomes in a variety of settings, including in 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.

2016, Freyberger 2020). 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).[10]

Table 1, Panel A, summarizes the characteristics of children and their families. Children were approximately five years old on the first day of kindergarten. Half of them are girls. At the time children enrolled in kindergarten, mothers were on average in their early thirties, and fathers were in their mid-thirties. Education levels are similar for both parents—just under nine years of school (which corresponds to completed middle school). The average child in our sample has a TVIP score that places her more than 1 standard deviation below the reference population that was used to norm the test, indicating that many children begin formal schooling with significant delays.[11]

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 females 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 between 35 and 40 students.

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. (2021), replicated in Appendix B of this paper, we present a test of random assignment developed by Jochmans (2020), analogous to a standard balance test but adapted 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

A. 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 grade (t) and school (s) according to its quality (Q_{sct}). To help visualize the essence of our data it is helpful to

[10] Table A1 shows percentiles of the distribution of grade equivalent scores at the end of each grade.
[11] 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).

consider a simple example where classroom quality is discrete and takes only two values, high or good (G), and low or bad (B): $Q_{sct} = \{G_{sct}, B_{sct}\}$. In practice, classrooms G_{sct} are those in which quality is above the grade- and school-specific mean, while classrooms B_{sct} 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 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 $\tilde{Q}^t = \{Q_{sc0}, \dots, Q_{sct}\}$. Average achievement (*Y*) at the end of grade *t* in each cell (where *j* is student) is:

$$E\left(Y_{sctj} \middle| \tilde{Q}^t\right) = E\left(Y_{sctj} \middle| Q_{sc0}, \dots, Q_{sct}\right)$$

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, and needs to be estimated. 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:

$$Y_{sctj} = X_{sct-1j}\gamma_t + \delta_{cst} + u_{sctj} \quad (1)$$

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

Let $v_{sctj} = Y_{sctj} - X_{sct-1j}\gamma$. v_{sctj} measures the amount of learning (or growth in achievement, since the model controls for Y_{sct-1j}) of student *j* in grade *t*. Then, we can compute VA as:

$$VA_{sct} = \frac{1}{N_{sct}} \sum_{k=1}^{N_{sct}} v_{sctk}$$

where N_{sct} is the number of students in school *s*, classroom *c*, 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:

$$\overline{VA}_{st} = \sum_{c=1}^{C_{st}} \frac{N_{sct}}{N_{st}} VA_{sct} \quad (2)$$

Finally, $\alpha_{sct} = VA_{sct} - \overline{VA}_{st}$ 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.

However, even if this assumption is relaxed (as in this paper), 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 at the end of a grade VA differs across classrooms, 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 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:

$$\alpha_{sct} > 0 \Rightarrow G_{cst} = 1$$

$$\alpha_{sct} \leq 0 \Rightarrow B_{cst} = 1 \quad (3)$$

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 $m_t = \left(G_{cs0j}, \ldots, G_{cstj}\right)$, where *t=K,…,6*.

In order to estimate average learning per cell we run the following regression for each *t*:

$$Y_{sctj} = \sum_{m_t} \kappa^{m_t} GB_{tj}^{m_t} + X_{sc0j}\zeta_t + \vartheta_{st} + w_{sctj} \quad (4)$$

where ϑ_{st} is a school by grade fixed effect; X_{sc0j} 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_{sctj} is the test score (math and language aggregate) at the end of grade *t*; and w_{sctj} 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 easily visualize how achievement depends on the sequence of classroom assignments up to a grade, and to what extent dynamic complementarities are likely to be important.[12]

There is one important caveat to the procedures we implement in this paper. Unfortunately, we do not observe parental inputs in our dataset. Therefore, we conflate the effects of classroom quality on learning with the effects of parental responses to classroom quality. It is possible, for example, that there are strong

[12] One potential issue with this procedure is that data on the same individuals shows up on both sides of the regression, since the *GB* indicators and sequences are constructed using all the observations in each classroom. The fact that the *GB* indicators are all discrete and the sequences are quite complex is likely to attenuate this problem substantially. In fact, we show below and in Appendix D 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: $VA_{sct} = \frac{1}{N_{sct}-1}\sum_{k=1,k\neq i}^{N_{sct}} v_{sctk}$. 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 as "exclusion bias" (e.g., Jochmans, 2021, Caeyers and Fafchamps, 2020). 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 Appendix E, and discuss below, estimates based on leave-one-out VA measures, and how they are very similar to the ones presented in the main body of the paper.

dynamic complementarities between school inputs in different grades, but which are partly or wholly undone by parental reactions to these inputs.

B. Testing for Additive Classroom Effects

The data visualization exercise just described is intuitive and simple to implement, but required 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_{sctj}) as a linear function of classroom indicators (δ_{cst}), student level controls (X_{sctj}), and a residual (u_{sctj}). A typical control variable is the lagged test score (Y_{sct-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) begins by taking the standard model which assumes additivity of classroom effects over time, and from which we get the following equation:

$$Y_{sc_0 \ldots c_t tj} = X_{sc_0 \ldots c_t tj}\gamma_t + \sum_{k=0}^{t} \delta_{c_k st} + u_{sc_0 \ldots c_t tj} \quad (5)$$

This is a simple extension of equation (1), where the indices of the variables in the regression have been changed to include the entire history of classroom assignment up to grade *t*: $c_0 \ldots c_t$. In this model, the impacts of classroom qualities in different grades on learning at the end of grade *t* is additive in classroom quality in each grade, ruling out any complementarities between classroom quality across 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 colinear with the main classroom effects. The model becomes:

$$Y_{sc_0 \ldots c_t tj} = X_{sc_0 \ldots c_t tj}\gamma_t + \sum_{k=0}^{t} \delta_{c_k st} + \varphi_{sc_0 \ldots c_t t} + u_{sc_0 \ldots c_t tj} \quad (6)$$

$\varphi_{sc_0 \ldots c_t t}$ 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 ($\delta_{c_k st}$).

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, $\varphi_{sc_0 \ldots c_t t}$, 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 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).

C. Estimating a CES 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:

$$Y_{sc_0 \ldots c_t tj} = A_{sc_0 \ldots c_t tj}\left(\sum_{k=0}^{t} \pi_{c_k st}\delta_{csk}{}^{\rho_t}\right)^{\frac{\theta_t}{\rho_t}} u_{sc_0 \ldots c_t tj} \quad (7)$$

As before, $Y_{sc_0 \ldots c_t tj}$ 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. 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 $\pi_{c_k st}$ give us the relative productivity of classroom quality in each grade (we use the following normalization: $\sum_{k=0}^{t} \pi_{c_k st} = 1$). $A_{sc_0 \dots c_t tj}$ is a productivity parameter that includes school fixed effects (ϑ_{st}) as well as individual level observables ($X_{sc_0 \dots c_t tj}$, 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 $lnA_{sc_0 \dots c_t tj} = \vartheta_{st} + X_{sc_0 \dots c_t tj}\gamma_t$. $u_{sc_0 \dots c_t tj}$ is an individual level i.i.d. shock.

As discussed above, classroom quality, $\delta_{c_k sk}$, is unobserved. However, unlike most papers in the teacher VA literature, we assume that learning is a potentially non-linear function of classroom quality in different grades. This means that we cannot rely on additive linear 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. The reason this is feasible here is because, as stated above, we assume that all students in a classroom receive the same level of input (see also the identification discussion in Appendix C). It is this important (but standard) assumption that allows us to recover simultaneously the classroom input (which shows up as a classroom fixed effect in the model) and the parameters of the production function. Although we assume the classroom input (or fixed effect) to be the same for all students in the classroom, we depart from the literature by allowing the impact of classroom input in one grade to vary across students in the same classroom, depending on the history of classroom inputs they have experienced up to that grade.

The production function in (7) is also different in one important aspect from the specifications in other recent papers such as, for example, Cunha et al (2010), Agostinelli and Wiswall (2016), or Attanasio et al (2020a). In those papers the production function has a first order Markov structure, 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 $f_t(.)$ is the period *t* production function. In other words, all interactions between inputs in period *t* (δ_{tj}) and inputs in prior periods ($\delta_{0j} \dots \delta_{t-1j}$) operate through lagged skills (Y_{t-1j}). Our paper relaxes this assumption, which is found to be too restrictive in Attanasio et al (2020b).

Equation

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

$$lnY_{sc_0 \dots c_t tj} = \mu_{st} + X_{sc_0 \dots c_t tj}\gamma_t + \frac{\theta_t}{\rho_t} ln\left(\sum_{k=0}^{t} \pi_{c_k st} \delta_{csk}{}^{\rho_t}\right) + v_{sc_0 \dots c_t tj} \quad (8)$$

We define $v_{sc_0 \ldots c_t tj} = ln\left(u_{sc_0 \ldots c_t tj}\right)$. 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:

$$lnY_{sc_0 0j} = \mu_{s0} + X_{sc_0 0j}\gamma_0 + \theta_0 \ln\left(\pi_{c_0 s0}^{\frac{1}{\rho_0}}\right) + \theta_0 ln(\delta_{cs0}) + v_{sc_0 0j} \quad (9)$$

This is a standard VA equation for kindergarten, where $lnY_{sc_0 0j}$ is a linear function of classroom assignment indicators, which are estimated to be $\theta_0 ln(\delta_{cs0})$. 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 (8) and (9) (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 C. We start from equation (9), $t = 0$, from which we recover estimates of δ_{cs0} for each classroom (and we 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 δ_{cs0} from the $t = 0$ equation, and we estimate all the parameters of the production function (θ_1, ρ_1, $\pi_{c_0 s1}$) together with δ_{cs1} (as well as the parameters on the controls). In grade *t*, we use $\{\delta_{cs0} \ldots \delta_{cst-1}\}$ obtained from the previous grade equations, and we estimate (θ_t, ρ_t, $\pi_{c_0 st}, \ldots, \pi_{c_{t-1} st}$) together with δ_{cst}.

4. Results

A. Sequences of High- and Low-Quality Classrooms

We start by a simple visualization of the data. 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, the first panel, 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 0.14 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 0.21 GE higher than those in a *BB* sequence.

Taken together, the panels in Figure 2 show that larger numbers of *G* classrooms in a sequence generally lead to higher achievement, as one would expect. At the end of 6th grade, students who were in a *G* classroom for seven years in a row (sequence *GGGGGGG*) have scores that are 1.74 GE higher than those who were in a *B* classroom in every grade, a remarkably large difference.

Keeping constant the number of *G* classrooms in the sequence, the bars are not 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. In particular, the bars are frequently taller in sequences in which *G* classrooms are more recent. For example, in 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 good 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 the 3rd grade panel), the impact of *BGGG* is lower than the impact of *GBGG*.[13]

It is possible to aggregate the impacts of the various sequences on learning in different ways, but one that is both particularly simple and instructive for the purposes of our paper is shown in Figure 3. In each panel of this figure, we average the height of the bars of the same color in the corresponding panels of Figure 2. Take, for example, the 3rd grade panel. The zero good classrooms case is represented in every figure as a benchmark, although the height of that bar is by definition zero. Then, the first actual bar in the 3rd grade panel of Figure 3 (labeled "One", for

[13] 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 D1 in Appendix D 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 of being assigned to a sequence on TVIP scores 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 D1 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.

one good classroom) shows the average height of the bars corresponding to the sequences of classrooms with only one good classroom in the 3rd grade panel of Figure 2 (first group of four bars). The second bar in the 3rd grade panel of Figure 3 (labeled "Two") averages the height of all the bars in the analogous panel of Figure 2 corresponding to sequences of classrooms with exactly two good classrooms. The third bar in the 3rd grade panel of Figure 3, corresponds to the average of bars for the three good classroom sequences in the 3rd grade panel in Figure 2. The last bars in both 3rd grade panels of Figures 2 and 3 are the same, because at the end of third grade there is only one sequence with four good classrooms.

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 the first three panels (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. Overall, we cannot reject for any grade that achievement is a linear function of the number of good classrooms in the sequence.

This (perhaps surprisingly) 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 previously in each sequence.[14]

B. 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 to consider). The number of classroom effects and interactions in the model increases with the grade we consider, since we interact classroom effects for two grades at the end of 1st grade, but we interact classroom effects for seven grades at the end of 6th grade. We start by using only controls (test scores, maternal education, and household wealth) measured at baseline, which

[14] Appendix E shows that results are essentially the same when using leave-one-out measures of classroom and school VA to construct the *GB* sequences.

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 2. 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 2 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 between classroom effects.[15] In Panel B we show that the results are very similar if we control for test scores in grade *t-1* ($Y_{sc_0 \ldots c_t t-1 j}$), as opposed to baseline test scores ($Y_{sc_0 \ldots c_t 0 j}$).

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.

C. 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 quality, or classroom input (two expressions that we use interchangeably in this section). As discussed above, we allow for grade-specific production functions. The estimated parameters (θ_t, ρ_t, $\pi_{c_k st}$) as well as the estimated classroom inputs (δ_{csk}) are reported in Tables C1 and C2 in the appendix. In Table 3 we show the implied estimates of the elasticity of substitution ($\frac{1}{1-\rho_t}$) between classroom inputs in different grades. The well-known Cobb-Douglas production function is a useful benchmark, since it has an elasticity of substitution equal to 1 ($\rho_t = 0$). For the case of the CES function we estimate, the elasticity of substitution can be as low as zero (perfect complements) or as high as $+\infty$ (perfect substitutes).

[15] 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.

Classroom inputs are most complementary across grades for test scores at the end of 1st grade, when we only have two inputs: classroom quality in kindergarten and 1st grade. Classroom inputs are considerably more substitutable across grades in the production of achievement at the end of all the remaining grades. Furthermore, the elasticity of substitution for all these grades is above 1.8.

It is helpful to visualize what these estimates imply for the production of achievement. To do so we simulate average predicted scores for different 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.[16] 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 the 1st grade panel, the P25 line shows how achievement at the end of 1st grade depends on 1st grade classroom quality when kindergarten classroom quality is fixed at the 25th percentile of the distribution. As another example, the P75 line in the 5th grade panel shows how achievement at the end of 5th grade depends on fifth grade classroom quality when classroom quality in each of the other grades is fixed at the 75th percentile of the within-grade distribution of classroom quality.

As expected, the three lines are upward sloping in all panels, indicating that the marginal product of classroom quality is always positive. For the same reason, in every panel the P75 line is everywhere above the P50 line, which in turn is above the P25 line. For 1st and 2nd grade it is difficult to distinguish the three lines. This is because the more recent inputs are considerably more important than the lagged inputs (especially in 1st grade), and because the most recent input has a higher variance than the lagged inputs (especially in 2nd grade). However, as can be seen in the figure, we see 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

[16] 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 over which we represent it, so we consider only values of the input between the 10th and 90th percentiles of its distribution.

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 moved 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 are the differences between the thick and thin lines across all panels. In other words, inputs in different grades are close to perfect substitutes. Overall, this is consistent with the estimated elasticities of substitution in Table 3, which are typically around 2 and 3.

The exception is the case of 1st grade, which has the lowest elasticity of substitution, but in spite of that we cannot distinguish the various lines in the corresponding panel of Figure 4. Recall that there are only two inputs to consider up to end of 1st grade achievement: kindergarten and 1st grade classroom quality. We estimate that even though these two inputs appear to be complementary, the productivity of the kindergarten input is much lower than that of the 1st grade input, so we can barely see any differences between the various lines in the corresponding panel of Figure 4.

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 especially salient if we consider large shifts in inputs. In other words, perhaps there is not much difference in the marginal product of inputs in *t* when inputs in previous grades are either at the 25th or 75th percentiles of their distributions, but there may be more visible differences in this same marginal product when we evaluate the production function at the 10th and 90th percentiles of the distributions of previous inputs.

We see that, in fact, 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 of accumulated skills at the end of 2nd grade). For both the P10 and P90 in later grades there are more visible differences across panels between the estimated production function and the simulated production function with perfect substitutability, and these departures from linearity suggest that there could be some dynamic complementarity between inputs (since the gap between the P10 and P90 lines open up slightly more with increases in current classroom quality for the actual production function, than for the simulated production function with perfect substitutability). That said, overall, these still represent small departures from the case of perfect substitutes.

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 inputs in different grades is unlikely to be an important feature of our data.

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, 2000) 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.[17]

Alternatively, teachers may not tailor instruction to individual students, but they may focus on material that is particularly relevant for lagging students.[18] 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).

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 additive in classroom inputs, at least in the setting that we study.

[17] 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.

[18] 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.

References

Agostinelli, Francesco, and Matthew Wiswall. 2016. "Estimating the Technology of Children's Skill Formation." NBER Working Paper 22442.

Agostinelli, Francesco, and Matthew Wiswall. 2016. "Identification of Dynamic Latent Factor Models: The Implications of Re-Normalization in a Model of Child Development." NBER Working Paper 22441.

Almond, Douglas, and Bhashkar Mazumder. 2013. "Fetal Origins and Parental Responses." *Annual Review of Economics* 5(1): 37-56.

Araujo, M. Caridad, Pedro Carneiro, Yyannu Cruz-Aguayo, and Norbert Schady. 2016. "Teacher Quality and Learning Outcomes in Kindergarten." *Quarterly Journal of Economics* 131(3): 1415-53.

Attanasio, Orazio, Sarah Cattan, Emla Fitzsimons, Costas Meghir, and Marta 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, Orazio, Costas Meghir, and Emily Nix. 2020a. "Human Capital Development and Parental Investment in India.". *Review of Economic Studies*, 87(6), 2511-2541.

Attanasio, Orazio, Raquel Bernal, Michele Giannola and Milagros Nores. 2020b. "Child Development in the Early Years: Parental Investments and the Changing Dynamics of Different Dimensions." NBER Working Paper 27812.

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

Caeyers, Bet, and Marcel Fafchamps. 2020. "Exclusion Bias in the Estimation of Peer Effects". Working Paper.

Carneiro, Pedro, Yyannu Cruz-Aguayo, Francesca Salvati and Norbert Schady. 2021. "The Effect of Classroom Rank Throughout Elementary School: Experimental Evidence from Ecuador." Working Paper.

Chetty, Raj, John Friedman, and Jonah Rockoff. 2014a. "Measuring the Impact of Teachers I: Evaluating Bias in Value-Added Estimates." *American Economic Review* 104(9): 2593-2632.

----------. 2014b. "Measuring the Impact of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review* 104(9): 2633-2679.

Cunha, Flavio, and James Heckman. 2007. "The Technology of Skill Formation." *American Economic Review, Papers and Proceedings* 97(2): 31-47.

Cunha, Flavio, James Heckman, and Susanne Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78(3): 883-931.

Duflo, Esther, Pascaline Dupas, and Michael 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, Lloyd, Delia Lugo, Eligio Padilla, and Leota Dunn. 1986. *Test de Vocabulario en Imágenes Peabody* (Circle Pines, MN: American Guidance Service).

Freyberger, Joachim. 2020. "Normalizations and Misspecification in Skill Formation Models." Working Paper.

Hanushek, Eric, and Steven Rivkin. 2012. "The Distribution of Teacher Quality and Implications for Policy." *Annual Review of Economics* 4(1): 131-57.

Heckman, James. 2000. "Policies to Foster Human Capital." *Research in Economics* 54(1): 3-56.

Jackson, Kirabo, Jonah Rockoff, and Douglas Staiger. 2014. "Teacher Effects and Teacher-Related Policies." *Annual Review of Economics* 6(1): 801-25.

Jacob, Brian, Lars Lefgren, and David Sims. 2010. "The Persistence of Teacher-Induced Learning Gains." *Journal of Human Resources* 45(4): 915-943.

Jochmans, Koen. 2021. "Testing Random Assignment to Peer Groups." Working Paper.

Johnson, Rucker, and Kirabo Jackson. 2019. "Reducing Inequality Through Dynamic Complementarity: Evidence from Head Start and Public School Spending." *American Economic Journal: Economic Policy*, 11(4), 310-49.

Kinsler, Josh. 2016. "Teacher Complementarities in Test Score Production: Evidence from Primary School." *Journal of Labor Economics* 34(1): 29-61.

Malamud, Ofer, Cristian Pop-Eleches and Miguel Urquiola. 2016. "Interactions Between Family and School Environments: Evidence on Dynamic Complementarities?" NBER Working Paper 22112.

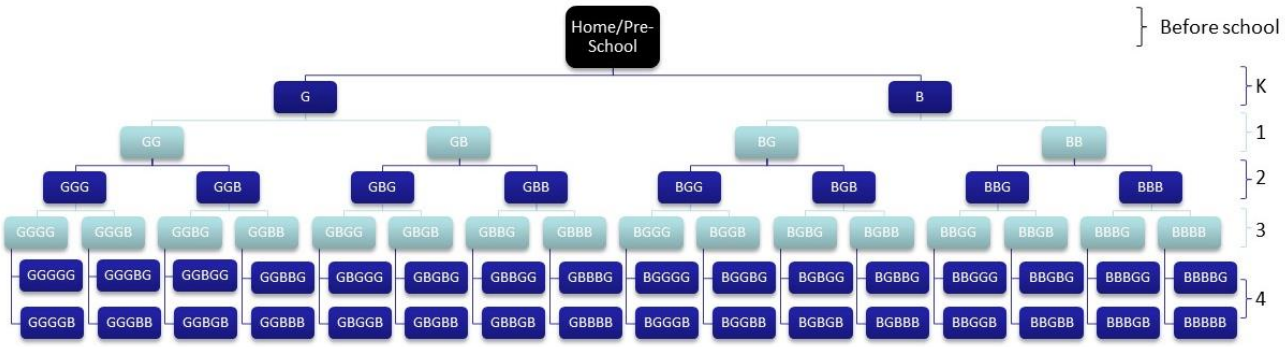
Pop-Eleches, Cristian, and Miguel Urquiola. 2013. "Going to a Better School: Effects and Behavioral Responses." *American Economic Review* 103(4): 1289-1324.

Schady, Norbert. 2012. "El Desarrollo Infantil Temprano en América Latina y el Caribe: Acceso, Resultados y Evidencia Longitudinal de Ecuador." In *Educación para la Transformación*, Marcelo Cabrol and Miguel Székely, eds. (Washington, DC: Inter-American Development Bank).

Schady, Norbert, Jere Behrman, M. Caridad Araujo, Rodrigo Azuero, Raquel Bernal, David Bravo, Florencia Lopez-Boo, Karen Macours, David Marshall, Christina Paxson, and Renos Vakis, "Wealth Gradients in Early Childhood Cognitive Development in Five Latin American Countries." 2015. *Journal of Human Resources* 50 (2): 446–63.

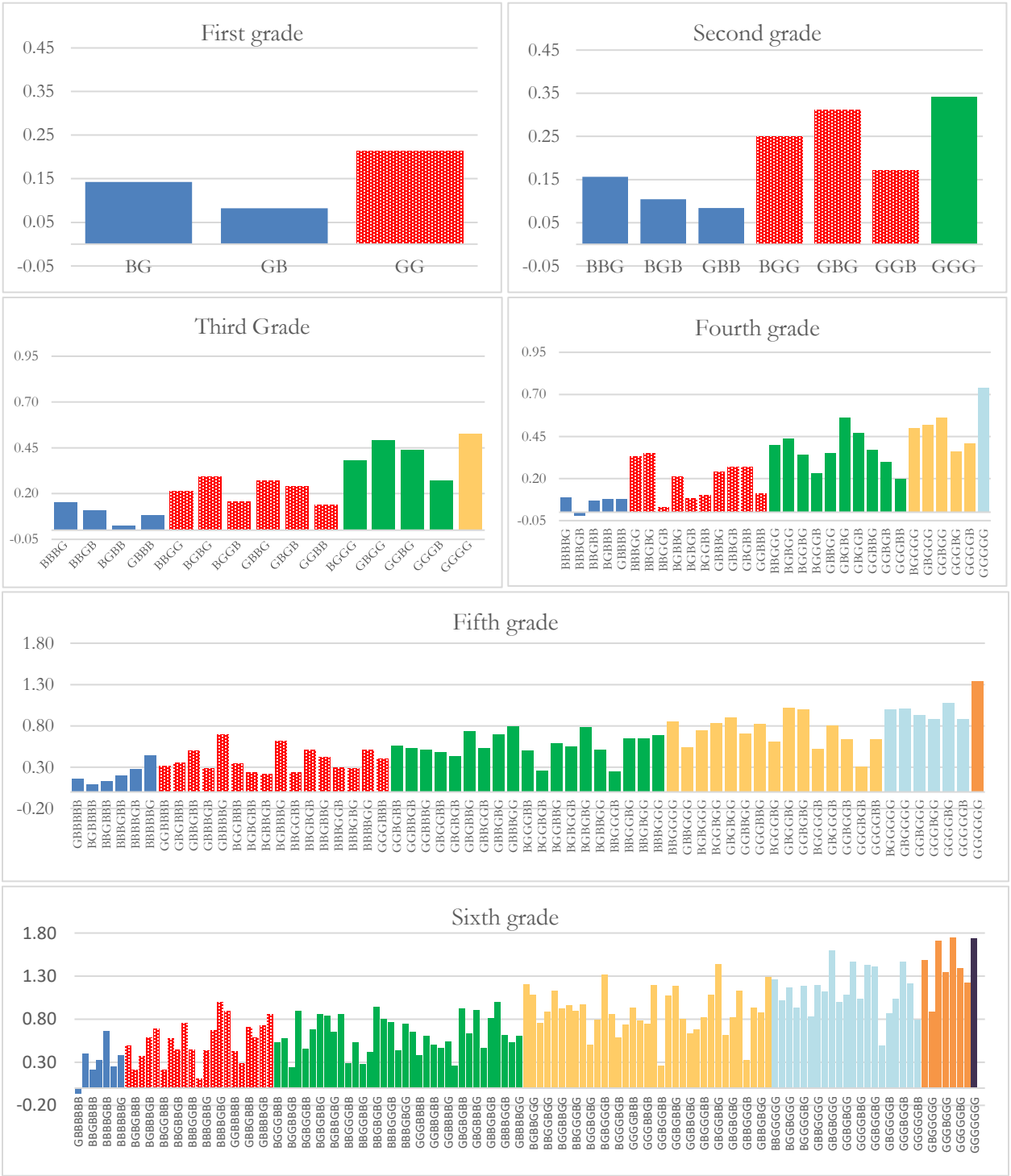
Figures and Tables

Figure 1 – Sequences



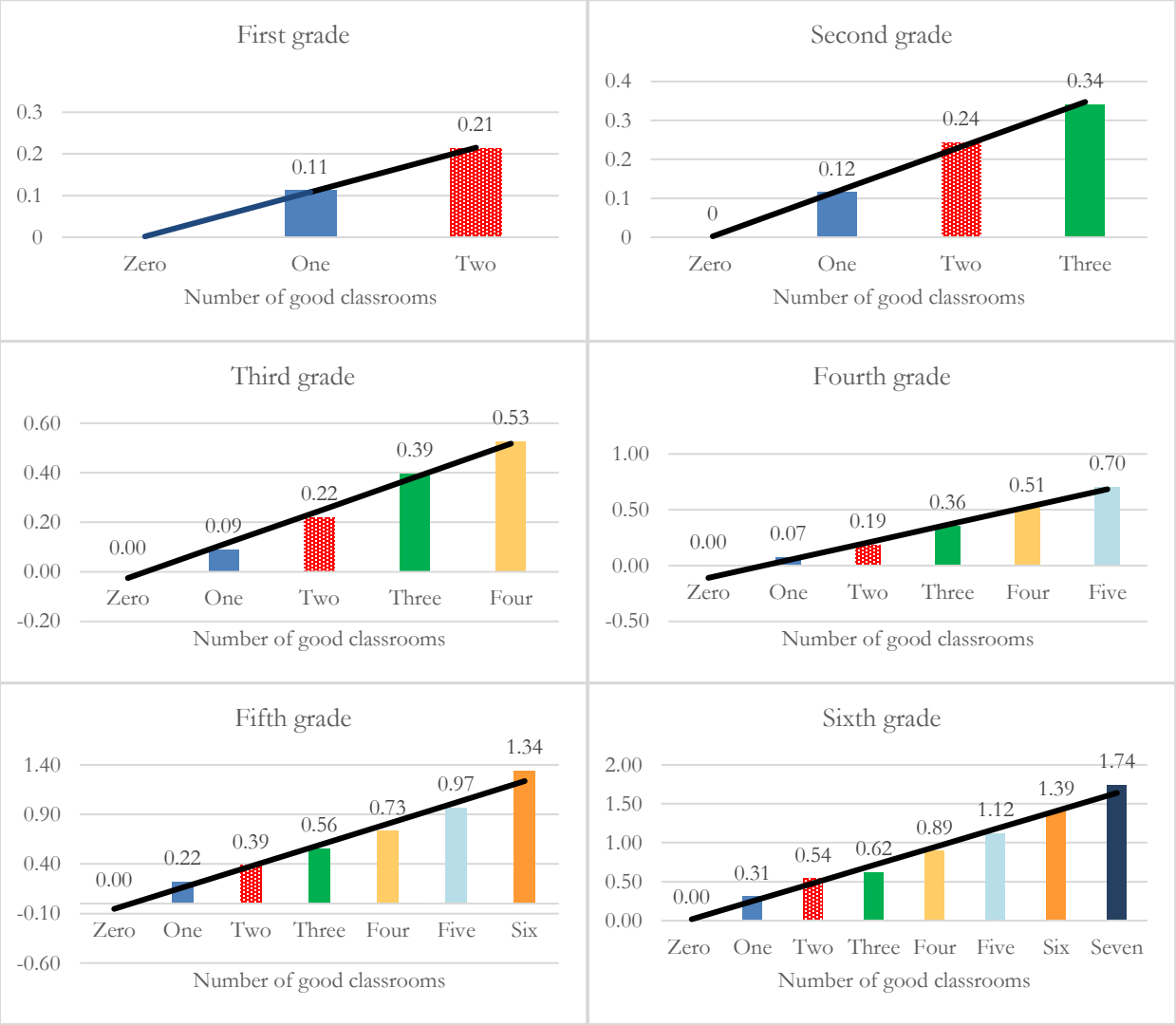
Note: This figure shows the sequences of classroom quality which are possible between kindergarten and fourth grade. G denotes a classroom with value added above the average in the school, where B denotes a classroom with below school average value added.

Figure 2: Impact of sequences of classroom quality on achievement



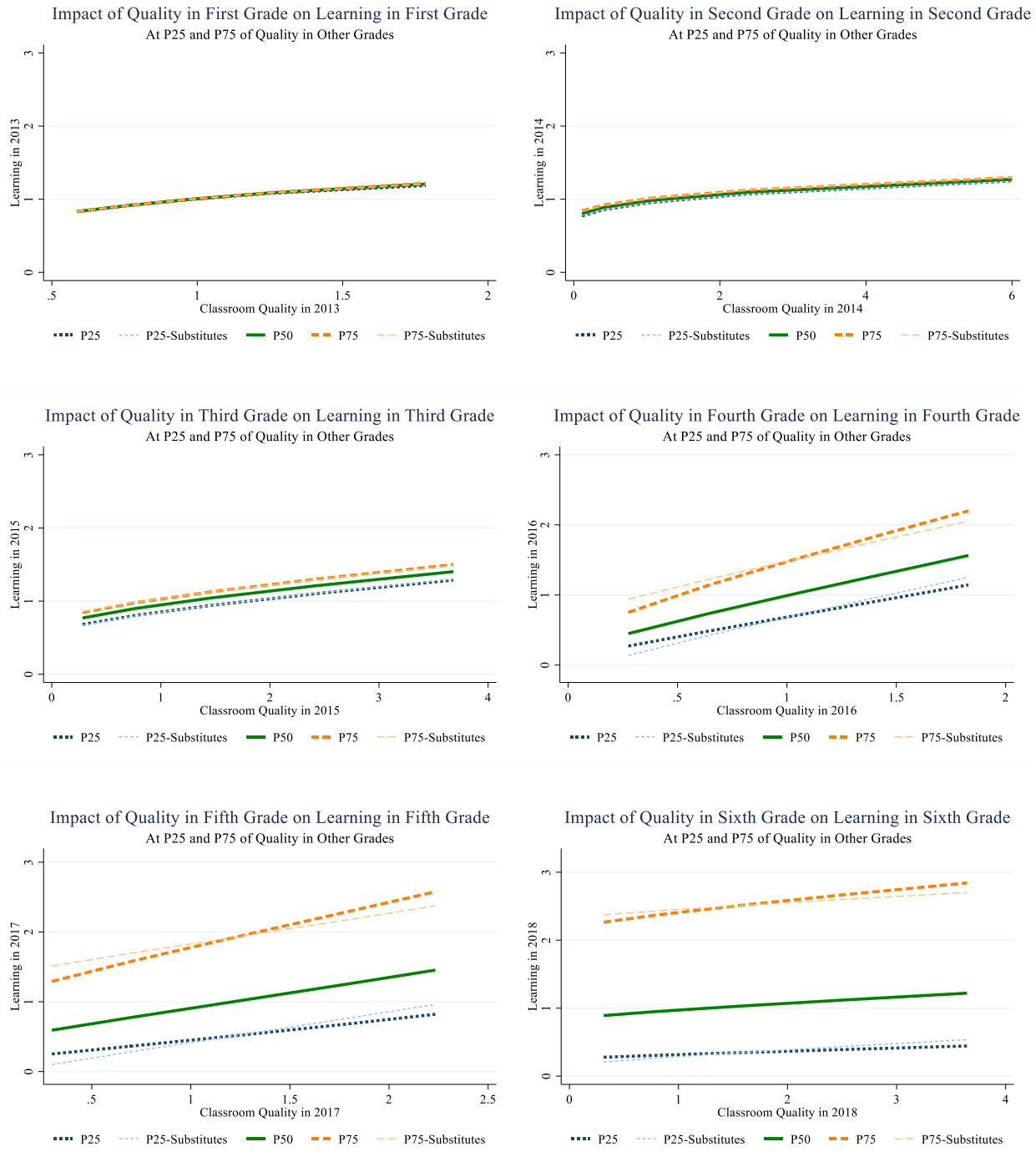
Note: Each panel in this figure (A, B, C, D, E and F) shows average residual learning at the end of each grade (1st, 2nd, 3rd, 4th, 5th, and 6th) for students in different sequences of classroom quality. B denotes a bad classroom in the sequence and G denotes a good classroom in the sequences (so, for example, GBBGG in panel D means that, by the end of 4th grade, students in this sequence experienced an above school average classroom in kindergarten, 3rd and 4th grade, and a below school average classroom 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 and wealth (as well as school fixed effects).

Figure 3: Impact of the number of good classrooms across grades on achievement



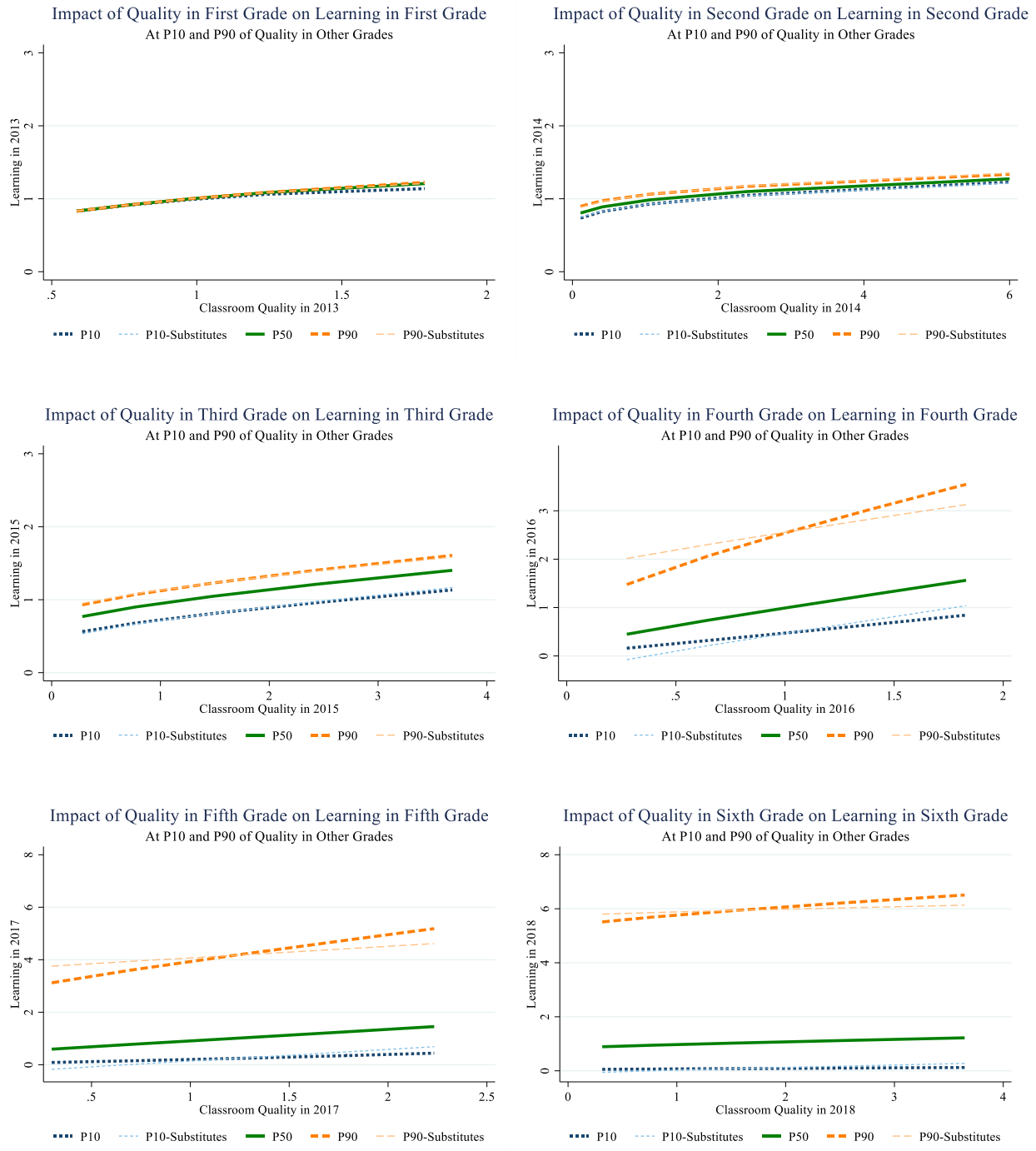
Note: Each panel in this figure shows average residual learning at the end of each grade (1st, 2nd, 3rd, 4th, 5th, 6th) for students in sequences with different numbers of good classrooms, relatively to students with zero good classrooms up to the grade achievement is measured. Residual learning is achievement in math and language at the end of a grade after controlling for age, gender, baseline TVIP, maternal education and wealth (as well as school fixed effects).

Figure 4: Impact of classroom quality on achievement at different values (percentiles 25 and 75) of past classroom quality



Note: 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 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



Note: 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 and wealth, as well as school fixed effects.

Table 1: Descriptive Statistics

Panel A: Children	
Age (months)	67.498
	(4.135)
Proportion female	0.497
	(0.500)
TVIP	83.961
	(16.685)
Mother's age	30.524
	(6.616)
Fathers's age	34.744
	(7.934)
Mother's years of schooling	8.866
	(3.756)
Fathers's years of schooling	8.493
	(3.747)

Panel B: Teachers

	(1)	(2)	(3)	(4)	(5)
	Kindergarten	1st grade	2nd grade	3rd grade	4th 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 tenured	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	17.491	18.013
	(1.932)	(2.061)	(2.530)	(2.277)	(2.431)
Class size	34.619	37.892	39.544	37.262	38.762
	(8.000)	(7.479)	(7.528)	(6.685)	(6.414)
CLASS score	3.407	3.289	3.337	3.281	3.394
	(0.283)	(0.232)	(0.242)	(0.240)	(0.185)

Note: Panel A shows means and standard deviations of student and family characteristics in our sample. 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. Standard errors in parentheses.

Table 2: Testing Interactions Between Classroom Effects in Different Grades

	Fraction of schools with p-value less than		
	10%	5%	1%
Panel A - Control for Baseline Scores			
First Grade	0.09	0.03	0.00
Second Grade	0.07	0.03	0.01
Third Grade	0.12	0.05	0.01
Fourth Grade	0.11	0.07	0.00
Fifth Grade	0.12	0.06	0.01
Sixth Grade	0.02	0.00	0.00
Panel B - Control for *t-1* Scores			
First Grade	0.09	0.06	0.01
Second Grade	0.10	0.05	0.01
Third Grade	0.10	0.07	0.01
Fourth Grade	0.10	0.04	0.01
Fifth Grade	0.17	0.07	0.02
Sixth Grade	0.05	0.02	0.02

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 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, mothers's education and and wealth index, as well as child gender and age.

Table 3: Estimates of the elasticity of substitution between classroom quality in different grades

	Grade					
	1	2	3	4	5	6
Elasticity of substitution $(1/(1-\rho))$	0.2 (0.04)	1.93 (0.13)	2.4 (0.09)	1.85 (0.01)	3.28 (0.04)	3.72 (0.04)

This table shows estimates of the elasticity of substitution between classroom quality in different grades. This corresponds to $\frac{1}{1-\rho}$, where ρ is a parameter in a CES production function (equation 8). 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. Standard errors are reported in parenthesis.

Appendices

Appendix A – 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 *j* in grade *t* looks like:

$$D_{i1st} = 1\left[a_{1st} + b_{1st}\theta_{ist} + \varepsilon_{ijst} > 0\right]$$

...

$$D_{iJst} = 1\left[a_{Jst} + b_{Jst}\theta_{ist} + \varepsilon_{iJst} > 0\right]$$

With logit errors we get the standard 2 parameter IRT model:

$$\Pr\left(D_{i1st} = 1|\theta_{ist}\right) = \frac{e^{a_{1st}+b_{1st}\theta_{ist}}}{1 + e^{a_{1st}+b_{1st}\theta_{ist}}}$$

...

$$\Pr\left(D_{iJst} = 1|\theta_{ist}\right) = \frac{e^{a_{Jst}+b_{Jst}\theta_{ist}}}{1 + e^{a_{Jst}+b_{Jst}\theta_{ist}}}$$

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.

$$\Pr\left(D_{i1sK} = 1|\theta_{isK}\right) = \frac{e^{a_{1sK}+b_{1sK}\theta_{isK}}}{1 + e^{a_{1sK}+b_{1sK}\theta_{isK}}}$$

...

$$\Pr\left(D_{iJsK} = 1|\theta_{isK}\right) = \frac{e^{a_{JsK}+b_{JsK}\theta_{isK}}}{1 + e^{a_{JsK}+b_{JsK}\theta_{isK}}}$$

From this procedure we obtain estimates $(\hat{a}_{1sK}, \dots, \hat{a}_{JsK}, \hat{b}_{1sK}, \dots, \hat{b}_{JsK})$. 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:

$$\Pr\left(D_{i1s1} = 1|\theta_{is1}\right) = \frac{e^{a_{1s1}+b_{1s1}\theta_{is1}}}{1+e^{a_{1s1}+b_{1s1}\theta_{is1}}}$$

...

$$\Pr\left(D_{iJs1} = 1|\theta_{is1}\right) = \frac{e^{a_{Js1}+b_{Js1}\theta_{is1}}}{1+e^{a_{Js1}+b_{Js1}\theta_{is1}}}$$

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 $(\hat{a}_{1s1}, \dots, \hat{a}_{Js1}, \hat{b}_{1s1}, \dots, \hat{b}_{Js1})$, 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:

$$\Pr\left(D_{i1s2} = 1|\theta_{is2}\right) = \frac{e^{a_{1s2}+b_{1s2}\theta_{is2}}}{1+e^{a_{1s2}+b_{1s2}\theta_{is2}}}$$

...

$$\Pr\left(D_{iJs2} = 1|\theta_{is2}\right) = \frac{e^{a_{Js2}+b_{Js2}\theta_{is2}}}{1+e^{a_{Js2}+b_{Js2}\theta_{is2}}}$$

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 it within our sample. We start by computing

$$M_K = median(\theta_{isK})$$

...

$$M_6 = median(\theta_{is6})$$

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 $(\theta_{isK}, \dots, \theta_{is6})$ are continuous variables. We need to fill in the remaining points in the function by fitting a function $g_s(.)$ to this data:

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

The function $g_s(.)$, which needs to be estimated, converts scores θ_{ist} into grade equivalents φ_{ist}. It turns out that a an exponential function provides a good fit for $g_s(.)$ (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 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 very low end of kindergarten scores, or 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 posterior distribution of factor for each individual, given their response patterns), which depend on how many items are there 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 A1 shows percentiles of the distribution of grade equivalents resulting from this procedure:

Table A1: Distribution of grade equivalent scores at the end of each grade

	Grade						
	K	1	2	3	4	5	6
Percentile							
10	0.74	1.27	2.05	2.71	3.43	4.21	4.82
25	0.96	1.48	2.59	3.34	4.10	5.06	5.83
50	1.09	1.84	3.14	4.02	4.90	6.09	7.04
75	1.22	2.26	3.70	4.74	5.79	7.22	8.42
90	1.40	2.74	4.24	5.47	6.69	8.33	9.75

This table shows percentiles 10, 25, 50, 75 and 90 of the distribution of grade equivalent scores at the end of each grade.

Appendix B – 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.[19] 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 (2020).

First, we briefly discuss the procedure outlined in Jochmans (2020). Consider our setting, in which we observe data on S schools, and each school has $n_1, \dots, n_s$ 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 (2020) 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

$$ts = \frac{\sum_{s=1}^{S}\sum_{i=1}^{n_s}\tilde{x}_{s,i}\left(\bar{x}_{s,j} + \frac{x_{s,i}}{n_s - 1}\right)}{\sqrt{\sum_{s=1}^{S}\left(\sum_{i=1}^{n_s}\tilde{x}_{s,i}\left(\bar{x}_{s,j} + \frac{x_{s,i}}{n_s - 1}\right)\right)^2}} \qquad (B.1)$$

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

[19] We use the word "random" as shorthand but, as discussed at length in Araujo et al. (2016) and Campos et al. (2020), 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.

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 B1 and B2, 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 B1: 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 (2020). 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 B2: 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 method- ology proposed by Jochmans (2020). 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.

Appendix C – Procedure for estimating the production function

Basic Model

Equation (4) defines a system of equations, one for each grade $t = 0 \dots 6$. In order to estimate it, we start by taking logs:

$$lnY_{sc_0 \dots c_t tj} = \mu_{st} + X_{sc_0 \dots c_t tj}\gamma_t + \frac{\theta_t}{\rho_t} ln\left(\sum_{k=0}^{t} \pi_{c_k st} {\delta_{csk}}^{\rho_t}\right) + v_{sc_0 \dots c_t tj} \quad (8)$$

We define $v_{sc_0 \dots c_t tj} = lnu_{sc_0 \dots c_t tj}$. 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:

$$lnY_{sc_0 0j} = \mu_{s0} + X_{sc_0 0j}\gamma_0 + \theta_0 \ln\left(\pi_{c_0 s0}^{\frac{1}{\rho_0}}\right) + \theta_0 ln(\delta_{cs0}) + v_{sc_0 0j} \quad (9)$$

This is a standard VA equation for kindergarten, where $lnY_{sc_0 0j}$ is a linear function of classroom assignment indicators, which are estimated to be $\theta_0 ln(\delta_{cs0})$. 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 main text of 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 equation per grade) and the vector of classroom qualities are identified, and should be estimated simultaneously.

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 δ_{cs0} for each classroom, *c=A,B,C*. Start from the production function for grade 1 achievement.

Define $Y_{sc_0c_11} = E\left(lnY_{sc_0c_11j}|c_0, c_1\right) = \frac{\theta_t}{\rho_t} ln\left[\pi_{1s1}{\delta_{c_0s0}}^{\rho_1} + (1-\pi_{1s1}){\delta_{c_1s1}}^{\rho_1}\right]$. Then:

$$Y_{sAA1} = E\left(lnY_{sc_0c_11j}|c_0 = A, c_1 = A\right) = \frac{\theta_t}{\rho_t} ln[\pi_{1s1}{\delta_{As0}}^{\rho_1} + (1-\pi_{1s1}){\delta_{As1}}^{\rho_1}]$$

$$Y_{sAB1} = E\left(lnY_{sc_0c_11j}|c_0 = A, c_1 = B\right) = \frac{\theta_t}{\rho_t} ln[\pi_{1s1}{\delta_{As0}}^{\rho_1} + (1-\pi_{1s1}){\delta_{Bs1}}^{\rho_1}]$$

$$Y_{sAC1} = E\left(lnY_{sc_0c_11j}|c_0 = A, c_1 = C\right) = \frac{\theta_t}{\rho_t} lnA[\pi_{1s1}{\delta_{As0}}^{\rho_1} + (1-\pi_{1s1}){\delta_{Cs1}}^{\rho_1}]$$

...

$$Y_{sBA1} = E\left(lnY_{sc_0c_11j}|c_0 = B, c_1 = A\right) = \frac{\theta_t}{\rho_t} ln[\pi_{1s1}{\delta_{Bs0}}^{\rho_1} + (1 - \pi_{1s1}){\delta_{As1}}^{\rho_1}]$$

...

$$Y_{sCC1} = E\left(lnY_{sc_0c_11j}|c_0 = C, c_1 = C\right) = \frac{\theta_t}{\rho_t} ln[\pi_{1s1}{\delta_{Cs0}}^{\rho_1} + (1 - \pi_{1s1}){\delta_{Cs1}}^{\rho_1}]$$

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

$$\frac{Y_{sAA1}}{Y_{sAB1}} = \frac{ln[\pi_{1s1}{\delta_{As0}}^{\rho_1} + (1 - \pi_{1s1}){\delta_{As1}}^{\rho_1}]}{ln[\pi_{1s1}{\delta_{Bs0}}^{\rho_1} + (1 - \pi_{1s1}){\delta_{As1}}^{\rho_1}]}$$

$$\frac{Y_{sAA1}}{Y_{sAC1}} = \frac{ln[\pi_{1s1}{\delta_{As0}}^{\rho_1} + (1 - \pi_{1s1}){\delta_{As1}}^{\rho_1}]}{ln[\pi_{1s1}{\delta_{Bs0}}^{\rho_1} + (1 - \pi_{1s1}){\delta_{Cs1}}^{\rho_1}]}$$

$$\frac{Y_{sAA1}}{Y_{sBA1}} = \frac{ln[\pi_{1s1}{\delta_{As0}}^{\rho_1} + (1 - \pi_{1s1}){\delta_{As1}}^{\rho_1}]}{ln[\pi_{1s1}{\delta_{Bs0}}^{\rho_1} + (1 - \pi_{1s1}){\delta_{As1}}^{\rho_1}]}$$

$$\frac{Y_{sAA1}}{Y_{sBB1}} = \frac{ln[\pi_{1s1}{\delta_{As0}}^{\rho_1} + (1 - \pi_{1s1}){\delta_{As1}}^{\rho_1}]}{ln[\pi_{1s1}{\delta_{Bs0}}^{\rho_1} + (1 - \pi_{1s1}){\delta_{Bs1}}^{\rho_1}]}$$

$$\frac{Y_{sAA1}}{Y_{sBC1}} = \frac{ln[\pi_{1s1}{\delta_{As0}}^{\rho_1} + (1 - \pi_{1s1}){\delta_{As1}}^{\rho_1}]}{ln[\pi_{1s1}{\delta_{Bs0}}^{\rho_1} + (1 - \pi_{1s1}){\delta_{Cs1}}^{\rho_1}]}$$

$$\frac{Y_{sAA1}}{Y_{sCA1}} = \frac{ln[\pi_{1s1}{\delta_{As0}}^{\rho_1} + (1 - \pi_{1s1}){\delta_{As1}}^{\rho_1}]}{ln[\pi_{1s1}{\delta_{Cs0}}^{\rho_1} + (1 - \pi_{1s1}){\delta_{As1}}^{\rho_1}]}$$

$$\frac{Y_{sAA1}}{Y_{sCB1}} = \frac{ln[\pi_{1s1}{\delta_{As0}}^{\rho_1} + (1 - \pi_{1s1}){\delta_{As1}}^{\rho_1}]}{ln[\pi_{1s1}{\delta_{Cs0}}^{\rho_1} + (1 - \pi_{1s1}){\delta_{Bs1}}^{\rho_1}]}$$

$$\frac{Y_{sAA1}}{Y_{sCC1}} = \frac{ln[\pi_{1s1}{\delta_{As0}}^{\rho_1} + (1 - \pi_{1s1}){\delta_{As1}}^{\rho_1}]}{ln[\pi_{1s1}{\delta_{Cs0}}^{\rho_1} + (1 - \pi_{1s1}){\delta_{Cs1}}^{\rho_1}]}$$

From here we have enough moments to recover π_{1s1}, ρ_1, δ_{As1}, δ_{Bs1}, δ_{Cs1}. Even though δ_{As1}, δ_{Bs1}, δ_{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 independent ratios across two schools;

in addition to θ_1, which can then be recovered from the level equations). Since we have many more than two schools, we can estimate all the parameters of the model, even if all 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_{1s1}). 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 only 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

In practice, instead of estimating the entire model for all grades simultaneously, it is computationally easier to proceed iteratively, one grade at a time, starting with the lower grades. We start from equation (9), $t = 0$, from which we recover estimates of δ_{cs0} for each classroom (and we estimate the remaining parameters of the model, which are not of substantial interest). From the equation for first grade (equation (8) for $t = 1$), we use δ_{cs0} from the $t = 0$ equation, and we estimate all the parameters of the production function (θ_1, ρ_1, $\pi_{c_0 s1}$) together with δ_{cs1} (as well as the parameters on the controls). In grade *t*, we use $\{\delta_{cs0} \dots \delta_{cst-1}\}$ obtained from the previous grades' equations, and we estimate (θ_t, ρ_t, $\pi_{c_0 st}, \dots, \pi_{c_{t-1} st}$) together with δ_{cst}.

Within each grade *t*, the procedure has four steps. Again, this greatly facilitates the computation of the estimates given the large number of classroom indicators included in the nonlinear CES specification of the production function. The steps are as follows:

1. In the first step we estimate γ_t from grade specific regressions of log test scores on classroom fixed effects and baseline controls: $lnY_{sc_0 \dots c_t tj} = \mu_{sct} + X_{sc_0 \dots c_t tj}\gamma_t + v_{sc_0 \dots c_t tj}$. In principle, instead of μ_{sct} one could use indicators for the whole sequence of classroom assignments. We show below that this increases substantially the number of parameters to be estimated without any substantial change in our results. From this equation we recover γ_t for each grade, and we can then also estimate $ln\tilde{Y}_{sc_0 \dots c_t tj} = lnY_{sc_0 \dots c_t tj} - X_{sc_0 \dots c_t tj}\gamma_t$. We then use this quantity to estimate the production function:

$$ln\tilde{Y}_{sc_0\ldots c_t tj} = \mu_{st} + \frac{\theta_t}{\rho_t} ln\left(\sum_{k=0}^{t} \pi_{c_k st} \delta_{csk}{}^{\rho_t}\right) + v_{sc_0\ldots c_t tj} \quad (10)$$

2. It is easier to describe the second step of the procedure by imagining that we start by guessing initial values for δ_{cst} ($\{\delta_{cs0} \ldots \delta_{cst-1}\}$ having already been estimated and therefore used as data in this step). One possible initial guess comes from estimates of classroom effects for grade *t* from linear VA models. Given initial guesses for δ_{cst}, we use nonlinear least squares to estimate the remaining parameters.
3. For the third step we take expectations on both sides of equation (7) and solve for δ_{cst}:

$$\delta_{cst} = \frac{e^{\left\{[E(ln\tilde{Y}_{sc_0\ldots c_t tj} | c_0 \ldots c_t) - (\mu_{st} + X_{sc_0\ldots c_t tj}\gamma_t)]\frac{\rho_t}{\theta_t}\right\}} - \sum_{k=0}^{t-1} \pi_{c_k st} \delta_{csk}{}^{\rho_t}}{1 - \sum_{k=0}^{t-1} \pi_{c_k st}} \quad (11)$$

   We then use the estimates in step 3 as initial values in step 2 and loop between these two steps until the procedure converges.
4. Finally, we restart the algorithm from step 2 using completely new initial values for δ_{cst}. To generate these new initial values, we first take random draws for (θ_t, ρ_t, $\pi_{c_0 st}$, …, $\pi_{c_{t-1} st}$), and we use them to generate values of δ_{cst} consistent with these random draws, using equation (10). After looping over 500 different starting values for δ_{cst}, we pick the set of estimates that minimizes the sum of squared residuals (SSR):

$$SSR = \sum_{j=1}^{N_t} \frac{1}{N_t} \left\{ ln\tilde{Y}_{sc_0\ldots c_t tj} - \left[\mu_{st} + \frac{\theta_t}{\rho_t} ln\left(\sum_{k=0}^{t} \pi_{c_k st} \delta_{csk}{}^{\rho_t}\right)\right]\right\}^2$$

The estimated parameters of the production function are reported in table C1 and the percentiles of the estimated distribution of classroom inputs in each grade are in table C2.

Table C1: Estimates of the parameters of the production function for each grade

	Grade					
	1	2	3	4	5	6
ρ	-4.11	0.48	0.58	0.46	0.70	0.73
	(1.13)	(0.04)	(0.02)	(0.003)	(0.003)	(0.003)
θ	0.37	0.26	1	1.29	1.78	2
	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	.
π_0	0.05	0.08	0.69	0.01	0.04	0.01
	(0.04)	(0.05)	(0.01)	.	(0.01)	.
π_1	0.95	0.45	0.08	0.11	0.11	0.09
	(0.04)	(0.03)	(0.01)	(0.004)	(0.002)	(0.002)
π_2		0.47	0.01	0.26	0.18	0.16
		(0.03)	(0.002)	(0.002)	(0.002)	(0.001)
π_3			0.22	0.06	0.22	0.25
			(0.004)	(0.002)	(0.002)	(0.001)
π_4				0.56	0.16	0.04
				(0.003)	(0.003)	(0.003)
π_5					0.29	0.40
					(0.003)	(0.003)
π_6						0.06
						(0.001)
Elasticity of substitution	0.2	1.93	2.4	1.85	3.28	3.72
	(0.04)	(0.13)	(0.09)	(0.01)	(0.04)	(0.04)

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 are reported in parenthesis. There are three instances where standard errors are not reported. This happened because these are close to the boundary points in the search grid (although the algorithm never stops at a corner).

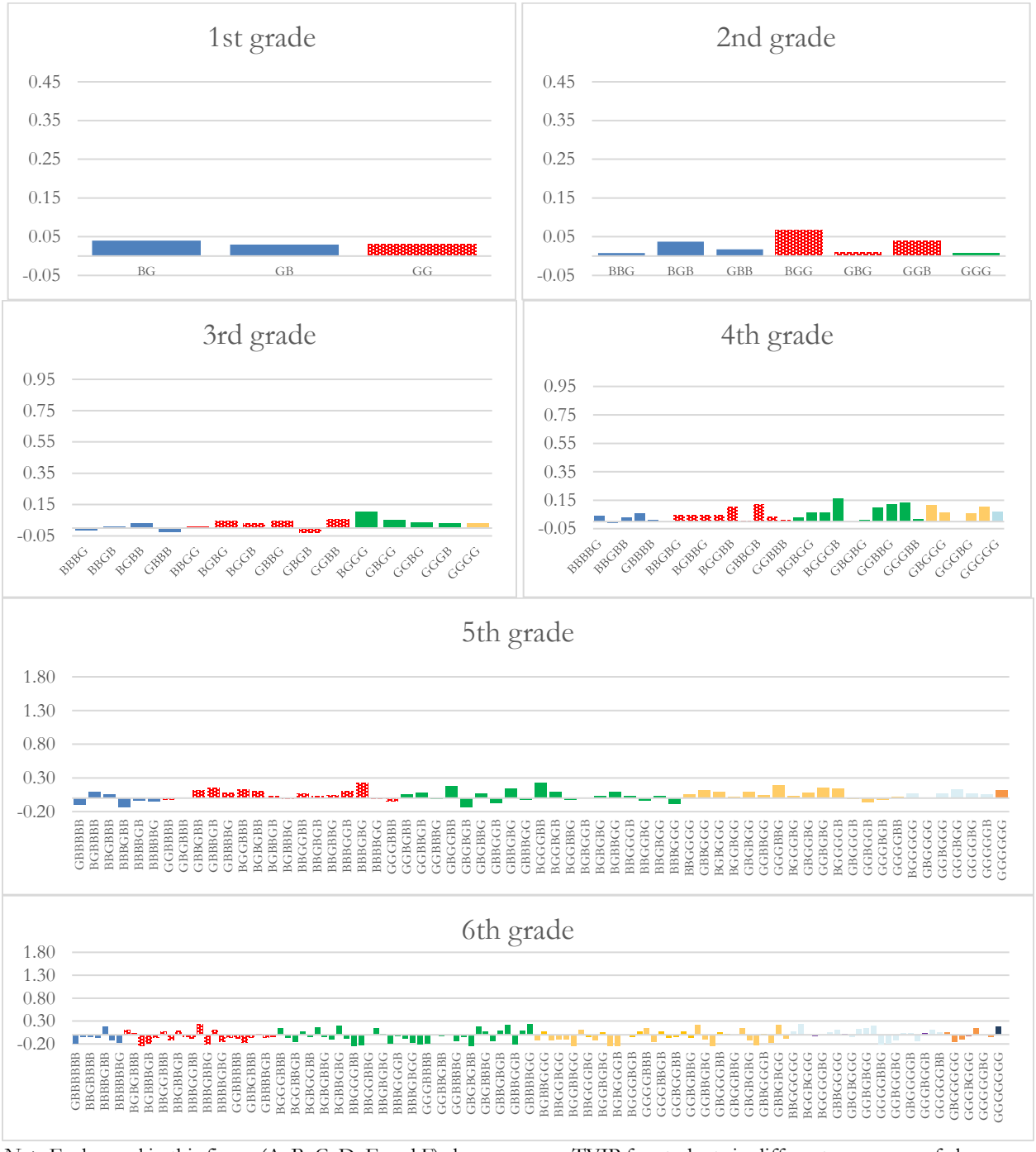
Table C2: Distribution of classroom inputs in each grade

	Grade						
	K	1	2	3	4	5	6
Percentile							
10	0.76	0.59	0.11	0.28	0.27	0.30	0.31
25	0.94	0.76	0.41	0.77	0.65	0.69	0.78
50	1.06	0.97	1.06	1.47	1.03	1.22	1.55
75	1.15	1.24	2.39	2.42	1.42	1.79	2.53
90	1.23	1.79	6.00	3.68	1.83	2.33	3.65

This table shows estimates of the distribution of classroom inputs in each grade estimated from the production function (and the implied elasticity of substitution between inputs in different grades). The production function is specified in equation (8).

Appendix D – Coefficients from the regressions of baseline TVIP on Good-Bad sequences

Figure D1: *Impact* (placebo) of sequences of classroom quality on baseline TVIP



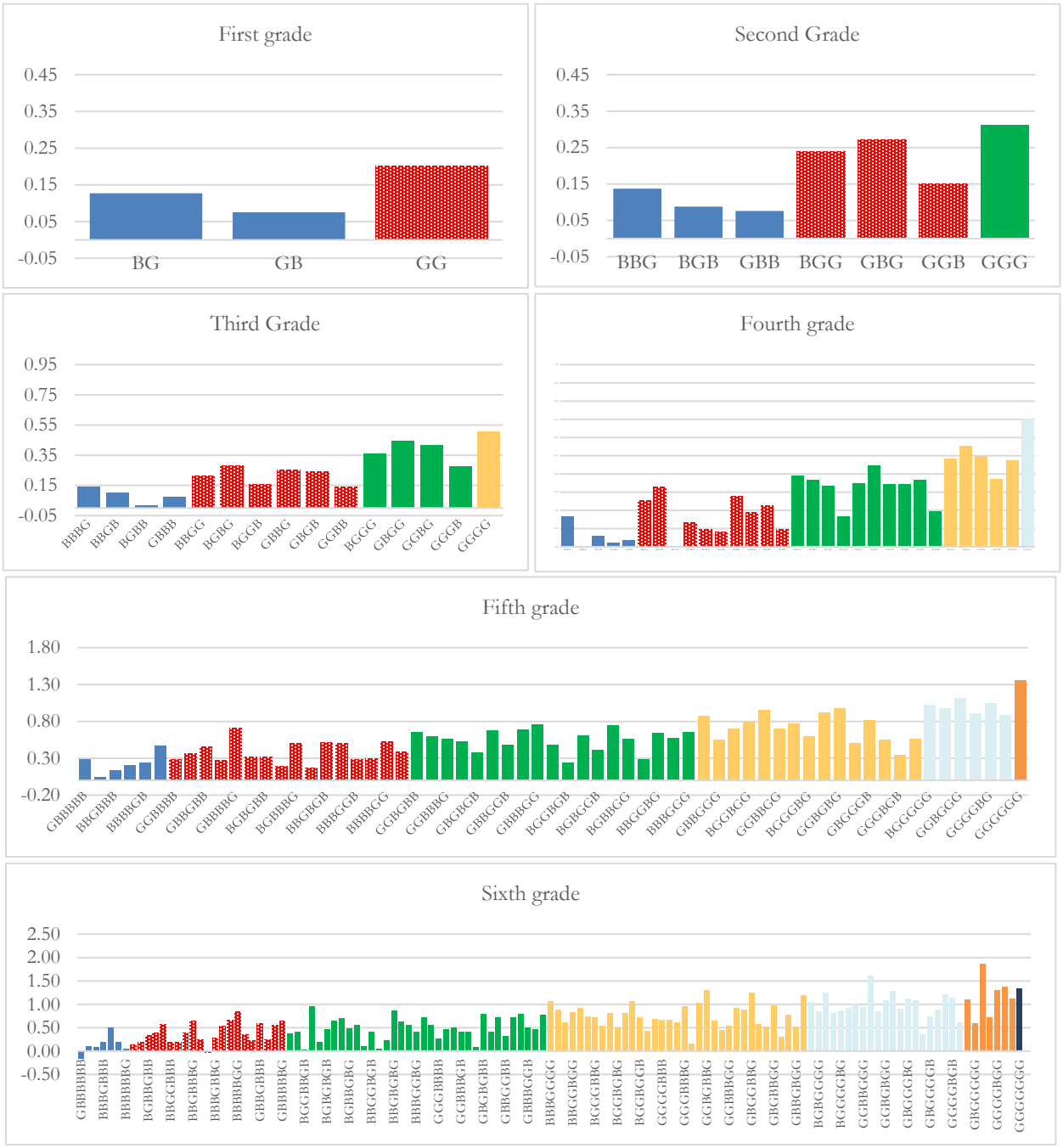
Note: Each panel in this figure (A, B, C, D, E and F) shows average TVIP for students in different sequences of classroom quality. The difference between each panel is the sample of students, corresponding to those in the balanced panel at the end of each grade. B denotes a bad classroom in the sequence and G denotes a good classroom in the sequences (so, for example, GBBGG in panel D means that, by the end of 4th grade, students in this sequence experienced an above school average classroom in kindergarten, 3rd and 4th grade, and a below school average classroom in the remaining grades). The regressions also control for age, gender, maternal education and wealth (as well as school fixed effects).

Figure D2: *Impact* (placebo) of the number of good classrooms across grades baseline TVIP

Note: Each panel in this figure shows average residual learning at the end of each grade (1st, 2nd, 3rd, 4th, 5th, 6th) for students in sequences with different numbers of good classrooms, relatively to students with zero good classrooms up to the grade achievemnt 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 and wealth (as well as school fixed effects).

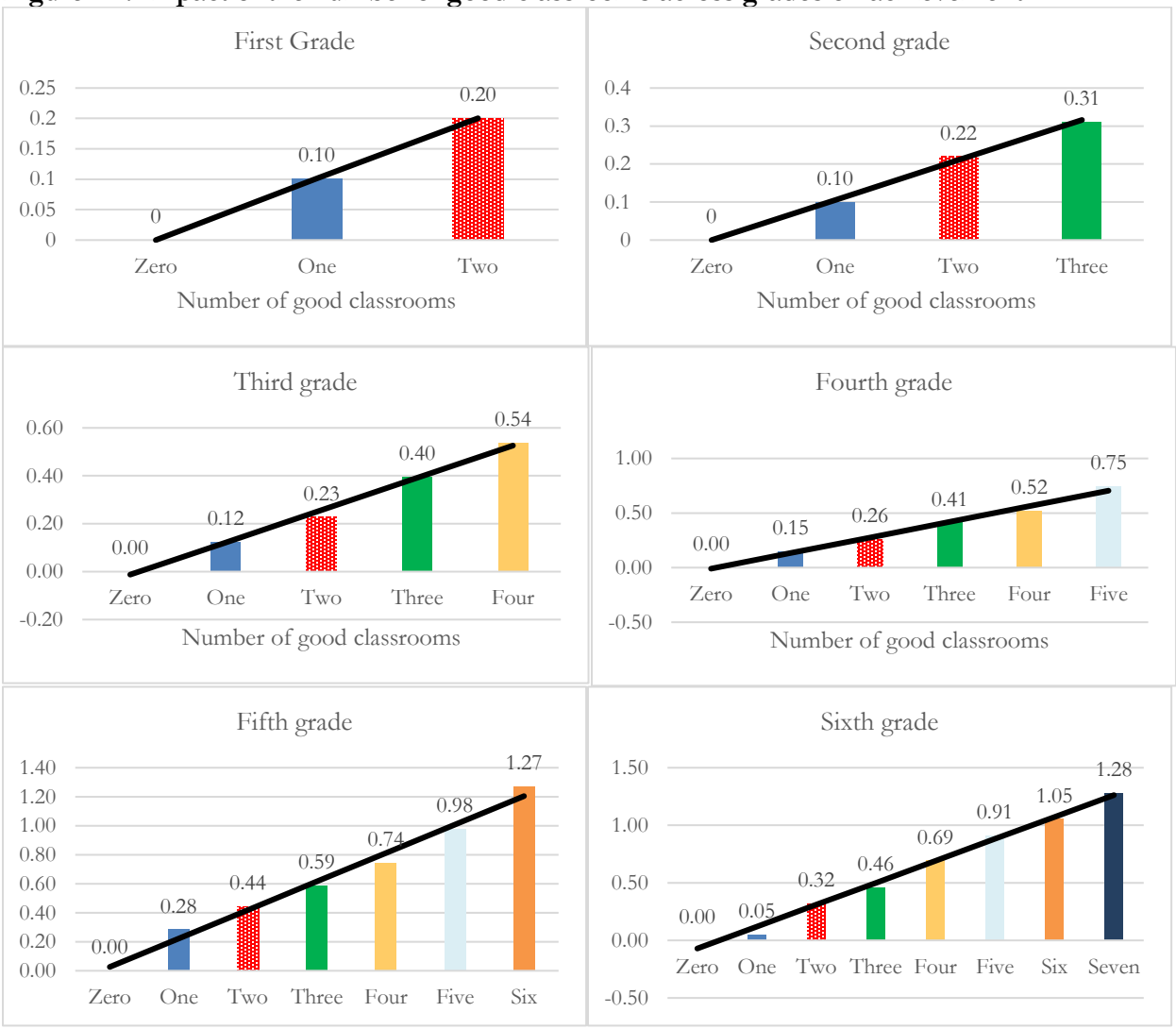
Appendix E – Sequences of classroom quality, leave-one-out (from classroom and school) estimates

Figure E1: Impact of sequences of classroom quality on achievement



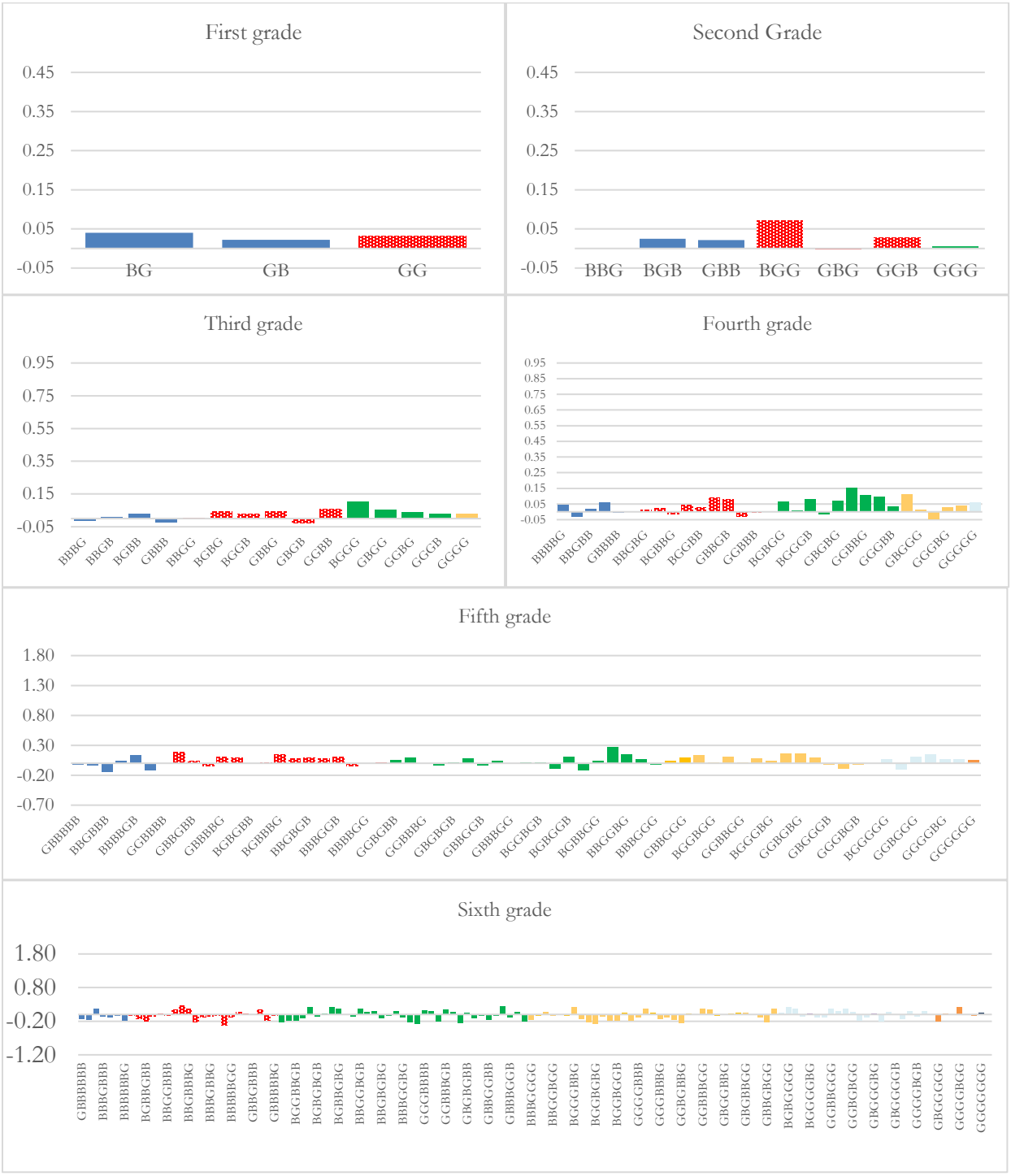
Note: Each panel in this figure (A, B, C, D, E and F) shows average residual learning at the end of each grade (1st, 2nd, 3rd, 4th, 5th, and 6th) for students in different sequences of classroom quality. B denotes a bad classroom in the sequence and G denotes a good classroom in the sequences (so, for example, GBBGG in panel D means that, by the end of 4th grade, students in this sequence experienced an above school average classroom in kindergarten, 3rd and 4th grade, and a below school average classroom 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 and wealth (as well as school fixed effects).

Figure E2: Impact of the number of good classrooms across grades on achievement



Note: Each panel in this figure shows average residual learning at the end of each grade (1st, 2nd, 3rd, 4th, 5th, 6th) for students in sequences with different numbers of good classrooms, relatively to students with zero good classrooms up to the grade achievemnt 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 and wealth (as well as school fixed effects).

Figure E3: *Impact* (placebo) of sequences of classroom quality on baseline TVIP



Note: Each panel in this figure (A, B, C, D, E and F) shows average TVIP for students in different sequences of classroom quality. The difference between each panel is the sample of students, corresponding to those in the balanced panel at the end of each grade. B denotes a bad classroom in the sequence and G denotes a good classroom in the sequences (so, for example, GBBGG in panel D means that, by the end of 4th grade, students in this sequence experienced an above school average classroom in kindergarten, 3rd and 4th grade, and a below school average classroom in the remaining grades). The regressions also control for age, gender, maternal education and wealth (as well as school fixed effects).

Figure E4: *Impact* (placebo) of the number of good classrooms across grades baseline TVIP

Note: Each panel in this figure shows average residual learning at the end of each grade (1st, 2nd, 3rd, 4th, 5th, 6th) for students in sequences with different numbers of good classrooms, relatively to students with zero good classrooms up to the grade achievemnt 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 and wealth (as well as school fixed effects).