The Lost Human Capital

Teacher Knowledge and Student Achievement in Africa[*]

Tessa Bold[a], Deon Filmer[b], Ezequiel Molina[c], Jakob Svensson[d]

April 16, 2019

Abstract: In many low-income countries, teachers do not master the subject they are teaching, and children learn little while attending school. Using unique data from nationally representative surveys of schools in seven Sub-Saharan African countries, we propose a methodology to assess the effect of teacher subject content knowledge on student learning when panel data on students are not available. We show that data on test scores of the student's current and the previous year's teachers, and knowledge of the correlation structure of teacher knowledge across time and grades, allow us to estimate two structural parameters of interest: the contemporaneous effect of teacher content knowledge, and the extent of fade out of teacher impacts in earlier grades. We use these structural estimates to understand the magnitude of teacher effects and to simulate the impacts of various policy reforms. Shortfalls in teachers' content knowledge account for 30 percent of the shortfall in learning relative to the curriculum, and about 20 percent of the cross-country difference in learning in the sample. Assigning more students to better teachers would potentially lead to substantial cost-savings, even if there are negative class-size effects. Ensuring that all incoming teachers have the officially mandated effective years of education, along with increasing the time spent on teaching to the officially mandated schedule, could almost double student learning within the next 30 years.

[*] We are grateful to the members of the SDI team and the many researchers, survey experts and enumerators who have worked on the country surveys that made this paper possible. We would like to especially thank Gayle Martin, Christophe Rockmore, Brian Stacy, and Waly Wane whose collaboration on earlier analysis of these data is gratefully acknowledged. We would like to thank Peter Fredriksson, David Strömberg and Jishnu Das for extensive discussions and suggestions. We appreciate comments from seminar participants at Stockholm School of Economics, NHH, Oxford University, UPF, University of Warwick, and the World Bank. We are grateful to the World Bank, and in particular the Service Delivery Indicators Trust Fund (funded in large part by the Hewlett Foundation) for supporting the research. We also thank the Riksbankens Jubileumsfond (RJ) and Handelsbankens forskningsstiftelser for financial support. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.
[a]IIES, Stockholm University and CEPR, tessa.bold@iies.su.se; [b]The World Bank, dfilmer@worldbank.org; [c]The World Bank, molina@worldbank.org; [d]IIES, Stockholm University and CEPR, jakob.svensson@iies.su.se.

1. Introduction

In many low-income countries, children learn little from attending school (World Bank 2018). For example, four out of five students in Mozambique and Nigeria cannot read a simple word of Portuguese and English, respectively, after more than three years of compulsory language learning. In India, only one in four fourth grade student manages tasks—such as basic subtraction—that are part of the curriculum for the second grade. In Uganda, roughly half of the students cannot order numbers between 0 and 100 after three years of mathematics teaching.[1]

A growing body of evidence—based on teacher value-added and experimental studies—suggests that teacher quality, broadly defined, is a key determinant of student learning.[2] Less is known, however, about what specific dimensions of teacher quality matter. In this paper we focus on teacher subject content knowledge, noting that there is a dearth of well-identified studies on the impact of this knowledge on learning outcomes in low- and middle-income countries (see Glewwe and Muralidharan 2015).

In particular, we investigate the role of teachers' knowledge of the subject they teach, using unique cross-sectional data collected from nationally representative surveys of schools in Kenya, Nigeria, Mozambique, Senegal, Tanzania, Togo, and Uganda (these countries, together, represent close to 40 percent of the region's total population). The data quantifies teacher quality along three core quality dimensions: time spent teaching, teachers' knowledge of the subject they are teaching, and teachers' pedagogical skills (see Bold et al. 2017). Unlike many other large-scale data collection efforts, which assess students and teachers using multiple choice items and thus introduce additional chance variation in test scores, the data we use is collected using one-on-one tests for students, and test scores for teachers that are derived from mock student tests marked by the teachers.

Our first contribution is methodological. We show that data on test scores of the student's current and the previous year's teachers, and knowledge of the correlation structure of teacher knowledge across time and grades, allow us to estimate two structural parameters of interest: the contemporaneous effect of teacher content knowledge, and the extent of fade out of teacher impacts in earlier grades.

[1] The estimates for Mozambique, Nigeria, and Uganda are derived using the data we present in this paper. The estimate for India is from ASER (2013).
[2] For a review of the experimental literature, see Kremer, Brannen, and Glennerster (2013), Murnane and Ganimian (2014), Glewwe and Muralidharan (2015), and Evans and Popova (2016).

Our second contribution is empirical. We show that teacher subject content knowledge has a large and significant contemporaneous effect on student performance. Our preferred specification implies that a 1 standard deviation (SD) increase in teacher content knowledge increases student learning by 0.06 SD in the short run (after one year of teaching). This implies that moving a student from a teacher at the 5th to the 95th percentile of the teacher content knowledge distribution increases student learning by 0.2 SD in one year. These effects do not persist fully, however, with about 60 percent of the short-run effect persisting from one grade to the next (an estimate that is similar to those derived from value-added models in other contexts).[3] As a result, being taught for four years by teachers whose content knowledge is 1 SD higher than the average student knowledge would increase by 0.12 SD, equivalent to two and a half times the short-run effect.

The empirical results allow us to assess the relative magnitude of the effects and to carry out various policy simulations. First, we investigate the extent to which shortfalls in teachers' content knowledge account for the large learning gaps observed within and across countries. Second, we ask by how much learning would increase if students were moved from teachers with lower to higher content knowledge, recognizing that this entails a potential cost through larger classes. Finally, we consider the longer run implications of complementary policy reforms aimed at improving teacher content knowledge along with incentives that increase the time devoted to teaching.

Our analysis suggests that raising teacher effective content knowledge to the typical official requirement for primary teachers in Africa (10 years) would, by itself, reduce the observed effective schooling gap after four years by around one-third. Around 20 percent of the gap in learning between the lowest and the highest ranked country in terms of student learning is explained by differences in teachers' content knowledge. The simulation results show that there are potentially substantial cost-savings that could result from assigning more students to better teachers, even if there are negative class-size effects. Last, they show that ensuring that all incoming teachers have the officially prescribed effective years of education (that is, they master the material covered in the officially prescribed years of education), along with increasing the time spent on teaching to the mandated schedule, would almost double student learning within the next 30 years. Either of these reforms alone would have substantially smaller effects.

[3] See Kane and Staiger (2008); Jacob, Lefgren, and Sims (2010); Rothstein (2010); and Andrabi et al. (2011).

The paper is structured as follows. We proceed in Section 2 by providing a short discussion of the literature we build on, and a description of the data we use. In Section 3 we provide summary statistics on student learning outcomes and teacher content knowledge. In Section 4 we present a statistical model of cumulative knowledge acquisition, show how to estimate the key structural parameters, and discuss how we identify these parameters empirically. In Section 5 we present reduced form evidence followed by the main results, show specification and placebo tests, and present the findings from various simulations that use the estimated structural parameters. Finally, we conclude in Section 6 with a short discussion of the implications of our findings.

2. Related literature, data, and context

2.1 Related literature

This paper is related to a growing literature on the impact of teacher quality on student learning outcomes. Several studies, primarily from the U.S but more recently also from other countries such as Ecuador and Pakistan have demonstrated the importance of teachers using a value-added approach, with effect sizes ranging from 0.1 to 0.2 SD (Rockoff 2004; Rivkin et al. 2005; Aaronson et al. 2007; Chetty et al. 2014b; Araujo et al. 2016; and Bau and Das 2017).[4] While data that track individual student achievement over time, and link students to their teachers, is becoming increasingly more common in high-income countries, this is usually not the case in low- and middle-income countries.[5] Therefore, finding alternatives to estimating value-added models that rely on those datasets is necessary in order to provide insights into the education production function.

Our results confirm the importance of teacher quality and demonstrate the external validity of the value-added findings to a low-income environment where both the average quality of teachers is low but also the variation in measured teacher quality is likely much greater.

Our work complements the literature in two additional ways. First, we estimate the causal effect of one key component of teacher quality; namely the extent to which teachers master the knowledge of the subject they are teaching. Second, we propose an alternative empirical approach that allows us to estimate the impact of teacher content knowledge using

[4] The effect size is the impact on standard deviations of student performance from a one standard deviation improvement in teacher value added.
[5] Singh (2019) presents grade- (but not teacher-) value added estimates for Ethiopia, India, Peru and Vietnam. Buhl-Wiggers et al. (2018) present teacher value added estimates in the context of a randomized evaluation in Northern Uganda. They find that a 1 SD increase in teacher effectiveness increases student learning by 0.09 SD.

cross-sectional data. We identify structural impacts building on Dee (2005; 2007) who use within-student across-subject comparisons. In addition, we build on Metzler and Woessmann's (2012) proposal to use contemporaneous within-student within-teacher across-subject variation to identify impacts. This approach, initially applied by Melzer and Woessmann (2012) to Peruvian data, was subsequently implemented by Bietenbeck et al. (2017) and Hanushek et al. (2018) in multi-country settings (in 13 Sub-Saharan African countries spanning years 2000 to 2007 in the former and 31 countries in the latter).

In order to obtain consistent estimates of the impact of teacher content knowledge using a contemporaneous specification one needs to assume either (i) that only contemporaneous teacher's knowledge matters to the production of current achievement or (ii) that there is no correlation between the current teacher's knowledge and past teachers' knowledge (Todd and Wolpin 2003). We relax both assumptions by exploiting the fact that we have data for both the current and previous teacher. As a result, our approach is more in line with the value-added approach where, as a rough approximation, prior achievement is used as a sufficient statistic for the history of prior teacher inputs.

2.2 Data

We use data from the Service Delivery Indicators (SDI)—an ongoing Africa-wide program that aims to collect informative and standardized measures of what primary teachers know, what they do, and what they have to work with. The SDI program—piloted in Tanzania and Senegal in 2010 (Bold et al. 2010, 2011)—grew out of concern about poor learning outcomes observed in various student tests, as well as evident shortcomings in school-level service delivery in fast-expanding education systems.

To date, the SDI program has collected data from a total of seven countries (eight surveys): Kenya (2012), Mozambique (2014), Nigeria (2013), Senegal (2010), Tanzania (2010, 2014), Togo (2013), and Uganda (2013).[6] In each country, representative surveys of between 150 and 760 schools were implemented using a multistage, cluster-sampling design. Primary schools with at least one fourth-grade class formed the sampling frame. The samples were designed to provide representative estimates for teacher effort, knowledge, and skills in public primary schools, broken down by urban and rural location. For four of the six non-pilot

[6] The survey in Nigeria covered only four states (Anambra, Bauci, Ekiti, Niger). The data for Nigeria are therefore not nationally representative, but representative of these four states. For simplicity we nevertheless refer to this dataset as Nigeria.

surveys, representative data were also collected for private primary schools. Across the eight surveys, the SDI collected data on 2,600 schools, over 21,000 teachers and 24,000 students.[7]

The SDI surveys collect a broad set of school, teacher, and student specific information, with an approach that relies as much as possible on direct observation rather than on respondent reports. Data are collected through visual inspections of fourth-grade classrooms and the school premises, direct physical verification of teacher presence by unannounced visits, and teacher and student tests. Bold et al. (2017) document how African teachers perform along three dimensions: Time spent teaching, teachers' knowledge of the subject they are teaching, and teachers' pedagogical skills. Teachers, on average, are absent from class 44 percent of the time and about half of that classroom absence is due to teachers not being at the school during regular teaching hours (Table 1). As a result, while the scheduled teaching time for fourth graders is relatively long—5 hours and 25 minutes—the actual time students are taught is about half that time—2 hours and 46 minutes. Pedagogical knowledge is low, with only one in ten teachers deemed to have minimum pedagogical knowledge, and even fewer teachers are judged to be able to properly assess students' learning shortcomings and progression.

In each school, ten students are sampled from a randomly selected grade 4 classroom. The choice to test students that had completed the third grade was made with the following objectives in mind: on the one hand a desire to assess cognitive skills at young ages when these are most malleable; and on the other hand a desire to assess the learning outcomes of students who have completed at least some years of schooling and to assess language learning at a time when all children would have had lessons in the official language of their country (English in Nigeria and Uganda, English and Swahili in Kenya and Tanzania, French in Senegal and Togo, and Portuguese in Mozambique). In each school, the students' current, and to the extent possible, previous language and mathematics teacher are selected for testing.[8] In more than 80 percent of the schools surveyed in the seven countries, and for 65 percent of the students, data were collected on both the current and previous teachers, i.e., the teachers in grade 4 and in grade 3.

The student test was designed as a one-on-one evaluation, with enumerators reading instructions aloud to students in their mother tongue. This was done to build up a differentiated picture of students' cognitive skills; i.e., oral one-to-one testing allows one to

[7] See Bold et al. (2017) for details of the sample.
[8] In five of eight surveys, teachers in other grades were also sampled.

test whether a child can solve a mathematics problem even when his/her reading ability is so low that he/she would not be able to attempt the problem independently.

The student language test, which evaluated ability in English (Kenya, Nigeria, Tanzania, and Uganda), French (Senegal and Togo), or Portuguese (Mozambique), ranges from simple tasks that tested letter and word recognition to a more challenging reading comprehension test.[9] The mathematics test ranges in difficulty from recognizing and ordering numbers, to the addition of one- to three-digit numbers, to the subtraction of one- and two-digit numbers, and to the multiplication and division of single-digit numbers. In both language and mathematics, the tests span items from the first four years of the curriculum.[10]

In contrast to other approaches to assess teachers' knowledge, where teachers take exams, teachers were asked to mark (or "grade") mock student tests in language and in mathematics. This method of assessment has two potential advantages. First, it aims to assess teachers in a way that is consistent with their regular teaching activities—namely, marking student work. Second, by using a different mode of assessment for teachers compared to students, it recognizes teachers as professionals. In the analysis, we use data on language knowledge of those teachers who teach language, and data on mathematics knowledge of those teachers who teach mathematics.

Both the language and mathematics tests for teachers covered items starting at Grade 1 level (simple spelling or grammar exercises, addition and subtraction) and included items up to the upper primary level (Cloze passages to assess vocabulary and reading comprehension, interpretation of information in a diagram and/or a graph and more advanced math story problem).[11] Both the student and the teacher tests have good reliability, with a reliability ratio (estimated by Cronbach's alpha) above 0.8 in both subjects on the student test and above 0.85 in both subjects on the teacher test.[12]

---

[9] For consistency with the other countries, in Tanzania and Uganda we do not use the part of the samples from those countries in which students were tested in Swahili.

[10] The teacher and student subject tests were designed by experts in international pedagogy and validated against 13 Sub-Saharan African primary curricula (Botswana, Ethiopia, Gambia, Kenya, Madagascar, Mauritius, Namibia, Nigeria, Rwanda, Seychelles, South Africa, Tanzania, and Uganda). See Johnson, Cunningham and Dowling (2012) for details. A few items in the tests also measured grade 5 knowledge.

[11] Cloze sentences/passages leave blank a word in a phrase/sentence and ask the test-taker to fill in that word (sometimes from a set of options) with one that completes the sentence in a way that is sensical and grammatically correct.

[12] Cronbach's alpha is defined as the square of the correlation between the measured test score and the underlying metric. A Cronbach alpha of 1 would indicate that the test is a perfect measure of the underlying metric (though not necessarily of student/teacher knowledge). As a rule of thumb, values between 0.8-0.9 are considered as good.

3. Student and teacher performance

We use two alternative approaches to deriving overall test scores. First, we aggregate the individual items (into the latent underlying factor that drives them) using item response analysis (IRT).[13] The teacher and student tests overlapped in the grade content covered in the tests; i.e. both tests spanned items from the first four years of the curriculum. The teacher test, however, also covered items from higher grades. The method of assessment also differed (students were administered an exam, while teachers were asked to mark mock student tests). For these reasons, the item response analyses were done separately for teachers and students.

Second, we derive a complementary test score measure which we label "effective years of schooling".[14] As with the IRT scores, this complementary test score measure scales the raw scores by difficulty. But unlike the IRT methodology which is essentially a data-driven approach which classifies a question as easy or difficult based on how many test-takers were able to answer it, the effective years of schooling approach classifies a question as easy or difficult based on which grade in the curriculum it supposed to be covered.[15]

This alternative test score measure has three advantages. First, the transformed data points are informative in-and-of themselves since they situate performance along the distribution of curriculum expectations at each grade.[16] Second, the transformation allows us compare students and teachers using the same scale, although it is important to keep in mind that the method of assessment differed (enumerator-administered for students versus self-

[13] Item response theory is a method to estimate a respondent's underlying ability/latent trait based on their answers to a series of items, in our case an estimate of the student's (teacher's) knowledge based on the pattern of correct/incorrect questions on the test. To do so, IRT specifies a parametric model for the probability of a correct answer given the test-takers latent trait and properties of the item. While models vary in the precise parameterization, they generally share the following features: the probability of a correct answer is decreasing in the difficulty of an item and increasing in the ability/latent trait of the test-taker (see Jacob and Rothstein, 2016). To estimate the item parameters for the student test, we specify a 2-parameter logistic model which describes each item by its difficulty and the extent to which it discriminates between students of different ability. To the teacher test, we apply a partial credit model, which allows for items that are scored on an ordinal (but not necessarily binary) scale. Given the estimated item parameters and patterns of correct/incorrect answers, we can then construct a measure of the underlying student and teacher subject knowledge.

[14] For teachers, we will refer to the measure as `effective years of education' to emphasize that their education is completed at the time at which we assess them.

[15] Both the teacher and student subject tests were validated against a large set of Sub-Saharan African primary curricula (see footnote 6). Thus, all items in the student tests covered items in the first four grades in the countries surveyed. We use the Kenyan curriculum to link test items to specific grade levels. This choice should be kept in mind when making cross-country comparisons using this outcome measure, as there might be some (albeit likely small) variations across countries in which grade each subject item was introduced. Mitigating this potential issue, however, is the fact that our empirical specification uses variation across students to identify the structural estimates.

[16] We should note, however, that the tests were designed to maximize the precision over a smaller range of abilities in each grade, rather than assess knowledge at each grade levels; the measure should therefore be viewed as a proxy measure of students/teachers grade level competencies.

administered for teachers). Finally, it allows us to extrapolate beyond the effective years of schooling observed in the sample in a meaningful way (an advantage we exploit for the policy simulations in section 5).

Table 2 (for students) and Table 3 (for teachers) show how the scores constructed with item response theory correspond to the curriculum-scaled years of schooling measure. There is a strong positive correlation between the two scores (the correlation coefficient is above 0.9 in both cases). Table 2, and Figure 1, further show the distribution of students across each effective year of schooling. On average, students have 1.5 effective years of schooling in language and mathematics after three and half years of studies. That is, the median mathematics student, after completing approximately three and half years of schooling, does not master the second-grade curriculum in mathematics. A quarter of fourth grade students have not acquired any effective years of schooling in language and one-third have not acquired any effective years of schooling in mathematics; 15 percent and 21 percent have acquired one effective year of schooling in language and mathematics, respectively; less than a quarter of the students have three years or more of effective years of schooling in language and mathematics. Comparing across countries for mathematics, the average student in Kenya (the top performer) has acquired 2.5 effective years of schooling after three and a half years of schooling, while the average student in Mozambique (the bottom performer) has acquired only 0.34 years of effective years of schooling (see also Figure 2).

Table 3 and Figure 3 show the distribution of teachers across each effective year of education.[17] On average, teachers have 3.8 years of effective years of education in language and 4 years mathematics.[18] A large share of teachers barely master the curriculum that they are expected to teach: 47 percent in language and 52 percent in mathematics have no more than 3 effective years of education. Like students, there are large differences across countries in teacher performance: Kenyan mathematics teachers (the top performers) have on average

[17] We here include all the teachers sampled, which includes a number of teachers in grades other than grade 3 and 4 (17% of the sample), in both public and private schools (19% of the schools). Restricting the sample to either teachers in grade 3 and 4, or in public schools, or both, reduces average effective years of education by about 10%.

[18] While these numbers are low, they are consistent with the alternative measures of teacher knowledge presented in Bold et al. (2017), who calculate that two thirds of teachers across Sub-Saharan Africa have subject knowledge equivalent to a fourth grader, defined as mastering 80% of the material covering grade 1 to 4 on the test. Our definition here is in one respect more stringent since a teacher has to score, for example, all grade 4 questions correctly in order to be categorized as having grade 4 knowledge (rather than 80% of questions covering grade 1-4). On the other hand, we do not require, again using grade 4 as an example, that the teacher also manages all tasks covering grade 1-3 in order to be categorized as having grade 4 knowledge. Adopting a slightly more lenient approach that allows for some margin of error at each grade level increases teachers' average effective years of education to 4.6 years. Nevertheless, we use the stricter measure here as it makes for a more transparent definition of the teacher test score.

5.7 effective years of education while mathematics teachers in Togo have only 2 effective years of education (see Figure 4).

4. Conceptual framework and methods

In this section, we first present a simple statistical model of students' learning achievement in language and mathematics as a function of their teachers' (content) knowledge in these subjects. We then discuss our core identifying assumptions and show that with data on test scores of the student's current and previous year's teachers, and knowledge of the correlation structure of teacher knowledge across time and grades in a subset of schools, we can estimate two structural parameters of interest: the contemporaneous effect of teacher content knowledge and the extent of fade out of the teachers' impact in earlier grades. Finally, we discuss how we make inferences about the parameters of interest.

4.1. Statistical model

Consider the outcomes of a student i who is in grade g at time t. Each student is taught two subjects, k and k'. Let $j = j(i,g,t)$ denote student i's teacher in grade g and time (year) t and $x_{j(i,g,t)k}$ teacher j's (content) knowledge of subject k. Assuming for now that that $g = t$, i.e. that grade level and time spent in school are the same, we can drop the time subscript t and write student i's test score in grade 4 as

$$y_{i4k} = \beta_{0k} + \sum_g^4 \beta_g x_{j(i,g)k} + \nu_{i4k} \quad (1)$$

where

$$\nu_{i4k} = \sum_g^4 \mu_{j(i,g)} + \eta_i + \tau_{j(i,4),k} + \varepsilon_{i4,k} \quad (2)$$

The error term ν_{i4k} is decomposed into four components: teacher-specific terms $\mu_{j(i,g)}$, accounting for all teacher-specific factors that vary across teachers but are constant across subjects for a given teacher, student fixed effects, η_i, accounting for all student-specific subject-invariant variation (including innate ability, family characteristics and other parental-supplied or school-supplied inputs), and idiosyncratic teacher- and student-level variation, $\tau_{j(i,4),k}$ and $\varepsilon_{i4,k}$.

In (1), children's achievement, as measured by test performance in grade g, is the outcome of a cumulative process of knowledge acquisition.[19] The reduced form coefficients

[19] In the cumulative student achievement function (1), $\beta_{g,k} = \beta_{g,k'} = \beta_g$; i.e., a one unit increase in teacher content knowledge, properly normalized, has the same marginal effect on student test scores in both subjects. In

β_g thus capture both a direct treatment effect, which we label the contemporaneous effect, α_g, and the link between achievement across periods. We capture this second effect by allowing for learning to fade out over time. Specifically, $\gamma_{g,g'} \leq 1$ captures the degree of persistence of teaching that took place in grade g' but was measured in grade g, with $g > g'$. That is, if α_{g-1} is the marginal effect of teacher content knowledge in grade $g-1$ on student achivement at the end of the same grade, then $\gamma_{g,g-1}\alpha_{g-1}$ is the marginal effect of teacher content knowledge in grade $g-1$ on student achivement at the end of grade g. With these assumptions, we can define the cumulative effect of teacher knowledge on student learning after four years as $CE = \alpha_4 + \sum_{g=1}^{3} \gamma_{4,g}\alpha_g$ and rewrite (1) as

$$y_{i4k} = \alpha_0 + \alpha_4 x_{j(i,4)k} + \alpha_3\gamma_{4,3}x_{j(i,3)k} + \alpha_2\gamma_{4,2}x_{j(i,2)k} + \alpha_1\gamma_{4,1}x_{j(i,1)k} + \nu_{i4k} \quad (3)$$

Equation (3) illustrates the three challenges we face in estimating the cumulative effect (CE), given the data we have. First, we have a cumulative model of learning, but teacher knowledge and student learning are measured at one point in time. Second, student achievement in grade (or year) four is a function of the whole history of the content knowledge of the students' teachers; i.e., x_4, $\ldots x_1$, but we observe the content knowledge of student i's teachers only in year 3 and 4. Third, students' innate ability and several school and parent-supplied inputs are inherently unobservable and may be correlated with x_j if, for instance, better students sort into schools with better teachers.

We now discuss three assumptions under which we can nevertheless estimate the (causal) cumulative effect of teacher content knowledge on student learning, as well as the two key structural parameters of interest (the contemporaneous effect and the degree of persistence).

Assumption 1: Measurement

A) Teacher content knowledge is time-invariant; i.e., $x_{j(t)} = x_j$

Assumption 2: Model restrictions

A) The contemporaneous effect is independent of the grade; i.e., $\alpha_g = \alpha_{g-n} = \alpha$

B) The effect of teacher content knowledge declines (geometrically) with distance.

Assumption 3: Identification

section 5, we test this model restriction, following Ashenfelter and Zimmerman (1997). We fail to reject the null of equality of the coefficients across the two subjects.

A) Teacher content knowledge in subject k is uncorrelated with all unobservables conditional on controlling for student and teacher invariant heterogeneity; i.e.;

$$\text{cov}\left(x_{j(i,g)k}, \varepsilon_{i,g,k} | \mu_{j(i,1)k}, \dots, \mu_{j(i,g)k}, \eta_i\right) = 0$$

$$\text{cov}\left(x_{j(i,g)k}, \tau_{j(i,g)k} | \mu_{j(i,1)k}, \dots, \mu_{j(i,g)k}, \eta_i\right) = 0$$

B) If a student is taught by a general teacher, who teaches both subjects k and *k' in primary grade g, then she is taught by a general teacher in grades* $g' < g$.

Assumption 1 implies that teacher content knowledge does not have to be measured at the same time at which it is applied; for example, measuring content knowledge of a student's grade 3 teacher at the time when the student is in grade 4 gives an accurate measure of the teacher knowledge the student would have been exposed to in the prior year. Our data can therefore be used to estimate a cumulative production function.[20]

Assumption 2 allows us to simplify the production function to:

(4) $$y_{i4k} = \alpha_0 + \alpha x_{j(i,4)k} + \alpha\gamma x_{j(i,3)k} + \alpha\gamma^2 x_{j(i,2)k} + \alpha\gamma^3 x_{j(i,1)k} + \nu_{i4k}$$

Reducing the number of structural parameters (from seven to two) is a necessary condition for them to be identifiable from the reduced form effects of teacher knowledge in grade 3 and 4 on student learning in grade 4.[21]

Assumption 3A is our core identifying assumption. It allows for teacher knowledge to be correlated with student unobservables (such as the child's aptitude for learning or parental support), for instance because better students sort into better performing schools, but it rules out that students systematically sort based on subject-specific abilities into schools with subject-specific teacher knowledge. That is, the assumption would be violated, if, for example, students in lower primary with relatively higher motivation for mathematics systematically sort into schools (or classrooms) with relatively more knowledgeable mathematics teachers.

Assumption 3A also places some restrictions on what parents and schools do. For example, while our identifying assumption does allow for parents (or schools) to respond to their children's low mathematics aptitude by providing additional teaching (or hiring a private tutor), they cannot do this to compensate for insufficient teacher mathematics knowledge.[22]

---

[20] This assumption is also frequently made in the teacher-value added literature (see Kane and Staiger, 2008). Chetty et al. (2008) show that estimates of teacher value added with and without an assumption of time-invariance give very similar results.

[21] As shown in Todd and Wolpin (2003), assumption 1 is also required in order to derive the lagged-score value-added model from a linearized cumulative student achievement function.

[22] Using data from kindergarten students in Ecuador, Araujo et al. (2016) find that while parents recognize better teachers, they do not change their behaviors to take account of differences in teacher quality. Note that our

More generally, while differential supply of school and parental inputs across subjects may occur, and may be correlated with various school and student characteristics, our maintained assumption is that these differential input flows are uncorrelated with the variation in teacher content knowledge across subjects.

Finally, assumption 3A also allows for teacher qualities/behaviors other than content knowledge to matter for student performance. Teachers with higher knowledge scores may, for example, have better pedagogical skills. However, the assumption rules out that class teachers systematically spend more effort teaching the subject they master relatively better, or likewise systematically put more effort into teaching the subject of which they are less knowledgeable.

Together these assumptions provide necessary conditions for the causal effect of teacher knowledge on student learning to be identified from two sources of available information: (i) within-student across-subject variation in learning in grade 4 and within-teachers across-subject variation in knowledge of the student's grade 3 and grade 4 teacher, and (ii) information on how teachers' knowledge in grade 3 and 4 is correlated with teachers' knowledge in earlier grades.

In practice, we can remove subject-invariant variation for grade 4 students and their current and previous teachers either by introducing a set of fixed effects (for students and for teachers teaching students in grade 3 and 4) or by taking the first-difference across subjects and restricting the sample to students who were taught by general teachers in grade 3 and 4, which removes the student's and teachers' subject-invariant terms, η_i and $\mu_{j(i,4)}$ and $\mu_{j(i,3)}$ from equation (3).

Since we do not know the students' teacher in grade 1 and 2, however, neither of these transformations is guaranteed to remove subject-invariant variation for the teachers in grades 1 and 2. This necessitates Assumption 3B, which is used to "impute" information on whether a student was taught by general teachers in grades 1 and 2 based on whether they were in grades 3 and 4.

Assumption 3B is grounded in the empirical context of primary schooling in Sub-Saharan Africa. Specifically, primary school students tend to have a general teacher who teaches both language and mathematics in lower primary (grades 1-3) while subject teachers, who specialize in either language or mathematics, become progressively more common as

identifying assumption here is even weaker. We assume parents do not respond to differential (across subjects) differences in the quality of the teacher.

students move to upper primary and secondary. In other words, if a student is taught by a general teacher in both mathematics and language in grade g, it is very likely that the student was also taught by a general teacher in $g-1$, while the opposite transition seldom happens.[23] In this case, fixed effects for the grade 3 and 4 teacher (or an equivalent sample restriction) remove all teacher subject-invariant variation also in grades 1 and 2.

We provide further discussion of these assumptions in Section 4.5.

4.2. Empirical implementation

Assumptions 1-3 are sufficient to identify the structural parameters from the available data. Our main estimation equation for the reduced-form is the following first-difference specification:

$$\Delta y_{i4} = y_{ij4,k} - y_{ij4,k'} = \beta_0 + \beta_4 \Delta x_{j(i)4} + \beta_3 \Delta x_{j(i)3} + \Delta \xi_{i4} \,, \tag{5}$$

where $\Delta \xi_{i4} = \Delta \varepsilon_{i4} + \Delta \tau_{j(i,4)} + \alpha\gamma^2 \Delta x_{j(i,2)} + \alpha\gamma^3 \Delta x_{j(i,1)}$.

Plugging the structural equation (4) into the OLS formulas for $\hat{\beta}_4$ and $\hat{\beta}_3$, we can express the reduced form coefficients in terms of the structural parameters of interest (see Appendix B for details). In this and all following derivations, we impose that the variance of teacher content knowledge is independent of the grade in which the teacher is deployed, i.e. $\text{var}(x_g) = \text{var}(x_{g'}) = \sigma_x^2$. We do this both to simplify the algebra and to easily place bounds on the reduced form bias.[24] This results in the following expressions $\hat{\beta}_4$ and $\hat{\beta}_3$:

$$\text{plim}\, \hat{\beta}_4 = \alpha + \alpha\gamma^2 \left(\frac{\rho_{4,2} - \rho_{3,2}\rho}{1-\rho^2}\right) + \alpha\gamma^3 \left(\frac{\rho_{4,1} - \rho_{3,1}\rho}{1-\rho^2}\right), \tag{6}$$

$$\text{plim}\, \hat{\beta}_3 = \alpha\gamma + \alpha\gamma^2 \left(\frac{\rho_{3,2} - \rho_{4,2}\rho}{1-\rho^2}\right) + \alpha\gamma^3 \left(\frac{\rho_{3,1} - \rho_{4,1}\rho}{1-\rho^2}\right), \tag{7}$$

where $\rho_{g,g'}$ measures the correlation between the regressors, namely the knowledge of student i's grade g teacher ($g = 3, 4$), and the omitted variables, namely the knowledge of the student i's grade g' teacher ($g' = 1, 2$). ρ is the correlation in knowledge between student i's grade 3 and grade 4 teacher (that is, $\rho = \rho_{4,3} = \rho_{3,4}$).

[23] Looking at the transition from lower primary (grade 3) to upper primary (grade 4) provides indirect support of the assumption: while 60% of the students taught by a class teacher in grade 3 are taught by subject-specific teachers in grade 4, more than 90% of the students that are taught by a class teacher in grade 4 also had a class teacher in grade 3.

[24] The assumption could be easily dispensed with. However, since it holds in our sample (see Table A4) we maintain it for notational and analytical convenience. With unequal variance of test scores across grades, equations (6) and (7) depend on the linear projection coefficients of regressing $\Delta x_{g'}$ on Δx_g, which are not bounded a priori and may differ according to the direction of the regression. With equal variances, equations (6) and (7) depend on the correlation coefficient, which are symmetric and less than 1 in absolute value.

We are now left with two equations, (6) and (7), which relate the reduced form coefficients β_4 and β_3 to seven unknowns ($\alpha, \gamma, \rho, \rho_{4,2}, \rho_{3,2}, \rho_{4,1}, \rho_{3,1}$). In order to recover the point estimates for the structural parameters, α, γ, we therefore need to know how teacher knowledge is correlated across grades.

For the grade 3 and 4 teachers, this is simply found from the variance-covariance matrix of the regressors in equation (5). To estimate the correlations between the included (Δx_4,Δx_3) and the omitted (Δx_2, Δx_1) teacher content knowledge (i.e. $\rho_{4,2}, \rho_{3,2}, \rho_{4,1}, \rho_{3,1}$), we exploit the fact that in a subset of schools we measure the test scores of a sample of teachers in grade 1 through 4 and their transition patterns across lower primary both in the current and previous year.

We estimate the correlation coefficient $\rho_{g,g'}$ with $g = 3, 4$ and $g' = 1, 2$, separately on two sub-samples: (i) the sub-sample where the student's grade g teacher taught the student already in grade g', and (ii) the sub-sample where the student's grade g teacher did *not* teach the student in grade g'. We then construct $\rho_{g,g'}$ as a weighted average of the correlation coefficients in the two sub-samples.

Observations are assigned to the relevant sub-sample based on the observed transition patterns of teachers in lower primary school. In particular, teachers can be divided into three groups: (a) teachers who have taught grade g both in the current and previous year are categorized as "grade teachers," who teach the same grade each year, (b) teachers who have taught grade g in the current year and grade $g - 1$ in the previous year are categorized as "class teachers," who cycle/transition with their class through lower primary, and (c) teachers who do not fit either of these patterns are labelled as "other". We denote the share of these teachers in grade $g = 3, 4$ as $s_{class,g}$, $s_{grade,g}$ and $s_{other,g}$.

From this classification, all class teachers in grade g are assigned to the first sample, since they have already taught student i in grade g' and all grade teachers in grade g are assigned to the second sample, since they have not taught student i in grade g'. Finally, for grade g teachers classified as `other', we assume that they are randomly allocated to one of the streams in lower primary each year and the share of such teachers who have already taught student i in grade g' is therefore inversely proportional to the (average) number of streams in lower primary, $\bar{n}_{streams}$.[25]

[25] The median number of streams per grade is 1 and the mean number is 1.5 in our sample. Note that we assume that class teachers do not switch streams between grades.

The values of $\rho_{g,g'}$ in the two sub-samples are then calculated as follows. In the first sub-sample, which consists of $s_g = s_{class,g} + \frac{1}{\bar{n}_{streams}} s_{other,g}$ teachers, the correlation between included and omitted test scores in the regression is simply 1.

In the second sub-sample, which consists of $(1 - s_g) = s_{grade,g} + \left(1 - \frac{1}{\bar{n}_{streams}}\right) s_{other,g}$ teachers, we need to calculate, $\rho_{g,g'|j(g)\neq j(g')}$ the test score correlation with the teacher who $4 - g'$ years ago taught student i in grade g'. Though in general we do not have this information, we can infer it for a subset schools: those where the grade g' teacher is a grade teacher and that have only one stream per grade. We could therefore estimate $\rho_{g,g'|j(g)\neq j(g')}$ in this subset. There are, however, two obvious concerns: (i) this subset of schools and teachers may not be representative of the second sub-sample, (ii) in general, there are only few pairs of teachers for each particular combination of grade g and g' and this is even more the case when we only consider certain categories of teachers and schools—these issues would lead to imprecise estimates. To alleviate these concerns, we show in the Appendix (Section C and Table A2), that the correlation in teacher knowledge between any two lower primary school teachers in a school is independent of the type of teacher (i.e. class, grade, other), grades (i.e. 1 through 4), school (i.e. one/multiple streams per grade), and dates (i.e. current or previous year) at which they teach. It therefore shouldn't matter whether we estimate $\rho_{g,g'|j(g)\neq j(g')}$ on all pairs of (distinct) teachers in a school or on a particular subset.[26,27]

Third, we estimate $\rho_{g,g'}$ as a weighted average of the correlations in these two sub-samples:[28]

$$\rho_{g,g'} \approx s_g \times 1 + (1 - s_g) \times \rho_{g,g'|j(g)\neq j(g')} \quad (8)$$

[26] In what follows, we estimate $\rho_{g,g'|j(g)\neq j(g')}$ for pairs where the grade g' is a grade teacher in schools that have only one stream per grade. The results using pairs of all teachers are virtually the same.

[27] The above argument suggests that we could match students to their teachers in grade 1 and 2 on the basis of these transition patterns and then estimate (5) with the full history of teacher knowledge. However, we have too few observations on teachers in lower grades to make this feasible. Specifically, in order to run such a regression, we need complete data on lower primary teachers in each school to calculate the 4-by-4 variance-covariance matrix, but even just imputing the knowledge of the grade 2 teacher reduces the sample to 280 students in 36 schools. In contrast, each pairwise correlation $\rho_{g,g'|j(g)\neq j(g')}$, can be estimated simply by creating pairs of grade g and grade g' teachers in each school, or, for that matter by creating pairs of all teachers who are not the same in each school, (since $\rho_{g,g'|j(g)\neq j(g')}$ was found not to depend on date, grade, teacher or school type) without being constrained by data availability for teachers in grades other than g and g'.

[28] This holds with equality if the first two moments of the test score distribution are the same in the two sub-samples.

With estimates for β_4, β_3, ρ, $\rho_{4,2}$, $\rho_{3,2}$, $\rho_{4,1}$, $\rho_{3,1}$ we can uncover α and γ from equation (5) and (6) giving us the contemporaneous effect of teacher knowledge on student learning, its persistence, and, by combining them, the cumulative effect.

4.3. A lower bound on the cumulative effect

An alternative approach to the estimation procedure described in the previous section is to compute the lower bound of the cumulative effect of teacher content knowledge. Being able to estimate such a lower bound is helpful when information on teacher knowledge in earlier grades is either not available at all, or when it is only partial. In particular, we show that with mild restrictions on the correlation matrix of teacher knowledge across grades we can estimate the lower bound of the cumulative effect of teacher knowledge on student learning from the sum of the reduced form coefficients.

Specifically, sufficient conditions (see Section B of the Appendix) for

$$(10) \qquad \hat{\beta}_4 + \hat{\beta}_3 = \alpha + \alpha\gamma + \alpha\gamma^2\left(\frac{\rho_{4,2}+\rho_{3,2}}{1+\rho}\right) + \alpha\gamma^3\left(\frac{\rho_{3,1}+\rho_{4,1}}{1+\rho}\right) \leq \alpha_4 + \sum_{t=1}^{3}\alpha_t\gamma_{t,4}$$

are: (i) the correlation in teacher content knowledge between any two grades is positive, $\rho_{g,g'} \geq 0$ for all g and g'; (ii) the further apart any two grades, the lower is the correlation in teacher content knowledge across grades; i.e., $\rho_{g,g'}$ is decreasing in $|g - g'|$. These conditions seem reasonable given the pattern of transitions of teachers across grades in primary school documented above, coupled with the assumption that these patterns are the main (though not necessarily exclusive) drivers of correlations across grades. In particular, if class teachers are prevalent in each grade but become less frequent in higher grades, both of which is the case in our data, then both conditions are likely to hold.[29]

4.4. Comparing the cumulative and the contemporaneous specifications.

In this subsection we compare our estimation procedure to the contemporaneous specification, which regresses subject differences in student knowledge on teacher knowledge in the current grade.

$$(11) \qquad \Delta y_{i4} = \beta_0 + \beta_4\Delta x_{j(i)4} + \Delta\xi_{i4} \,,$$

where $\Delta\xi_{i4} = \Delta\varepsilon_{i4} + \Delta\tau_{j(i,4)} + \alpha\gamma\Delta x_{j(i,3)} + \alpha\gamma^2\Delta x_{j(i,2)} + \alpha\gamma^3\Delta x_{j(i,1)}$.

Under identical assumptions to the ones made here (i.e. Assumption 3A & 3B), such a specification would identify a causal effect of teacher knowledge on student learning. The magnitude of this effect, however, is uninformative about the actual parameters of the

[29] We show in Section C of the Appendix that these restrictions hold in our data set.

education production function and cannot therefore be used for policy analysis. Specifically, the coefficient on current teacher knowledge in the contemporaneous specification measures either α_4 or the cumulative effect of teacher knowledge on learning $\alpha_4 + \sum_{t=1}^{3} \alpha_t \gamma_{t,4}$, or anything in between. It identifies α_4 under the strong assumption that there is no correlation between included and omitted teacher test scores. It identifies the cumulative effect if the grade 4 teacher has taught the student throughout primary. If neither of these scenarios applies, it is not possible to interpret the size of the estimated coefficient, since, even with information on the correlation between included and omitted teacher knowledge, one cannot separately identify the two crucial parameters, α and γ, needed for policy analysis.

Using the respective reduced form coefficients in the cumulative (5) and the contemporaneous specification (11), and imposing the same restrictions as in the previous sub-section (namely that $\rho_{g,g'} \geq 0$ and that $\rho_{g,g'}$ is decreasing in $|g - g'|$), it is trivial to show that the asymptotic bias in $\hat{\beta}_4$ as an estimate of the contemporaneous effect of teacher knowledge on student learning will always be smaller in our cumulative specification than in a contemporaneous specification. Perhaps more interestingly, the same assumptions also guarantee that the asymptotic bias on the reduced form coefficients $\hat{\beta}_4$ and $\hat{\beta}_3$ is not symmetric but loads mainly on the latter (see Appendix C). Together, this implies an improved accuracy in $\hat{\beta}_4$ as an estimate of α in a cumulative specification with two years of teacher test scores relative to a contemporaneous specification.

4.5. Discussion of assumptions necessary for identification

While untestable, we deem our core identifying assumption (3A) plausible in the context of African primary schools. In support of this assertion, section 5 provides estimates from various versions of the empirical model (5), including controlling for region-by-subjects fixed effects. This allow us to examine whether differences between teacher language and mathematics scores might be driven by a common underlying factor that also affects student subject differences. We also conduct placebo tests using teachers in higher grades and conditioning on length of exposure. The results rule out sorting as the (exclusive) driver of the relationship between teacher and student knowledge. Moreover, we can show that including other subject-invariant measures of teacher knowledge does not change the coefficient on teacher knowledge. These findings provide us confidence that our results can be interpreted as the effect of teacher knowledge on student learning.

One potential concern in our analysis is the fact that test scores are a noisy measure of the latent variable of interest, namely student and teacher knowledge. We refrain from

adjusting for this and simply remark that, in our context, classical measurement error attenuates the reduced form coefficients $\hat{\beta}_4$ and $\hat{\beta}_3$ as well as the resulting structural estimate.[30] Consequently, the estimates in Section 5.3 should be interpreted as a lower bound for the contemporaneous and the cumulative effect of teacher knowledge on student learning.

4.6. Inference

To make inference about the structural parameters of interest in equation (4), $\widehat{\boldsymbol{\theta}} = \{\hat{\alpha}, \widehat{\alpha\gamma}, \widehat{\alpha\gamma^2}, \widehat{\alpha\gamma^3}\}$, we need to estimate their standard errors. The asymptotic variance-covariance matrix is given by $V = \sigma_\epsilon^2 (E(\Delta\tilde{x}'\Delta\tilde{x}))^{-1}/N$, where $\Delta\tilde{x} = \{\Delta\tilde{x}_4, \Delta\tilde{x}_3, \Delta\tilde{x}_2, \Delta\tilde{x}_1\}$ is the matrix of demeaned test score subject differences during grades 1 to 4.

We have shown how to estimate the correlation matrix of included and excluded regressors in Section 4.2 and the variance-covariance matrix, $E(\Delta\tilde{x}'\Delta\tilde{x})$, can be obtained analogously. In estimating σ_ϵ^2, the error variance in the structural regression (4), we again face the issue that we do not have linked teacher and student data in grades 1 and 2. In Appendix D, we show that σ_ϵ^2 can be estimated from the reduced form error variance σ_ξ^2 and knowledge of the variance-covariance matrix of the included and excluded regressors, $E(\Delta\tilde{x}'\Delta\tilde{x})$.[31]

5. Results and policy simulations

5.1. Reduced form relationship between student and teacher content knowledge

We start by providing empirical support for two of the assumptions we rely on to derive structural parameters that have a causal interpretation from our core empirical specification (4), namely the assumption that the coefficients on all (teacher knowledge) inputs are subject-invariant and the assumption that the test score variance is homoscedastic (across grades).

---

[30] Ignoring omitted variable bias and assuming that test scores are measured with classical measurement error, so that $\Delta x_g = \Delta x_g^* + \Delta v_g$, $\Delta v_g \sim N(0, \sigma_{\Delta\xi}^2)$ for $g = 3,4$, and Δv_4 and Δv_3 independent for any two teachers who are not the same, the asymptotic expressions for $\hat{\beta}_4$ and $\hat{\beta}_3$ are: plim $\hat{\beta}_4 = \alpha - \frac{\alpha(1-\gamma\rho)}{(1-\rho^2)}\frac{\sigma_{\Delta v}^2}{\sigma_{\Delta x}^2} - \frac{\alpha(1-s_{43})(\gamma-\rho)}{(1-\rho^2)}\frac{\sigma_{\Delta v}^2}{\sigma_{\Delta x}^2}$ and plim $\hat{\beta}_3 = \alpha\gamma - \frac{\alpha(\gamma-\rho)}{(1-\rho^2)}\frac{\sigma_{\Delta v}^2}{\sigma_{\Delta x}^2} - \frac{\alpha(1-\gamma\rho)(1-s_{43})}{(1-\rho^2)}\frac{\sigma_{\Delta v}^2}{\sigma_{\Delta x}^2}$, where s_{43} is the share of teachers who switch between grade 3 and 4. $\hat{\beta}_4$ will be attenuated as long as persistence is between 0 and 1, while $\hat{\beta}_3$ is attenuated unless persistence is very small (much smaller than ρ), which is not the case in our data.

[31] Note that applying the delta method to the non-linear functions of $\hat{\beta}_4$ and $\hat{\beta}_3$ is not possible here. First, there are no easy closed form solutions for the parameters of interest in terms of the reduced form coefficients. Second, the delta method would not provide the full variance-covariance matrix (only its diagonal) that is needed to test hypotheses about the total cumulative effect.

To assess the empirical validity of the first assumption, we follow Ashenfelter and Zimmerman (1997) and test the restriction of subject-invariant effects by rewriting our main specification (5) as a correlated random effects model. As discussed in Appendix D, and reported in Table A3, we cannot reject the hypothesis that the effect of teacher content knowledge is the same in the two subjects, thus providing support for the first differenced, or fixed effect specification.

Table A4 in the appendix reports summary statistics on the subject differences in content knowledge of teachers in grade 1 through 4 and performs tests on the equality of standard deviations (variances) of the two distributions. As is evident, we cannot reject that the variance of the test score distribution is constant across grades, hence justifying our choice to write the structural parameters of interest as a function of the correlation coefficients ρ and $\rho_{g,g'}$ and to derive coefficient bounds based on this.

In Table 4, we begin to explore the relationship between teacher and student knowledge. We start in column (1) by regressing student achievement on teacher subject knowledge, controlling only for a set of country fixed effects. There is a large positive association. In column (2) we introduce student fixed effects, and in specification (3) we also introduce teacher fixed effects for the current teacher by restricting the sample to students who were taught by class teachers in grade 4. This specification thus controls for sorting of students to schools (or teachers) based on subject invariant characteristics, as well as other unobserved student and teacher subject invariant, characteristics. In this specification, the effect of teacher knowledge on student test scores can only be driven by differences between the two subjects. The results suggest that part of the association in column (1) is driven by better students sorting into better schools.[32]

In column (4) we include past teacher knowledge, resulting in a fall in the estimated coefficient on current teacher knowledge by about 30 percent.[33] We also report the effects of the sum of the teacher content knowledge estimates, which provides a lower bound on the cumulative effect. The estimated cumulative effect implies that being taught by a teacher with a 1 SD higher content knowledge throughout the first four years of schooling (or 3.5 to be more precise) increases student learning by at least 0.10 standard deviations.

[32] Note that the fixed effects specification tends to inflate existing measurement error, so the smaller effect size could also (partly) be a consequence of the decreased signal to noise ratio in this specification.
[33] The reduction is estimated by comparing the point estimates on the current teacher knowledge with and without past teacher knowledge as an additional regressor, holding the sample constant; i.e., using the sample with information on both current and past teachers in both specifications.

Finally, in column (5), our preferred specification, we also introduce teacher fixed effects for both the current and previous teacher by restricting the sample to students who were taught by class teachers in each of these years. With student and teacher fixed effects, the magnitude of the relative effect of the current and prior teacher knowledge change, although the change in the estimated lower bound; i.e., the sum of the two content knowledge coefficients, is much smaller (the lower bound on the cumulative effect falls by approximately 16 percent, from 0.099 to 0.083).

5.2. Robustness checks of the reduced form estimates

We have argued that, under mild conditions, knowledge of the teachers' current and previous test scores in grade 3 and 4 is sufficient to estimate a lower bound for the cumulative effect, and also, with additional information on the correlation pattern of teacher knowledge across grades, point estimates for the contemporaneous effect of teacher content knowledge on student achievement, as well as the extent of fade out of the teachers' impact in earlier grades. However, for these estimates to be interpreted as causal, the estimated reduced form coefficients of current and previous teachers content knowledge in Table 4, column (5), must not be biased by omitted variables (other than grade 1 and grade 2 teacher knowledge) or sorting.

To recap the assumptions stated in Section 4, there are two conditions for identification: (i) there must not be other factors (at teacher level or otherwise) that drive both student and teacher subject differences in knowledge; (ii) there is no sorting by students and teachers based on subject differences. In other words, students that are better in language than in mathematics are not systematically more likely to select into schools where teachers are better in language than in mathematics (or vice versa).

While we cannot unambiguously rule out either of these concerns, we present additional evidence in Table 5 and 6 to support the identification assumptions. In column (1) of Table 5, we repeat our main specification using first differences across subjects and restricting the sample to students who have the same teacher in both subjects in grade 3 and 4. In columns (2)-(4), we then examine whether differences between teacher language and mathematics scores might be driven by a common underlying factor that also affects student subject differences. For example, it might be the case that language knowledge of both students and teachers varies systematically across contexts, such as districts, or urban and rural areas, simply because of differences in the prevalence of the official language. To assess this, we include district (column 2) and urban/rural dummies (column 3). The estimates, including the lower bound for the

cumulative effect, change only marginally after the introduction of these additional controls (compared to the main specification reported in column 1).[34]

Similarly, other teacher behaviors and skills that vary by subject might be correlated with teacher knowledge and therefore affect learning. While we do not have any measure of teacher behaviors that vary across subjects for a given teacher, we can test directly how teacher subject knowledge correlates with other teacher skills and behaviors if we also include students taught by different teachers in language and mathematics in the sample. In column (4) of Table 5, we add a measure of teachers' pedagogy knowledge as an additional explanatory variable to the student fixed effects specification reported in column (4) of Table 4.[35] Pedagogy knowledge has a positive and statistically significant effect on student learning. However, the coefficients of interest remain significant and change little in magnitude relative to the estimates without teacher pedagogy (compare the fourth column in Tables 4 and 5). The lower bound on the cumulative effect is only marginally affected by the inclusion of pedagogy knowledge (0.099 versus 0.094 of a standard deviation). Hence, we would argue that unmeasured differences in teacher skills—at least pedagogical skills—across subjects are unlikely to confound the coefficient of interest.

To further test for sorting across (and within) schools we also report the results of a specifications in which we constrain the sample to schools with only one classroom (column 5); thus effectively ruling out sorting, or tracking, into different classes within schools. While the estimated coefficient on current teacher knowledge falls slightly and the estimate on previous teacher knowledge increases somewhat, the lower bound on the cumulative effect remains largely unchanged (at 0.086 of a standard deviation).

To support our causal interpretation of results, Table 6 presents a test of the identifying assumptions in line with Chetty et al. (2014a) and Rothstein (2010) using test scores of teachers in higher grades as a placebo. If there is a purely causal relationship between teacher knowledge and student achievement, then including the test scores of teachers in the same school who have not (yet) taught the student should not change the coefficients on current or previous teacher knowledge, and should in itself have no significant impact on student test scores. The coefficients on current and prior teachers' content knowledge in columns (1) and (2) of Table

[34] A Mundlak (1978) test indicates that we cannot reject the null that the additional district fixed effects are redundant. Results available upon request.

[35] Pedagogical knowledge is measured as the score on a lesson preparation exercise that was administered to all teachers. The assessment of pedagogy knowledge and skills is described in Bold et. al. (2017). While we have data on teacher absence, we do not add this as a regressor here, as it is extremely noisy at the individual teacher level.

6 broadly conform to this pattern (column 2 reports results for the same sample, but omitting the content knowledge of higher grade teachers).[36] That is, there is a small positive, but not significant effect of the subject knowledge of teachers in higher grades on student achievement and including this variable leaves the effect of current and previous teacher knowledge basically unchanged (albeit slightly reduced).[37] Hence, the estimated effects are linked to actual exposure to a teacher, not just the school environment in which the teacher teaches—giving credence to the causal interpretation. Similarly, if the relationship between teacher and student knowledge is purely due to sorting, then the length of exposure to a given teacher should not matter. We test this in columns (3) and (4) of Table 6, where we compare the coefficient on current teacher knowledge for those who have kept their grade class teachers in grade 4 and those who changed class teacher. The coefficient is larger in the first case, implying that length of exposure does indeed matter, again bolstering the causal interpretation.

5.3. Structural effects of teacher content knowledge on student learning

We now turn to solving for the structural estimates of the contemporaneous effect of teacher content knowledge (α), the extent of fade out of the teachers' impact in earlier grades (γ), and by combining them, the cumulative effect of teacher knowledge on student achievement ($\alpha \sum_{t=0}^{3} \gamma^t$). Table 7 reports these structural parameters, which are non-linear functions of the reduced form coefficients in Table 4, column (5), the correlation in teacher knowledge of the student's teacher in grade 3 and 4, estimated to be $\hat{\rho} = 0.58$ and the correlation matrix for the observed and unobserved subject differences in test scores with typical element $\hat{\rho}_{g,g'}$ (see Table 8).[38]

The contemporaneous effect is estimated as 0.06 SD.[39] As a comparison, effect sizes in the value-added literature, which estimates the impact of a broader measure of teacher quality, range from 0.1 to 0.2 SD (Rockoff, 2004; Rivkin et al., 2005; Aaronson et al., 2007; Chetty et al., 2014b; Araujo et al., 2016; and Bau and Das, 2017, Buhl-Wiggers et al. 2018).[40] The degree of persistence is estimated as 0.57 SD, implying that just over half of the short-run effect persists

[36] The sample size falls because not all schools tested teachers in higher grades.
[37] Note that teachers in higher grades are less likely to have taught the grade 4 students, but we cannot rule out that students were exposed to them in grade 1 and 2, which could explain the non-zero effect.
[38] After inserting the estimated $\hat{\rho}_{g,g'}$ and $\hat{\rho}$ in equations (6) and (7), we use MATLAB's fsolve command to solve for α and γ.
[39] When solving for the contemporaneous effect of teacher knowledge from the reduced form coefficients, we adjust for the fact for the fact that test scores are measured half way through year 4, while α measures the impact of teaching after a full year of teaching in each grade.
[40] Bau and Das (2017) show that teacher content knowledge is significantly correlated with estimated VA. Their preferred (IV) specification suggests that a one SD increase in teacher test scores increases VA by roughly 30 percent, or a 0.048 SD increase student test scores.

between one grade and the next. Again, this is consistent with what has been reported using data from other countries (see Kane and Staiger 2008; Jacob, Lefgren, and Sims 2010; Rothstein 2010; and Andrabi et al., 2011). These estimates combine to give a cumulative effect of 0.123 SD.

As seen from Table 8, the magnitudes and patterns of the estimated correlation coefficients are consistent with the assumptions made in Section 4.3, namely the correlation in teacher content knowledge between any two grades is positive, $\rho_{g,g'} \geq 0$ for all g and g'; and (ii) the further apart any two grades, the lower is the correlation in teacher content knowledge across grades; i.e., $\rho_{g,g'}$ is decreasing in $|g - g'|$. We can therefore interpret the sum of the reduced form coefficients, $\hat{\beta}_4 + \hat{\beta}_3$, as a lower bound on the cumulative effect. Moreover, this lower bound is fairly tight: Adjusting the structural estimates for the fact that student test scores are measured half way through year 4, so that $\widetilde{CE} = \alpha \times 0.5 + \alpha \sum_{t=1}^{3} \gamma^t = 0.094$, we see that this number is 13% larger than the sum of the reduced form coefficients, $\hat{\beta}_4 + \hat{\beta}_3 = 0.083$ (Table 4, column 5).

Second, as discussed in Section 4.4, introducing previous teacher test scores as a regressor not only reduces the overall bias, but also shifts a disproportionate amount of the remainder on the corresponding coefficient: The bias in $\hat{\beta}_3$ is 0.015 or 50% of the true $\alpha\gamma$. In contrast, the bias in $\hat{\beta}_4$ is just 0.004 or 15% of the true α (again, adjusting for the fact that student test scores are measured half way through the year). This compares to a bias of 0.038 or 130% of the true α, if $\hat{\beta}_4$ had been estimated based on a contemporaneous specification.

5.4. Implications of the structural parameters

In this subsection, we use the structural parameters to carry out analyses which further shed light on the magnitude of the role of teacher knowledge in student learning. First, we use the structural estimates of the returns to teacher content knowledge to simulate the learning impacts of increasing the content knowledge of all teachers to a minimum threshold. Building on the development accounting approach (see Caselli 2005) we then use an aggregate production for learning to calculate the extent to which shortfalls in teacher knowledge account for the observed low levels of student achievement.

Second, and inspired by the "misallocation" literature (see Banerjee and Duflo 2005; Hsieh and Klenow 2009), we simulate the impacts of reallocating students to teachers in a way that balances the benefits of assigning more students to better teachers with the (potential) costs of the larger class sizes that would result.

Third, we simulate the impacts of the simultaneous implementation of two policy reforms—ensuring that teachers reach a minimum threshold of effective years of education along with increasing teacher effort and reducing absenteeism. In particular, we focus on the longer run impacts of a reform that would ensure that all new teachers hired, keeping pace with population growth and adjusting for teacher retirement, enter the profession with the effective equivalent of lower secondary education.

Accounting for the learning gap

To what extent does the shortfall in teacher content knowledge account for the gap in student learning? To answer this question, we begin by estimating the extent of the teacher content knowledge shortfall, using the measures based on "effective years of schooling" (see section 3). This transformation, as discussed above, allows us to extrapolate beyond content knowledge outcomes observed in the sample in a meaningful and informative way. [41]

Based on our estimates, teachers have an average schooling gap across language and mathematics of 6.1 years relative to the official requirement to 10 years of schooling.[42] Our parameter estimates imply that being taught by teachers with a minimum knowledge of ten years of effective education would, over four years, increase effective years of schooling of grade 4 students by almost three quarters of a year.[43] Compared to the current situation where grade 4 students have reached a level of 1.5 years of effective schooling, this represents an almost 50 percent increase. These magnitudes, in turn, imply that the teacher content knowledge shortfall (relative to 10 years of effective education) accounts for 30 percent of the observed shortfall in student learning (relative to 4 effective years).[44] Alternatively, the difference in teacher content knowledge in mathematics between Kenya, which has the highest student performance in that subject, and Mozambique, which has the lowest, accounts for one fifth of the gap in students' effective years of schooling.

[41] We repeat both the reduced form and the structural analysis with this alternative measure. The results are very similar to the estimates using IRT scores and are presented in Table A.5-A.7 (reduced form estimates) and Table 7, column (2) (structural estimates) and Table A8 (matrix for $\rho_{g,g'}$).

[42] De jure all countries in our sample have well-established systems for teacher training, which confer training at or below the post-secondary non-tertiary level and the large majority of teachers hold such a training certificate. The minimum entry requirement for teacher training is lower secondary education, equivalent to ten years of schooling, which 90% of teachers in our sample have completed.

[43] This result is arrived at by multiplying the cumulative effect of four years of teaching in the second column of Table 7 by the number of effective years of education required to increase from the current average to the minimum requirement (10 years).

[44] This result is arrived at by dividing students' shortfall in effective years of schooling after four years, $2.3 = (4 - 1.5 \times 4/3.5)$, by the amount of learning acquired after four years if teachers increased their human capital by one year.

Together, these results suggest that variation in teacher knowledge explains a sizeable share of the learning gap (either relative to the curriculum or across countries). The results, however, also make clear that the largest share of the gap in learning outcomes (again either relative to the curriculum or across countries) is explained by other shortfalls, potentially including shortfalls in complementary dimensions of teacher quality, such as effort, pedagogical skills, and teaching practice and focus, that are held constant in the analysis. Without improvements in these complementary dimensions, it will not be possible to close the learning gap: our parameter estimates imply that reforms that focus purely on increases in teacher knowledge would require teachers in Sub-Saharan Africa to acquire effective years of education that exceed university level.[45]

Reducing "misallocation"

In the second policy simulation, we examine the efficiency loss due to (one type of) misallocation of students to teachers. Specifically, in a simple model without peer effects, an optimal allocation of students across teachers with heterogeneous skills would imply that teachers who know more about their subject should teach more students.[46] In the data, however, we find a negative correlation between teacher content knowledge and class size.[47] We first examine how much student learning is lost because of this misallocation, and then simulate the effects of moving students from the worst performing teachers to those with relatively better content knowledge, for different assumptions about the effects of class size (and with varying constraints on the feasibility of reallocating students across space).

In this simulation exercise (presented in full in appendix F) we consider several effects that partly offset each other: (i) a positive effect on learning for those students who are transferred from a teacher with low content knowledge to a teacher with higher content knowledge; (ii) a potentially negative effect on learning for the students taught by teachers with high knowledge because of increased class size; (iii) an ambiguous effect on learning for the transferred students arising from increased class size; and (iv) a potentially positive effect

[45] The same argument applies to other inputs. For example, teachers are absent from classroom roughly half of the scheduled teaching time (see Bold et al., 2017) and reducing absenteeism can be an effective way to improving learning (see Duflo, Hanna, Ryan, 2012; Duflo, Dupas, and Kremer, 2015; Bold et al., 2016; and Muralidharan and Sundararaman, 2013). However, to close the remainder of the gap by itself, the return to an additional hour of instruction would have to be counterfactually high, given that teaching time could at most be doubled.

[46] An implicit assumption is that the return to knowledge is approximately constant across the distribution of teacher and student knowledge. The data suggest, if anything, that the returns to teacher knowledge increase in its level, which would further amplify the estimates presented here.

[47] Implicit in this argument is that teachers with higher knowledge do not perform significantly worse along other dimensions that matter, such as effort and classroom skills. We find little evidence of this.

on learning for the students left behind with the low knowledge teacher arising from decreased class sizes.

To calibrate the learning gains from transferring (i), we use the median estimate of the contemporaneous effect of teacher knowledge on student learning across the estimated structural parameters, $\alpha = 0.06$. To calibrate the class size effects (ii) to (iv), we derive a range of estimates from the existing literature. In particular, we assume that class size effects, denoted σ, are linear in log-changes and consider impacts on learning ranging from zero, as estimated experimentally by Duflo, Dupas and Kremer (2015) in Kenya, to an upper bound of a 0.15 SD for a doubling of class size, which is slightly lower than the estimate by Muralidharan and Sundararaman (2013) for India.[48] We first constrain the movement of students to be within a school, and then allow movement across all schools within a district.

Figure 5 illustrates the costs of misallocation in terms of foregone student learning. If there are no costs from larger class sizes, then the human capital lost—relative to the optimal allocation in which all students from below average teachers are moved to better teachers within the district—amounts to 0.04 standard deviations after one year. Naturally, as class size effects increase, the potential gains from reallocation decrease: for the largest σ we consider, the gain in student learning from reallocating students at the district level would be 0.012 SD. Reallocation at the school level still leads to sizeable gains, moving students from the three worst performing teachers in each school to the better teachers would increase student learning by 0.025 of a standard deviation (equivalent to increasing teacher knowledge by one third of a standard deviation) if increasing class sizes is costless and 0.005 of a standard deviation if doubling class sizes reduces learning by .1 of a standard deviation. While these effects are smaller than reallocation at the district level, they are also much easier to implement by school management.

Even though the learning gains from efficient allocation decrease as the cost of learning in larger classes increases, there are sizeable fiscal inefficiencies even when σ is at its upper bound. To calculate these inefficiencies, we conduct a second, related, simulation in which we investigate the learning implications of reallocating *all* students from teachers with less than the median content knowledge to the remaining teachers. We emphasize that this is a statistical simulation not a policy one since there are, of course, substantial human and political economy implications of such an approach that we are not taking into account here.[49]

[48] In either case, we assume that class sizes cannot exceed 200 students.

[49] At the same time, such an approach is not unthinkable: a proposal very much in line with this simulation was recently put forward by the governor of the state of Kaduna, Nigeria. He proposed to dismiss over 20,000

Implicitly, this simulation results in a cost saving of 50 percent (removing half the teachers). Our estimates suggest that learning outcomes could be maintained if a doubling of class sizes reduces learning by no more than 0.04 SD. A similar exercise suggests that even for the highest class-size effect considered, a cost saving of 12 percent could be achieved (by transferring the students of the 12 percent of teachers with the lowest content knowledge) without compromising learning.

Long-run effects of increasing teacher knowledge and time spent teaching

The last policy simulation we consider is the joint effect of increasing both teacher knowledge and effort. The provision of primary education has expanded greatly in low-and middle-income countries in the last two decades, including countries in Sub Saharan Africa. Among current teachers in the SDI data, we find that twice as many teachers have entered the profession in the last ten years than in the decade before. This expansion of the teaching force will likely continue: According to recent population projections, close to half the world's children will live in Africa by the end of the 21st century (UNICEF, 2014). The number of primary school age children in Sub-Saharan Africa is set to rise from about 170 million to 206 million between 2019 and 2029, reaching 270 million by 2050.[50] As illustrated in Figure 6, simply to keep pace with population growth—adjusting for teacher retirement—and to maintain pupil teacher ratios at a rough benchmark of 40 students per teacher, would require the hiring of 1.6 million new teachers by 2030 and close to 5 million by 2050. As the large majority of teachers are employed on permanent and pensionable civil service contracts, and as upgrading current teachers' knowledge is probably significantly more costly than raising the quality of new cohorts of teachers, an expansion of the teaching force at low knowledge represents a lost potential with durable impacts.[51]

We use the estimate of the cumulative impact of content knowledge to simulate the impact of a reform that raises teachers' knowledge over time to the minimum entry requirement for teacher training; i.e., lower secondary education (10 years of effective years of education). We treat the knowledge of current teachers as fixed, which means that average teacher content knowledge can only be raised by hiring new qualified teachers, and we assume that teachers work for 40 years. We assume teachers follow their class up to grade six

teachers in response to findings from an evaluation showing that two-thirds of the lower primary teachers failed to score 75% or higher on assessments/exams set for their students. Perhaps unsurprisingly, the proposal led to substantial controversy and political push-back.

[50] See appendix for details for these projections and the policy simulations.

[51] Note though that in countries where contract teachers are prevalent, and for teachers with less than ten years of experience, almost half of the teachers are employed on short-term contracts. This partly reflects an age and a cohort effect, as many contract teachers graduate to civil service status over time (Bold et al., 2017).

when they again start teaching first grade students and new teachers are placed to replace old teachers and to ensure a constant student-teacher ratio.[52] We also simulate the impact of increasing the time teachers spend teaching, from the current average of 2 hours and 46 minutes per school day (see Table 1) to the OECD average (4 hours and 30 minutes of compulsory instructional time per school day); i.e. an increase of 63%. To estimate this effect, we interpret α as the contemporaneous effect of teacher content knowledge on student achievement of, on average, 2 hours and 46 minutes of instructional time per day. We then assume that the effect is linear in time spent teaching, implying that the marginal effect of teacher content knowledge, when teachers teach 4 hours and 30 minutes, is 1.63α. Finally, we consider the effect of combining the two reforms. The results of the policy simulations are summarized in Figure 7.

Four key results of this simulation stand out. First, even if all newly hired teachers have the required minimum content knowledge (a stark difference from the current system) the reform has only relatively small effects even after 10 years: effective years of student learning are simulated to increase by approximately 14 percent in the first decade. This relatively small effect, in turn, is caused primarily by the fact that it takes time to improve grade 4 outcomes, given that in the shorter run most students will still be taught by teachers already hired before the reform. For example, in the first year (2019), only 3 percent of the grade 4 students will benefit from having a fully qualified teacher.

Second, the simulation suggests that the reform would increase effective years of schooling by approximately 36 percent by 2050. We note that this is likely a lower bound, since higher subject content knowledge may also raise other aspects of teacher quality.

Third, increasing the time teachers teach raises student achievement directly (by assumption). But the magnitude of this impact is small (18 percent in the longer run), mostly because of the poor content knowledge of teachers. Note that in contrast to the reform aimed at improving new teachers' content knowledge, this reform is assumed to influence all teachers.[53] Therefore, the simulation shows starkly how having teachers that know little teach more only marginally raises student achievement.

Fourth, combining the reforms leads to big short and longer run impacts. After 5 years, for example, students' effective years of schooling increases by 26 percent and by

[52] The simulations reported below are robust to various alternative assumptions about how new teachers are allocated to schools.

[53] Here we project a reform that broadly increases the incentives to teach and thus a reform that affects all teachers.

2050, the simulation suggests that it will have increased by more than 60 percent (to almost 3 years). These simulations (illustrated in Figure 7) make clear the complementarity between the reforms: combining them leads to substantially larger marginal impacts than either reform alone.

6. Discussion

Recent estimates suggest that the quality of human capital can explain a dominant share of world income differences (Jones 2014; Malmberg 2017) in part through its effects on economic growth (Hanushek and Woessmann 2015; World Bank 2018). Thus, the fact that many children in low income countries learn little from attending school may be one of the most pressing development challenges. In this analysis, we focus on one component of the education production function—teachers' knowledge of the subject they are teaching. While a growing literature has shown that teachers matter, much less is known about the link between specific teacher characteristics and student learning (Glewwe and Muralidharan 2015). Here we show that teachers' content knowledge, or lack thereof, is an important part of the reason why primary school students in Sub-Saharan Africa are already well behind the expected curriculum—and their peers in other parts of the world—after only a few years of schooling. This implies a loss of potential human capital for numerous cohorts of students.

Our results have implications for both research and policy. Regarding the former, our results strongly suggest that research on how to improve teacher content knowledge should be a priority—there are few well identified studies on how to improve the quality of pre-service training (Glewwe and Muralidharan 2015). Moreover, we show that the marginal gain from collecting matched student and teacher data for one additional grade is large. Specifically, with additional data on teacher knowledge in other lower primary grades, one can recover the structural parameters of interest under the assumption that the cumulative education production function is fully characterized by two parameters, α and γ.

Our simulation results suggest that reforms that ensure that upcoming cohorts of teachers have greater mastery of the content they are expected to teach, combined with incentives to increase time spent teaching, could make a big improvement to student learning outcomes—almost doubling them. But the (aggregate) impact of such reforms would take time to materialize—working through these channels alone it would take on the order of 30 years to double learning outcomes. For that reason, it is also important to experiment with and roll-out shorter-term approaches. A number of interventions have been shown to produce promising results, including programs to supplement current teachers with additional

instructors, to leverage computer-aided learning programs to supplement teachers, or to support teachers with scripted lesson plans (Banerjee, et al. 2017).[54]

The data reveal large variation both within and across countries, with implications for policy. For example, in countries like Kenya where teacher content knowledge is relatively high, but teacher absence is also high, a focus on increasing teacher effort (reducing absence) should be the priority. On the other hand, a similar approach in countries like Togo and Mozambique would be unlikely to yield the same results given the low teacher content knowledge.

Overall, the results presented here—along with our other work based on these same data (Bold et al. 2017)—suggest that there are important complementarities between various dimensions of teacher quality: Increasing the effort of teachers who are well-trained and educated will likely go a long way toward increasing human capital accumulation.

[54] See, also, for example, the reviews in Murnane and Ganimian (2014), Glewwe and Muralidharan (2015), and Evans and Popova (2016), as well as the discussion in World Bank (2018).

References

Aaronson, Daniel, Lisa Barrow, William Sander. 2007. “Teachers and Student Achievement in the Chicago Public High Schools.” *Journal of Labor Economics* 25 (1), pp. 95–135.

Andrabi, Tahir, Jishnu Das, and Asim Ijaz Khwaja. 2009. “Do Value-Added Estimates Add Value? Accounting for Learning Dynamics, Harvard Kennedy School Faculty Research Working Papers Series, RWP09-034

ASER. 2013. *Annual Status of Education Report (Rural) 2013*. ASER Center. New Delhi.

Araujo. M. Caridad, Pedro Carneiro, Yyannú Cruz-Aguayo, and Norbert Schady. 2016. “Teacher Quality and Learning Outcomes in Kindergarten.” Quarterly Journal of Economics, 131(3): 1415–1453.

Ashenfelter, Orley and David J. Zimmerman. 1997. “Estimates of the returns to schooling from sibling data: fathers, sons, and brothers.” *The Review of Economics and Statistics* 79 (1), pp. 1–9.

Banerjee, Abhijit and Esther Duflo. 2005. “Growth Theory through the Lens of Development Economics”. In Aghion, P., Durlauf, S. (Eds.) *Handbook of Economic Growth*, vol 1: 473-552, Elsevier.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton. 2017. “From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application.” *Journal of Economic Perspectives*, 31(4): 73-102.

Bau Natalee and Jishnu Das. 2017. “The Misallocation of Pay and Productivity in the Public Sector: Evidence from the Labor Market for Teachers.” Policy Research Working Paper no. 8050, The World Bank.

Behrman, Jere R. 2010. “Investment in Education: Inputs and Incentives.” In Dani Rodrik and Mark Rosenzweig, eds., *Handbook of Development Economics*, Vol. 5, Elsevier, pp. 4883–4975.

Bietenbeck, Jan, Marc Piopiunik, and Simon Wiederhold. 2017. “Africa’s Skill Tragedy: Does Teachers’ Lack of Knowledge Lead to Low Student Performance?” *Journal of Human Resources* (forthcoming).

Bold, Tessa, Deon Filmer, Gayle Martin, Ezequiel Molina, Brian Stacy, Christophe Rockmore, Jakob Svensson, and Waly Wane. 2017. “Enrollment Without Learning: Teacher Effort, Knowledge, and Skill in Primary Schools in Africa”. *Journal of Economic Perspectives*, 31(4): 185-204.

Bold, Tessa, Bernard Gauthier, Jakob Svensson, and Waly Wane. 2010. “Delivering Service Indicators in Education and Health in Africa: A Proposal.” Policy Research Working Paper No. 5327, The World Bank.

Bold, Tessa, Bernard Gauthier, Jakob Svensson, and Waly Wane. 2011. “Service Delivery Indicators: Pilot in Education and Health Care in Africa.” CMI report no. 2011:8, Chr. Michelsen Institute, Bergen, Norway.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a and Justin Sandefur. 2019. "Experimental Evidence on Scaling Up Education Reforms in Kenya.", *Journal of Public Economics*, 168: pp.1-20

Buhl-Wiggers, Julie, Jason T. Kerwin, Jeffrey A. Smith and Rebecca Thornton. 2018 "Teacher Effectiveness in Africa: Longitudinal and Causal Estimates", *IGC Working Paper*, S-89238-UGA-1.

Caselli, Francesco. 2005. Accounting for Cross-Country Income Differences. In Aghion, P., Durlauf, S. (Eds.) *Handbook of Economic Growth*, vol 1: 679–741, Elsevier.

Chaudhury, Nazmul, Jeffrey Hammer, Michael Kremer, Karthik Muralidharan and F. Halsey Rogers. 2006. "Missing in action: teacher and health worker absence in developing countries." *Journal of Economic Perspectives* 20:1, pp. 91–116.

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

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014b. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood". *American Economic Review*, 104(9): 2633–2679.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2010. "Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects." *Journal of Human Resources*, 45, 655–681.

Dee, Thomas S. 2005. "A Teacher like Me: Does Race, Ethnicity, or Gender Matter?" *American Economic Review, Papers and Proceedings*, 95, 158–165.

Dee, Thomas S. 2007. "Teachers and the Gender Gaps in Student Achievement." *Journal of Human Resources*, 42, 528–554.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2015. "School governance, teacher incentives, and pupil-teacher ratios: Experimental evidence from Kenyan primary schools." *Journal of Public Economics*, Volume 123, pp: 92–110.

Duflo, Esther, Rema Hanna, and Stephen P. Ryan. 2012. "Incentives Work: Getting Teachers to Come to School." *American Economic Review*, 102(4): 1241–78.

Evans, David K. and Anna Popova. 2016. "What Really Works to Improve Learning in Developing Countries: An Analysis of Divergent Findings in Systematic Reviews". *World Bank Research Observer*. 31:2, pp. 242-270.

Ganimian, Alejandro J. and Richard J. Murnane. 2016 "Improving Education in Developing Countries: Lessons from Rigorous Impact Evaluations." *Review of Educational Research*, 86(3): 719–755.

Glewwe, Paul, Michael Kremer. 2006. "Schools, Teachers, and Education Outcomes in Developing Countries". In Eric A. Hanushek, Finis Welch (eds.), *Handbook of the Economics of Education*, Volume 2, pp. 945–1017. Amsterdam: North-Holland.

Glewwe, Paul and Karthik Muralidharan. 2015. "Improving School Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications." RISE Working Paper No. 15/001.Glewwe and Muralidharan (2015)

Hanushek, Eric A. and Steven G. Rivkin. 2006. “Teacher Quality.” In Eric A. Hanushek, Finis Welch (eds.), *Handbook of the Economics of Education*, Volume 2, pp. 1051–1078. Amsterdam: North-Holland.

Hanushek, Eric A., and Ludger Woessmann. 2015. The Knowledge Capital of Nations: Education and the Economics of Growth. CESifo Book Series. Cambridge, MA: MIT Press.

Hanushek, Eric A., Marc Piopiunik and Simon Wiederhold. 2019, “The Value of Smarter Teachers: International Evidence on Teacher Cognitive Skills and Student Performance”, *Journal of Human Resources*, forthcoming

Hsieh, Chang-Tai and Peter J. Klenow. 2009. “Misallocation and Manufacturing TFP in China and India.” *Quarterly Journal of Economics*, 124(4): 1403–1448.

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

Jacob, Brian and Jesse Rothstein. 2016. “The measurement of student ability in modern assessment systems.” *Journal of Economic Perspectives*, 30(3): 85-108.

Jennrich, R.I. 1970. “An asymptotic χ^2test for the equality of two correlation matrices.” *Journal of the American Statistical Association*, 65, 904-912.

Johnson, David, Andrew Cunningham and Rachel Dowling. 2012. “Teaching Standards and Curriculum Review”. Mimeo, The World Bank.

Jones, Benjamin F. 2014. “The Human Capital Stock: A Generalized Approach.” *American Economic Review*, 104(11): 3752-77.

Kane, Thomas J. and Douglas O. Staiger. 2008. “Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation.” NBER Working Paper No. 14607.

Kremer, Michael, Conner Brannen and Rachel Glennerster. 2013. “The Challenge of Education and Learning in the Developing World”, *Science* 340: 297-300.

Lavy, Victor. 2015. “Do Differences in Schools’ Instruction Time Explain International Achievement Gaps? Evidence from Developed and Developing Countries.” *Economic Journal*, 125, F397–F424.

Malmberg, Hannes. 2017 “Human Capital and Development Accounting Revisited”, Working Paper, IIES.

Metzler, Johannes and Ludger Woessmann. 2012. “The impact of teacher subject knowledge on student achievement: Evidence from within-teacher within-student variation.” *Journal of Development Economics*, 99, 486–496.

Mundlak, Yair. 1978. “On the Pooling of Time Series and Cross Section Data.” *Econometrica*, 46(1): 69-85.

Muralidharan, Karthik and Venkatesh Sundararaman. 2011. “Teacher Performance Pay: Experimental Evidence from India”. *Journal of Political Economy*, 119, No. 1, pp. 39-77.

Muralidharan, Karthik and Venkatesh Sundararaman. 2013. “Contract Teachers: Experimental Evidence from India”. NBER Working Paper No. 19440.

Rivkin, Steven G., Eric A. Hanushek, John F. Kain. 2005. “Teachers, Schools, and Academic Achievement”. *Econometrica* 73 (2), pp. 417–458.

Rockoff, Jonah E. 2004. “The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data”. *American Economic Review* 94 (2), pp. 247–252.

Rothstein, Jesse. 2010. “Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement.” *Quarterly Journal of Economics* 125 (1): 175–214.

StataCorp 2015. “Stata Multivariate Statistics Reference Manual, Release 14”, Stata Press, College Station, Texas.

Singh, Abhijeet. 2019. “Learning more with every Year: School year productivity and international learning gaps.” *Journal of the European Economic Association*, forthcoming

Todd, Petra E. and Kenneth I. Wolpin. 2003. “On the Specification and Estimation of the Production Function for Cognitive Achievement.” *Economic Journal*, 113, pp. F3-F33.

UIS. 2011. “Financing Education in Sub-Saharan Africa: Meeting the Challenges of Expansion, Equity and Quality”. UNESCO/UIS. Montreal.

UIS. 2018. “One in Five Children, Adolescents and Youth is Out of School.” UIS fact sheet No. 48. UNESCO/UIS. Montreal.

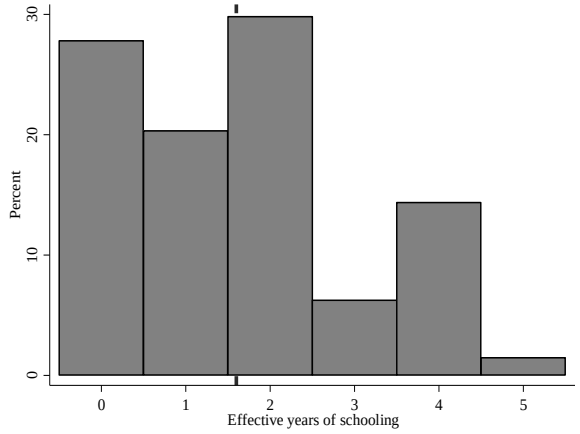
UNICEF. 2014. *Generation 2030/Africa*. UNICEF. New York.

UNICEF. 2015. *EFA Global Monitoring Report*. UNICEF. Paris.

World Bank. 2018. *World Development Report 2018: Learning to Realize Education’s Promise.* Washington, DC.

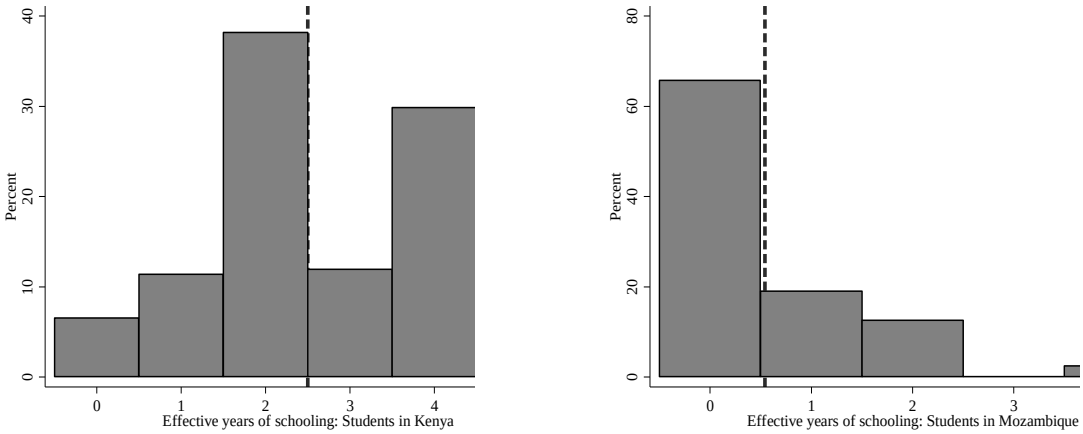
World Bank. 2019. *World Development Indicators*. Washington DC.

Figure 1: Effective years of schooling for students after four years of primary education



Note: Distribution of effective years of schooling for students (pooled data across countries and subjects). Dashed vertical line depicts mean.

Figure 2: Effective years of schooling for students after four years of primary education: Kenya and Mozambique

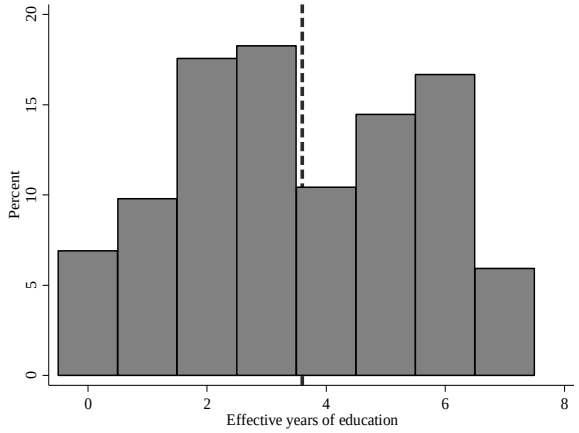


Panel A: Kenya

Panel B: Mozambique

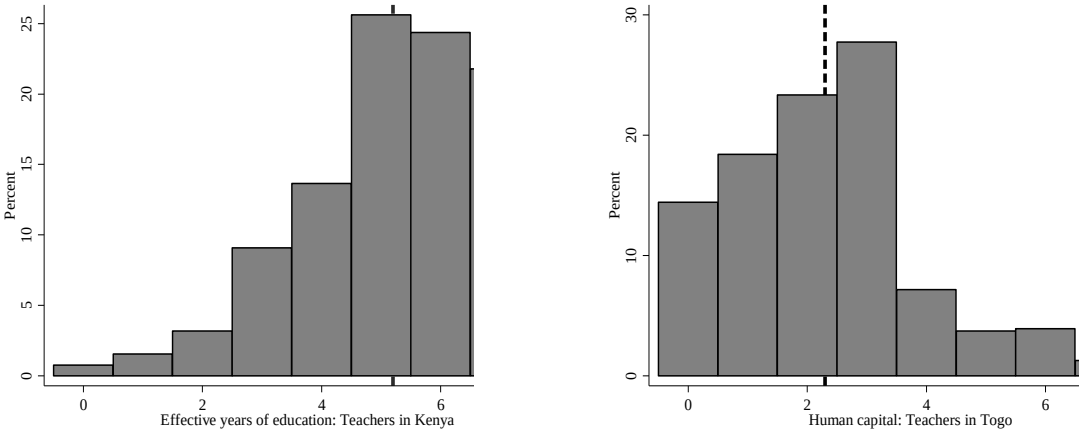
Note: Distribution of effective years of schooling for students in Kenya (Panel A) and Mozambique (Panel B). Pooled data across subjects for each country. Dashed vertical lines depict means.

Figure 3: Effective years of education for primary school teachers



Note: Distribution of effective years of education for primary school teachers (pooled data across countries and subjects). Dashed vertical line depicts mean.

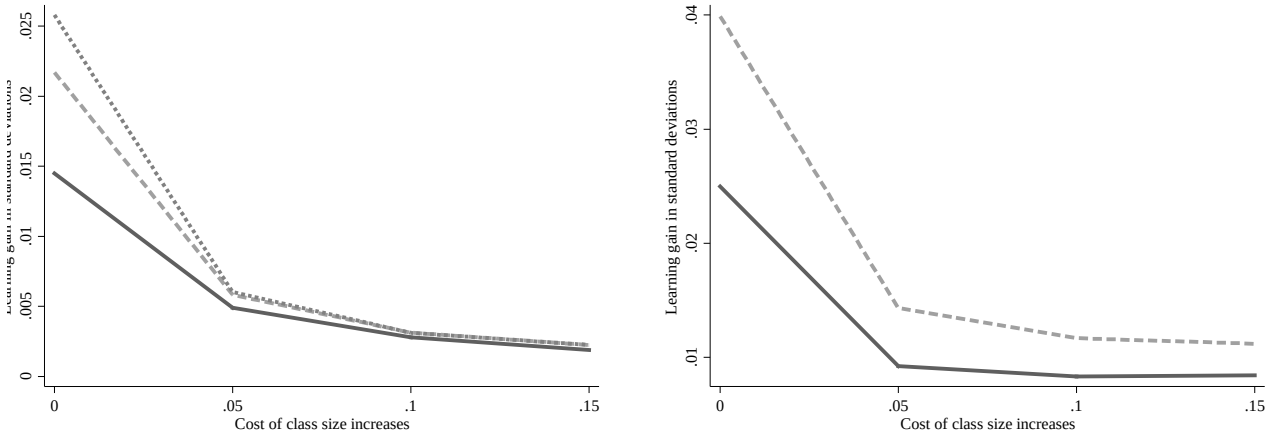
Figure 4: Effective years of education for primary school teachers: Kenya and Togo



Panel A: Kenya Panel B: Togo

Note: Distribution of effective years of education for primary school teachers in Kenya (Panel A) and Togo (Panel B). Pooled data across subjects for each country. Dashed vertical lines depict means.

Figure 5: Misallocation



Panel A: Reallocation at the district level

Panel B: Reallocation at the school level

Note: The figure shows the learning gain of reallocating from the teachers with low knowledge to teachers with higher knowledge as a function of the cost of increasing class sizes. Panel A shows the effects for reallocating at the district level and panel B shows the effects for reallocating at the school level. The solid (dashed) line in Panel A shows the effect on learning of reallocating from those teachers below the 25th (50th) percentile of the knowledge distribution. The solid (dashed/dotted) line in Panel B shows the effect of reallocation from 1 (2/3) teachers with the lowest knowledge in each school.

Figure 6: Teacher projections, 2019-2050

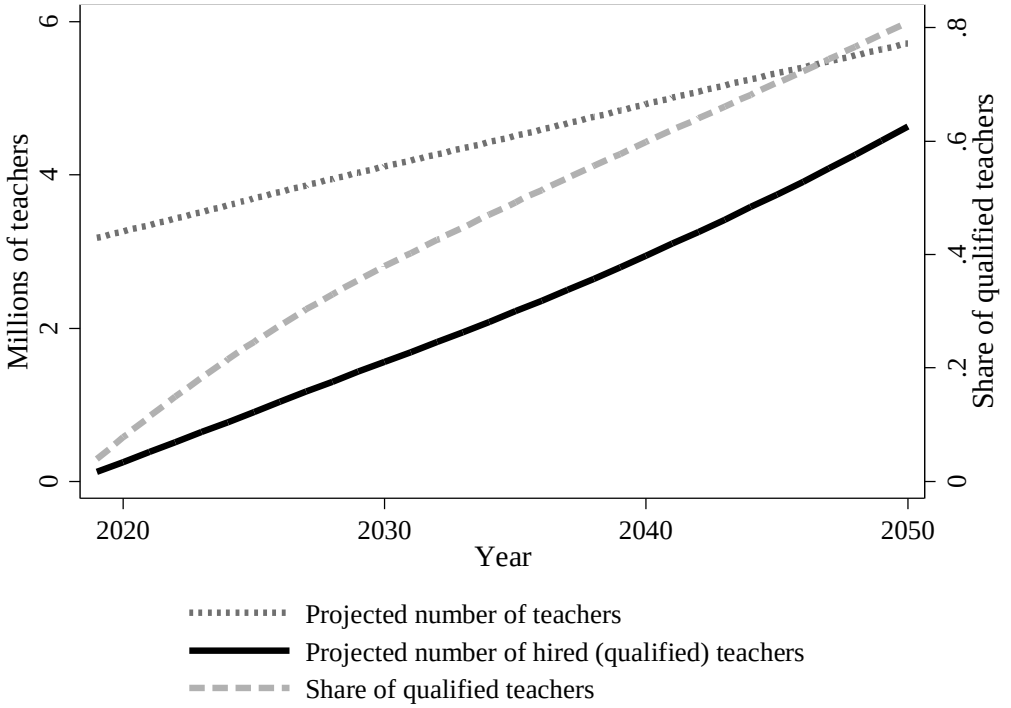


Figure 7: Projected impact of three policy reforms, 2019-2050

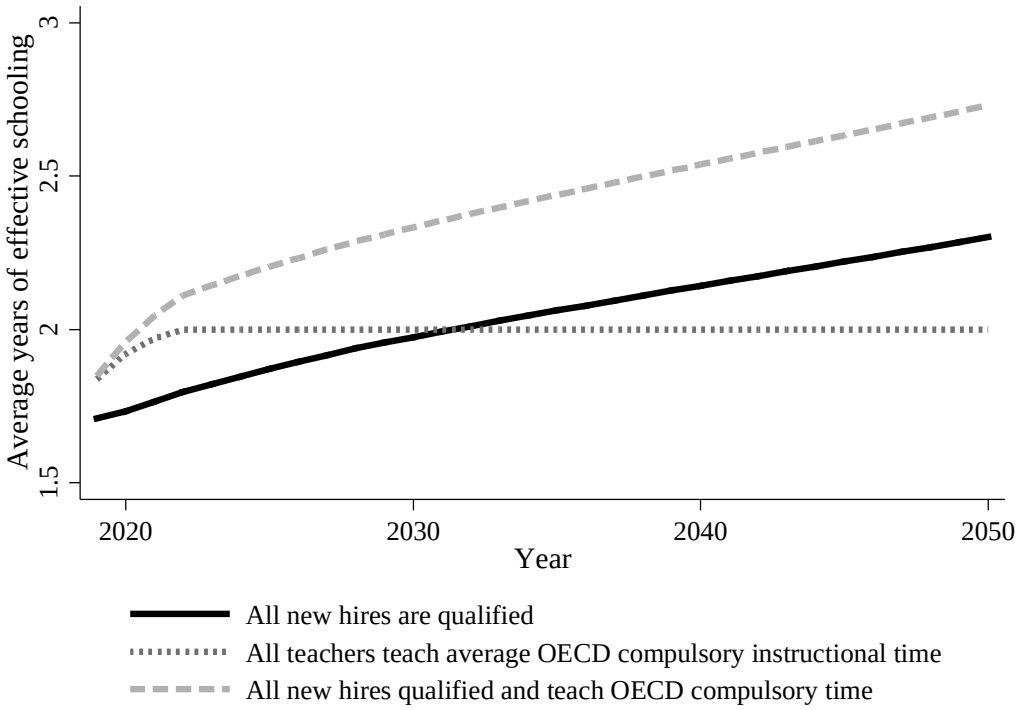


Table 1: Summary statistics

	Mean
Absence from class (%)	44
Absence from school (%)	23
Scheduled teaching time (h min)	5h 27mins
Time spent teaching (h min)	2h 46mins
Minimum general pedagogy knowledge (%)	11
Minimum knowledge assessing students (%)	0

Note: See Bold et al. (2017) for details. Pooled data for Kenya, Mozambique, Nigeria, Senegal, Tanzania, Togo, and Uganda on teacher quality. All individual country statistics are calculated using country-specific sampling weights. The average for the pooled sample is taken by averaging over the country averages. Teachers are marked as absent from school if during an unannounced visit they are not found anywhere on the school premises. Otherwise, they are marked as present. Teachers are marked as absent from class if during an unannounced visit, they are absent from school or present at school but absent from the classroom. Otherwise, they are marked as present. The scheduled teaching time is the length of the school day minus break time. Time spent teaching adjusts the length of the school day by the share of teachers who are present in the classroom, on average, and the time the teacher spends teaching while in the classroom. A teacher is defined as having minimum knowledge of general pedagogy if she scores at least 80% on the tasks that relate to general pedagogy (factual text comprehension and being able to formulate learning outcomes and lesson aims). A teacher is defined as having minimum knowledge for assessing students if they score least 80% on the tasks that relate to assessment (comparing students' writing and monitoring progress among a group of students).

Table 2: IRT score and effective years of schooling for students

	IRT scores		Distribution of effective years of schooling	
Effective years of schooling	Language	Mathematics	Language	Mathematics
0	-1.19	-0.99	27% (27%)	36% (36%)
1	-0.56	-0.06	15% (42%)	21% (57%)
2	0.15	0.49	45% (87%)	18% (75%)
3	0.98	0.92	5% (92%)	6% (81%)
4	1.48	0.94	8% (100%)	17% (98%)
5	-	1.57	n/a	3% (100%)
N	23,884	23,016	23,884	23,016

Note: Columns (2) and (3): Item Response Scores (IRT) for students, conditional on effective years of schooling, calculated using country-specific sampling weights. Columns (4) and (5): Distribution of effective years of schooling for students, calculated using country-specific sampling weights, cumulative distribution in brackets.

Table 3: IRT scores and effective years of education for teachers

	IRT scores		Distribution of scores	
Effective years of education	Language	Mathematics	Language	Mathematics
0	-1.54	-2.06	5% (5%)	7% (7%)
1	-0.95	-0.84	4% (9%)	14% (21%)
2	-0.43	-0.41	12% (21%)	13% (34%)
3	-0.03	-0.003	19% (40%)	12% (46%)
4	0.34	0.54	20% (60%)	6% (52%)
5	0.88	0.85	32% (92%)	2% (54%)
6	1.08	0.66	2% (94%)	32% (86%)
7	1.50	1.09	6% (100%)	15% (100%)
N	5758	6019	5758	6019

Note: Columns (2) and (3): Item Response Scores (IRT), conditional on effective years of education, calculated using country-specific sampling weights. Columns (4) and (5): Distribution of effective years of education, calculated using country-specific sampling weights, cumulative distribution in brackets.

Table 4: Relationship between student and teacher content knowledge

	(1)	(2)	(3)	(4)	(5)
Dep. variable	Student test scores				
Content knowledge of current teacher	0.175***	0.082***	0.068***	0.060***	0.034
	(.013)	(.013)	(.017)	(.022)	(.027)
Content knowledge of prior teacher				0.038*	0.049*
				(.021)	(.027)
Language	-0.038***	-0.024**	0.023	-0.014	0.017
	(.013)	(.012)	(.016)	(.018)	(.021)
Constant	0.504***	-0.030***	-0.184***	-0.069***	-0.185***
	(.030)	(.006)	(.011)	(.009)	(.016)
Lower bound (total effect)				0.099***	0.083***
				(.021)	(.024)
Observations	29,809	29,809	15,128	16,363	8,896
Adj. R-squared	0.188	0.081	0.042	0.073	0.038
Number of schools	1,960	1,960	904	1,458	626
Number of students	16,794	16,794	7,638	9,981	4,503
Country FE	X				
Student FE		X	X	X	X
Same teacher in both subjects			X		X

Note: Dependent variable: Student test score (IRT rescaled). Lower bound (total effect) is the test of the sum of the estimated coefficients on the content knowledge of current and prior teacher. Clustered, by school, standard errors in parenthesis. *** 1% , ** 5% , * 10% significance.

Table 5: Specification tests

	(1)	(2)	(3)	(4)	(5)
Dep. Variable	Student test scores				
Content knowledge of current teacher	0.034 (.028)	0.029 (.023)	0.027 (.027)	0.055** (.022)	0.029 (.029)
Content knowledge of prior teacher	0.049* (.027)	0.056** (.023)	0.053** (.026)	0.038* (.021)	0.057** (.028)
Teacher pedagogy score				0.242** (.123)	
Constant	-0.017 (.021)	-0.017 (.018)	-0.018 (.026)	0.015 (.018)	0.018 (.022)
Lower bound (total effect)	0.083*** (.024)	0.085*** (.023)	0.080*** (.024)	0.094*** (.021)	0.086*** (.024)
Specification/Sample	Main	District FE	Urban FE	Student FE	One gr. 4 class
Same teacher in both subjects	X	X	X		X
Observations	4,393	4,393	4,393	6,382	3,831
Number of schools	605	605	605	951	523

Note: Dependent variable: Student test score (IRT rescaled). Lower bound (total effect) is the test of the sum of the estimated coefficients on the content knowledge of current and prior teacher. First difference (across subjects) specification, with clustered, by school, standard errors in parenthesis. Specification: (1) main specification with the sample of students with the same teacher in language and mathematics in a given year (grade 3 and 4); (2) main specification with subject variant district fixed effects; (3) main specification with subject variant urban fixed effects; (4) sample of all students with data on current and previous teacher score and current teacher pedagogy score; (5) main specification on sample of schools with one grade 4 classroom. *** 1% , ** 5% , * 10% significance.

Table 6: Placebo tests

	(1)	(2)	(3)	(4)
Dep var.	Student test scores			
Content knowledge of current teacher	0.071*	0.075*	0.121***	0.050***
	(0.041)	(0.042)	(0.038)	(0.018)
Content knowledge of previous teacher	0.055	0.065*		
	(0.037)	(0.037)		
Content knowledge of higher grade teachers	0.037			
	(0.035)			
Constant	-0.150***	-0.151***	0.022	-0.038***
	(0.034)	(0.034)	(0.036)	(0.017)
Lower bound (total effect)	0.126***	0.140***		
	(0.040)	(0.040)		
Specification/Sample	Placebo	Comparison	Same class teacher in grade 3 & 4	Different class teachers in grade 3 & 4
Observations	1,645	1,645	1,804	5,686
Number of schools	214	214	280	727

Note: First difference (across subjects) specification, with clustered, by school, standard errors in parenthesis. Specification: (1) main specification controlling for teacher content knowledge of higher grade teachers in school; (2) main specification using the same sample as in column (1); (3) Sample of students with the same teacher in both subjects in both years (grade 3 and grade 4); (4) Sample of students with a new teacher teaching both subjects in grade 4. *** 1% , ** 5%, * 10% significance.

Table 7: Structural parameters

	(1)	(2)
Contemporaneous effect (α)	0.059	0.043
	(0.040)	(0.029)
Persistence (γ)	0.574	0.731
	(.805)	(0.787)
Total effect after 4 years	0.123***	0.115***
	[0.000]	[0.000]

Note: Estimates for α, γ and the cumulative effect across all correlation structures. Column (1) reports estimates using IRT scores for students as the dependent and for teachers as the explanatory variable. Column (2) uses effective years of education of teachers (explanatory variables) and effective years of schooling of students (dependent variable). Standard errors in parenthesis, with p-value in brackets giving the probability of the null hypothesis that the cumulative effect is zero.

Table 8: Correlation in teacher knowledge (subject-differenced) for student i's teachers in grade 3 and 4 and teachers in lower grades (IRT measure)

$\rho_{g,g'}$	Grade 3	Grade 4
Grade 1	0.55	0.43
Grade 2	0.53	0.32
Grade 3	-	0.58

Note: The table reports $\rho_{g,g'}$, the correlation in teacher knowledge (subject-differenced) for teachers in grade $g = 3,4$ and $g' = 1,2,3$. For $g' = 1,2$, the correlation coefficient is estimated as $\rho_{g,g'} \approx s_g \times 1 + (1 - s_g) \times \rho_{g,g'|j(g)\neq j(g')}$ using the inputs in Table A1 and A2. For $g' = 3$, it is estimated from the variance-covariance matrix of the regressors in equation (5).

Appendix:

A. Definition of curriculum-adjusted years of human capital

We define a student to have 0 years of human capital in language, if they cannot read three letters. A student is defined as having one year of human capital in language, if they can read three letters, but cannot do more advanced tasks. They are scored as having two years of human capital in language, if they can read three words, but cannot do any more advanced tasks. They are scored as having accumulated three years of human capital if they have basic vocabulary, can read a sentence, half a paragraph and answer a basic comprehension question, but cannot do more advanced tasks. Finally, they are scored as having four years of human capital if they can read the whole paragraph and answer an advanced comprehension question.

In mathematics, we score a student as having zero years of human capital if they cannot recognize numbers or cannot do single digit addition or cannot do single digit subtraction. We score them as having one year of human capital if they can recognize numbers, do single digit addition and single digit subtraction, but not any of the more advanced tasks. We score them as having two years of human capital if they can perform double digit addition, triple digit addition and order numbers between 0 and 999. We class them as having three years of human capital if they can multiply single digits, divide single digits and do double digit subtraction. We class them as having accumulated four years of human capital if they can divide double by single digits and compare fractions and as having five years of human capital if they can multiply double digits.

On the teacher side, we score teachers as having no years of human capital, if they could not answer the simplest grammar question, namely forming a question with "Where is…?" and using 'who' in order to define what person is doing. We scored them as having one year of human capital if they could formulate such a question, but could not do any of the more advanced material. We scored them as having two years of human capital if they could use 'when' as a conjunction, could form a sentence asking 'how much' and used 'which' correctly. We scored them as having three years of human capital if they could use because and so correctly as conjunctions, and we scored them as having four years of human capital if they could form a sentence with a conditional statement, use past passive and use unless correctly. We score them as having 5 years of human capital if they could complete more than 70% of an unprompted Cloze passage, as six years if they could correct more than 70% of the mistakes in a letter written by a fourth grade students and as seven years if they could complete both these tasks.

For mathematics, we score teachers as having 1 year of human capital if they could not add double digits (without borrowing). We score them as having one year of education if they could add double digits (without borrowing), but could not do any of the more advanced tasks. We scored them as having two years of human capital if they could add triple digits and recognize basic geometric shapes, but could not do any of the more advanced tasks. We score them as having three years of human capital if they could subtract double digits (with borrowing), and divide a double digit by a single digit. We scored them as having four years of human capital if they could add decimals, solve a multiplication problem involving

monetary unity, subtract decimals. We scored them as having five years of human capital if they could multiply double digits, manipulate fractions and solve a problem involving units of time. We scored them as having six years of human capital if they could solve square roots up to twelve, solve for an unknown in an algebraic equation. We scored them as having seven years of human capital if they could analyze data in a graph, divide fractions, and calculate the perimeter and area of a rectangle.

B. Expression for the OLS estimator of β_3 and β_4

The OLS estimators of β_4 and β_3 in equation (5) are

(A1) $$\hat{\beta}_4 = \left(1 - \frac{\sum \Delta\tilde{x}_3\Delta\tilde{x}_4}{\sum \Delta\tilde{x}_3^2}\frac{\sum \Delta\tilde{x}_3\Delta\tilde{x}_4}{\sum \Delta\tilde{x}_4^2}\right)^{-1}\left(\frac{\sum \Delta\tilde{x}_4\Delta\tilde{y}}{\sum \Delta\tilde{x}_4^2} - \frac{\sum \Delta\tilde{x}_3\Delta\tilde{x}_4}{\sum \Delta\tilde{x}_4^2}\frac{\sum \Delta\tilde{x}_3\Delta\tilde{y}}{\sum \Delta\tilde{x}_3^2}\right),$$

and

(A2) $$\hat{\beta}_3 = \left(1 - \frac{\sum \Delta\tilde{x}_3\Delta\tilde{x}_4}{\sum \Delta\tilde{x}_3^2}\frac{\sum \Delta\tilde{x}_3\Delta\tilde{x}_4}{\sum \Delta\tilde{x}_4^2}\right)^{-1}\left(\frac{\sum \Delta\tilde{x}_3\Delta\tilde{y}}{\sum \Delta\tilde{x}_3^2} - \frac{\sum \Delta\tilde{x}_3\Delta\tilde{x}_4}{\sum \Delta\tilde{x}_4^2}\frac{\sum \Delta\tilde{x}_4\Delta\tilde{y}}{\sum \Delta\tilde{x}_4^2}\right),$$

where a tilde above the variable denotes a demeaned variable. To find their probability limits, substitute the true model for $\Delta\tilde{y}$, divide each of the summed terms by N, the number of observations, and let N go to infinity. The resulting expressions when $var(\Delta x_g) = \sigma^2_{\Delta x}$ for all g are depicted in (6) and (7).

Note that

(A3) $$\hat{\beta}_4 + \hat{\beta}_3 = \alpha_4 + \alpha_3\gamma_{4,3} + \alpha_2\gamma_{4,2}\left(\frac{\rho_{4,2}+\rho_{3,2}}{1+\rho}\right) + \alpha_1\gamma_{1,4}\left(\frac{\rho_{4,1}+\rho_{3,1}}{1+\rho}\right)$$

If all $\rho_{g,g'} \geq 0$ for all g and g' and $\rho_{g,g'}$ is decreasing in $|g - g'|$, then $\rho_{3,2} \leq 1$; $\rho_{4,2} \leq \rho_{4,3} = \rho$; $\rho_{3,1} \leq 1$; $\rho_{4,1} \leq \rho$, which in turn implies that $\frac{\rho_{4,2}+\rho_{3,2}}{1+\rho} < 1$ and $\frac{\rho_{3,1}+\rho_{4,1}}{1+\rho} < 1$. That is, $\hat{\beta}_4 + \hat{\beta}_3$ provides a lower bound on the cumulative effect $\alpha_4 + \alpha_3\gamma_{4,3} + \alpha_2\gamma_{4,2} + \alpha_1\gamma_{4,1}$.

Second, note that the expression for the probability limit of $\hat{\beta}_4$ in a contemporaneous specification is

(A4) $$\text{plim}\ \hat{\beta}_4 = \alpha + \alpha\gamma\rho_{4,3} + \alpha\gamma^2\rho_{4,2} + \alpha\gamma^3\rho_{4,1},$$

and the difference in the asymptotic bias between the contemporaneous and the cumulative specification is therefore:

(A5) $$Bias(\beta_{4,cum}) - Bias(\beta_{4,cont}) =$$

$$\alpha\gamma^2\left(\frac{\rho_{4,2}-\rho_{3,2}\rho}{1-\rho^2}\right) + \alpha\gamma^3\left(\frac{\rho_{4,1}-\rho_{3,1}\rho}{1-\rho^2}\right) - \left(\alpha\gamma^2\rho_{4,2} + \alpha\gamma^3\rho_{4,1}\right) - \alpha\gamma\rho =$$

$$\alpha\gamma^2\left(\frac{\rho(\rho_{4,2}\rho-\rho_{3,2})}{1-\rho^2}\right) + \alpha\gamma^3\left(\frac{\rho(\rho_{4,1}\rho-\rho_{3,1})}{1-\rho^2}\right) - \alpha\gamma\rho \leq 0,$$

since $\rho_{4,2}\rho - \rho_{3,2}$ and $\rho_{4,1}\rho - \rho_{3,1}$ are both negative by assumption. Further,

$$Bias(\hat{\beta}_3) - Bias(\hat{\beta}_4) = \alpha_2\gamma_{4,2}\left(\frac{\rho_{3,2}-\rho_{4,2}}{1+\rho}\right) + \alpha_1\gamma_{1,4}\left(\frac{\rho_{3,1}-\rho_{4,1}}{1+\rho}\right) \geq 0$$

since $\rho_{3,2} - \rho_{4,2}$ and $\rho_{3,1} - \rho_{4,1}$ are both positive by assumption.

C. Estimating $\rho_{g,g'}$

In panel A of Table A1, we use information on the current and previous grade in which grade 3 and grade 4 teachers were deployed to categorize them as grade teachers, class teachers and those who do not fall into either of these categories ("others"). Specifically, teachers who report having taught grade g in the current year and grade $g-1$ in the previous year are categorized as class teachers, teachers who report having taught grade g both in the current and previous year are categorized as grade teachers, and teachers who do not fit either of these patterns are categorized as "other". In Panel B, we use these numbers to calculate the share of teachers in the first sub-sample, where the student's grade g teacher taught the student already in grade g', as

$$s_g = s_{class,g} + \frac{1}{\bar{n}_{streams}}\ s_{other,g} \tag{A6}$$

In Table A2, we report the correlation coefficients $\rho_{g,g'|j(g)\neq j(g')}$; i.e., the correlation in subject differences in teacher knowledge in the second sub-sample, where the student's grade g teacher did *not* teach the student in grade g'. We calculate this quantity on the subset of pairs of teachers where the grade g' teacher is a grade teacher in schools where there is only one stream per grade. We test for (and fail to reject) equality of $\rho_{g(\tau,h,m),g'(\tau',h',m')|j(g)\neq j(g')}$ for all possible combinations of pairs of distinct teachers in grades g and g', at dates τ and τ' (i.e. current and previous) of type h (i.e. grade, class, "other") and in school m (i.e. one stream per grade or several) using a test suggested by Jennrich (1970) (see details in StataCorp, 2015).[55]

D. *Inference*

The asymptotic variance-covariance matrix of the parameters $\hat{\theta} = \{\hat{\alpha},\ \widehat{\alpha\gamma},\ \widehat{\alpha\gamma^2},\ \widehat{\alpha\gamma^3}\}$ in the structural model (5) is given by

$$V = \sigma_\epsilon^2 E(\Delta\tilde{x}'\Delta\tilde{x})^{-1}/N = \sigma_\epsilon^2/N \begin{pmatrix} \sigma^2_{\Delta x_4} & \rho_{4,3}\sigma^2_{\Delta x_3} & \rho_{4,2}\sigma^2_{\Delta x_2} & \rho_{4,1}\sigma^2_{\Delta x_1} \\ \rho_{4,3}\sigma^2_{\Delta x_3} & \sigma^2_{\Delta x_3} & \rho_{3,2}\sigma^2_{\Delta x_2} & \rho_{3,1}\sigma^2_{\Delta x_1} \\ \rho_{4,2}\sigma^2_{\Delta x_2} & \rho_{3,2}\sigma^2_{\Delta x_2} & \sigma^2_{\Delta x_2} & \rho_{2,1}\sigma^2_{\Delta x_1} \\ \rho_{4,1}\sigma^2_{\Delta x_1} & \rho_{3,1}\sigma^2_{\Delta x_1} & \rho_{2,1}\sigma^2_{\Delta x_1} & \sigma^2_{\Delta x_1} \end{pmatrix}^{-1}, \tag{A6}$$

[55] As detailed in StataCorp, 2015, the test assumes that each vector of N_j pairs of teachers, with the first teacher teaching in grade g at date τ and being of type h and the second teacher teaching in grade g' at date τ' and being of type h', is drawn from a bivariate normal distribution $MVN_2(\mu_j, \mathbf{\Sigma_j})$ with sample correlation matrix $\boldsymbol{R_j}$, for $j = 1, \dots m$. Also denote $N = \sum_{j=1}^{m} N_j$. To test for the equality of the correlation matrices across m independent samples, where each sample consists of vectors of pairs of teachers with differing values of g, h, τ and g', h', τ', the Wald test statistic is constructed as follows:

$$W = \sum_{j=1}^{m} \left\{ \frac{1}{2}\ trace(\boldsymbol{Z_j^2}) - diag(\boldsymbol{Z_j})'(\boldsymbol{I} + \bar{\boldsymbol{R}} \circ \bar{\boldsymbol{R}}^{-1})^{-1}\ diag(\boldsymbol{Z_j}) \right\}$$

where $\bar{\boldsymbol{R}} = 1/N \sum_{j=1}^{m} N_j\ \boldsymbol{R_j}$, $\boldsymbol{Z_j} = \sqrt{N_j}\bar{\boldsymbol{R}}^{-1}(\boldsymbol{R_j} - \bar{\boldsymbol{R}})$, and $\circ$ denotes the Hadamard product. As shown by Jennrich (1970), W has an asymptotic χ^2 distribution with $(m-1)k(k-1)/2$ degrees of freedom.
To test equality of the correlation matrices for all grades, dates and types of teachers, we construct the test statistic on the basis of $\boldsymbol{R_j} = \begin{bmatrix} 1 & \rho_{g(\tau,h),g'(\tau',h')|switch} \\ \rho_{g(\tau,h),g'(\tau',h')|switch} & 1 \end{bmatrix}$ $j = 1, \dots, 60$ for all possible combinations of $g(\tau,h)$ and $g'(\tau',h')$.

where σ_ϵ^2 is the error variance, $\Delta\tilde{x} = \{\Delta\tilde{x}_4, \Delta\tilde{x}_3, \Delta\tilde{x}_2, \Delta\tilde{x}_1\}$ is the matrix of de-meaned test scores (differenced across subjects) in each year, and $\sigma^2_{\Delta x_g}$ is the population variance of $\Delta x_g \; \forall g = 1,\ldots,4$.

We now show that each of the terms in the product can be written as a function of known population moments. Consider first $E(\Delta\tilde{x}'\Delta\tilde{x})$: with the assumption that the variance of test score subject differences is constant across grades, we can write this as

$$E(\Delta\tilde{x}'\Delta\tilde{x}) = \sigma^2_{\Delta x}\begin{pmatrix} 1 & \rho_{4,3} & \rho_{4,2} & \rho_{4,1} \\ \rho_{4,3} & 1 & \rho_{3,2} & \rho_{3,1} \\ \rho_{4,2} & \rho_{3,2} & 1 & \rho_{2,1} \\ \rho_{4,1} & \rho_{3,1} & \rho_{2,1} & 1 \end{pmatrix}, \tag{A6}$$

which can be estimated by replacing $\sigma^2_{\Delta x}$ and $\rho_{g,g'}$ by their sample analogues.[56]

Next consider the error variance, σ_ϵ^2. If we could observe $\Delta\tilde{x}_2$ and $\Delta\tilde{x}_1$ for student i's teachers in grade 1 and 2, then we could simply estimate this as $\widehat{\sigma_\epsilon^2} = \frac{1}{N}\sum\left(\Delta\tilde{y}_i - \hat{\alpha}\,\Delta\tilde{x}_{4i} - \widehat{\alpha\gamma}\,\Delta\tilde{x}_{3i} - \widehat{\alpha\gamma^2}\,\Delta\tilde{x}_{2i} - \widehat{\alpha\gamma^3}\,\Delta\tilde{x}_{1i}\right)^2$. However, since we cannot link students and teachers in earlier grades, we relate σ_ϵ^2 to σ_μ^2, the error variance in the reduced form model, which we can estimate.

Writing

$$\hat{\mu} = \Delta\tilde{y} - \hat{\beta}_4\,\Delta\tilde{x}_4 - \hat{\beta}_3\Delta\tilde{x}_3, \tag{A7}$$

and adding and subtracting $\epsilon = \Delta\tilde{y} - (\alpha\Delta\tilde{x}_4 + \alpha\gamma\Delta\tilde{x}_3 + \alpha\gamma^2\Delta\tilde{x}_2 + \alpha\gamma^3\Delta\tilde{x}_1)$, gives

$$\hat{\mu} = \epsilon + (\alpha - \hat{\beta}_4)\,\Delta\tilde{x}_4 + (\alpha\gamma - \hat{\beta}_3)\,\Delta\tilde{x}_3 + \alpha\gamma^2\,\Delta\tilde{x}_2 + \alpha\gamma^3\Delta\tilde{x}_1\,. \tag{A8}$$

Hence, we get

$$\begin{aligned}\operatorname{plim}\frac{\hat{\mu}'\hat{\mu}}{N} = \sigma_\mu^2 = \sigma_\epsilon^2 + B_4^2\sigma^2_{\Delta x_4} + B_3^2\sigma^2_{\Delta x_3} + \alpha^2\gamma^4\sigma^2_{\Delta x_2} + \alpha^2\gamma^6\sigma^2_{\Delta x_1} + \\ 2B_4B_3\rho_{4,3}\sigma^2_{\Delta x_3} - 2B_4\alpha\gamma^2\rho_{4,2}\sigma^2_{\Delta x_2} - 2B_4\alpha\gamma^3\rho_{4,1}\sigma^2_{\Delta x_1} - \\ 2B_3\alpha\gamma^2\rho_{3,2}\sigma^2_{\Delta x_2} - 2B_3\alpha\gamma^3\rho_{3,1}\sigma^2_{\Delta x_1} + 2\alpha^2\gamma^5\rho_{2,1}\sigma^2_{\Delta x_1},\end{aligned} \tag{A9}$$

where we have used the fact that $\operatorname{plim}\frac{1}{N}\sum\epsilon_i\,(\alpha\Delta\tilde{x}_{4i} + \alpha\gamma\Delta\tilde{x}_{3i} + \alpha\gamma^2\Delta\tilde{x}_{2i} + \alpha\gamma^3\Delta\tilde{x}_{1i}) = 0$ by assumption, and $B_4 = \alpha\gamma^2\left(\frac{\rho_{4,2}-\rho_{3,2}\rho}{1-\rho^2}\right) + \alpha\gamma^3\left(\frac{\rho_{4,1}-\rho_{3,1}\rho}{1-\rho^2}\right)$ and $B_3 = \alpha\gamma^2\left(\frac{\rho_{3,2}-\rho_{4,2}\rho}{1-\rho^2}\right) + \alpha\gamma^3\left(\frac{\rho_{3,1}-\rho_{4,1}\rho}{1-\rho^2}\right)$ are the asymptotic biases on the OLS estimates of $\hat{\beta}_4$ and $\hat{\beta}_3$.

Once again imposing $\operatorname{var}(\Delta x_g) = \operatorname{var}(\Delta x)$, the asymptotic error variance is,

$$\begin{aligned}\sigma_\epsilon^2 = \sigma_\mu^2 - \big(B_4^2 + B_3^2 + \alpha^2\gamma^4 + \alpha^2\gamma^6 + 2B_4B_3\rho_{4,3} - 2B_4\alpha\gamma^2\rho_{4,2} - \\ 2B_4\alpha\gamma^3\rho_{4,1} - 2B_3\alpha\gamma^2\rho_{3,2} - 2B_3\alpha\gamma^3\rho_{3,1} + 2\alpha^2\gamma^5\rho_{2,1}\big)\sigma^2_{\Delta x}.\end{aligned} \tag{A10}$$

This can be estimated by replacing σ_μ^2 with the mean of the sum of squared residuals, $\frac{\hat{\mu}'\hat{\mu}}{N}$, in the reduced form model, $\sigma^2_{\Delta x}$ with $\frac{\sum_i \Delta\tilde{x}_{i,4}{}^2}{N}$ (or test scores in any or all of the other

[56] Again, the assumption that $\operatorname{var}(\Delta x_g) = \operatorname{var}(\Delta x)$ for $g = 1,\ldots,4$ is not strictly necessary here, but holds in our sample and simplifies the algebra.

grades), $\rho_{4,3}$ with $\hat{\rho}$, and $\rho_{g,g'}$ with $\hat{\rho}_{g,g'}$, estimation of which is detailed in Section 4.2 and Section C of the Appendix. With an estimate of the asymptotic variance-covariance matrix, $\hat{V}$, we can then compute the usual Wald statistic of the hypothesis that the cumulative effect of teacher knowledge on student learning is zero. Second, we can compute the standard error on the persistence parameter by applying the delta method to, for instance, the ratio of $\widehat{\alpha\gamma}$ and $\hat{\alpha}$.

In all calculations, we adjust for the fact that test scores are measured half way through year 4, while the structural coefficient of interest, α, measures the effect of one year of schooling. We also adjust parametrically for clustering of the standard errors at the school level by multiplying the estimate of the asymptotic variance-covariance matrix by the Moulton factor (which equals 2.35 in our data set).

E. *Testing the effects of teacher knowledge on student learning across subjects*

Consider the level equations version of the first difference equation (4):

(A11) $$y_{i,k} = \beta_{4,k}x_{4,k} + \beta_{3,k}x_{3,k} + \beta_v v_i + \omega_i + \epsilon_{i,k}$$

(A12) $$y_{i,k'} = \beta_{4,k'}x_{4,k'} + \beta_{3,k'}x_{3,k'} + \beta_v v_i + \omega_i + \epsilon_{i,k'}$$

where v_i is a vector of non-subject specific teacher, parent and school specific components and ω_i is a student-specific (ability) component. Assume further, following Ashenfelter and Zimmerman (1997), that the potential correlation of the unobserved student effect, ω_i, with the observed inputs is given by:

(A13) $$\omega_i = \pi_{4,k}x_{4,k} + \pi_{4,k'}x_{4,k'} + \pi_{3,k}x_{3,k} + \pi_{3,k'}x_{3,k'} + \pi_v v_i + \xi_i$$

where ξ_i is assumed to be uncorrelated with the observed inputs. Substituting equation (A13) into equations (A11) and (A12) yields the following correlated random effects models

(A14) $$y_{i,k} = \beta_{4,k}^{cre}x_{4,k} + \pi_{4,k'}x_{4,k'} + \beta_{3,k}^{cre}x_{3,k} + \pi_{3,k'}x_{3,k'} + (\beta_v+\pi_v)v_i + \bar{\epsilon}_{i,k}$$

(A15) $$y_{i,k'} = \pi_{4,k}x_{4,k} + \beta_{4,k'}^{cre}x_{4,k'} + \pi_{3,k}x_{3,k} + \beta_{3,k'}^{cre}x_{3,k'} + (\beta_v+\pi_v)v_i + \bar{\epsilon}_{i,k'}$$

where $\beta_{t,k}^{cre} = \beta_{t,k} + \pi_{t,k}$ and $\bar{\epsilon}_{i,k} = \epsilon_{i,k} + \xi_i$. The restriction implied by the fixed effect model; i.e. the first-differenced representation given in equation (5), that the (reduced form) effects of teacher content knowledge on student content knowledge, in a given grade, are the same across subjects ($\beta_{4,k} = \beta_{4,k'}$) and ($\beta_{3,k} = \beta_{3,k'}$), can be directly tested after estimating the system (A14)-(A15). That is, testing $\beta_{t,k}^{cre} - \pi_{t,k} = \beta_{t,k'}^{cre} - \pi_{t,k'}$, provides a test of $\beta_{t,k} = \beta_{t,k'}$.

As reported in Table A3, we cannot reject the hypotheses that the effects of teacher content knowledge is the same in the two subjects for both grade 3 teachers, $\beta_{3,math} = \beta_{3,language}$, and grade 4 teachers, $\beta_{4,math} = \beta_{4,language}$, thus providing support for our fixed effect specification.

F. *Reducing misallocation*

We consider reallocation at the district level (of which reallocation at the school level is a special case). Let there be T_l teachers with knowledge below the q'th percentile in $o_1, \ldots, o_m$ origin schools and T_h teachers with knowledge above the q'th percentile in

$d_1, \ldots, d_n$destination schools. The set of origin and destination schools may or may not overlap depending on the distribution of teacher knowledge across schools in the district.

Denote the flow of students from teacher l_r in origin school o_i, where $r = 1 \ldots T_{l,o_i}$ to teacher h_s in destination school d_j, where $s = 1 \ldots T_{h,d_j}$, by x_{l_r,o_i,h_s,d_j} .

An optimal allocation of students is given by the vector x, with typical element x_{l_r,o_i,h_s,d_j}, $r = 1 \ldots T_{l,o_i}$ $i = 1 \ldots m$, $s = 1 \ldots T_{h,d_j}$, $j = 1 \ldots n$, chosen to maximize the per-capita effect on student learning.

We now consider each of the effects listed in section 5.4 and how many students are exposed to it. In finding the optimal allocation, we consider only flows from teachers with knowledge below the qth percentile to those above (that is, we ignore flows in the other direction that might potentially achieve a more optimal allocation of class sizes). In calculating per-capita gains, we assume that teachers are a random sample from the school, that each teaches one classroom and that the pupil-teacher ratio (reported as the school level average in our data) is constant across class rooms in the school.

First, the effect of moving $x_{l_{r\prime},o_{i\prime},h_{s\prime},d_{j\prime}}$ from teacher $l_{r\prime}$ in origin school $o_{i\prime}$ to teacher $h_{s\prime}$ in destination school $d_{j\prime}$ is

(A16) $$x_{l_{r\prime},o_{i\prime},h_{s\prime},d_{j\prime}} \times \alpha \left(K_{h_{s\prime},d_{j\prime}} - K_{l_{r\prime},o_{i\prime}}\right)$$

where $K_{h_{s\prime},d_{j\prime}} - K_{l_{r\prime},o_{i\prime}}$ is the difference in knowledge between the two teachers and α is the causal effect of exposing students to a more knowledgeable teacher for one year.

When assessing the effect of changes in class size, we assume that the effect, if any, is linear in % changes of class size (see Muralidharan and Sundararaman, 2013). The total inflow to the classroom of teacher $h_{s\prime}$ in destination school $d_{j\prime}$ is $\sum_{i=1}^{m} \sum_{r=1}^{T_{l,o_i}} x_{l_r,o_i,h_{s\prime},d_{j\prime}}$ (i.e potential non-negative inflows from all teachers below the qth percentile in all origin schools) and hence the new class size in her classroom is $\sum_{i=1}^{m} \sum_{r=1}^{T_{l,o_i}} x_{l_r,o_i,h_{s\prime},d_{j\prime}} + PTR(d_{j\prime})$. For the teacher's existing students, learning is therefore reduced by

(A17) $$\sigma\left(\log\left(\sum_{i=1}^{m} \sum_{r=1}^{T_{l,o_i}} x_{l_r,o_i,h_{s\prime},d_{j\prime}} + PTR(d_{j\prime})\right) - \log\left(PTR(d_{j\prime})\right)\right) \times PTR(d_{j\prime})$$

For the students who move to this teacher from origin school i, the class size effect is:

(A18) $$\sigma\left(\log\left(\sum_{i=1}^{m} \sum_{r=1}^{T_{l,o_i}} x_{l_r,o_i,h_{s\prime},d_{j\prime}} + PTR(d_{j\prime})\right) - \log(PTR(o_i))\right) \times \sum_{r=1}^{T_{l,o_i}} x_{l_r,o_i,h_{s\prime},d_{j\prime}},$$

which takes into account that class sizes in origin school o_i may differ from class sizes in destination schools d_j and that there are students being reallocated from T_{l,o_i} teachers in origin school o_i.

Hence, the total class size effect on students (existing ones and new ones from the m origin schools) in the classroom of teacher $h_{s\prime}$ in destination school $d_{j\prime}$ is

(A19) $$\sigma \sum_{i=1}^{m} \left(\log\left(\sum_{i=1}^{m} \sum_{r=1}^{T_{l,o_i}} x_{l_r,o_i,h_{s\prime},d_{j\prime}} + PTR(d_{j\prime})\right) - \log(PTR(o_i))\right) \times \sum_{r=1}^{T_{l,o_i}} x_{l_r,o_i,h_{s\prime},d_{j\prime}}$$
$$+ \sigma\left(\log\left(\sum_{i=1}^{m} \sum_{r=1}^{T_{l,o_i}} x_{l_r,o_i,h_{s'},d_{j'}} + PTR(d_{j'})\right) - \log\left(PTR(d_{j'})\right)\right) \times PTR(d_{j'})$$

Second, the effect of removing students from the classroom of teacher $l_{r\prime}$ in origin school $o_{i\prime}$ is

$$\text{(A20)}\qquad \sigma\left(\log\left(PTR(o_{i\prime}) - \sum_{j=1}^{n}\sum_{s=1}^{T_{h,d_j}} x_{l_{r\prime},o_{i\prime},h_s,d_j}\right) - log\left(\ PTR(o_{i\prime})\right)\right)\times \left(PTR(o_{i\prime}) - \sum_{j=1}^{n}\sum_{s=1}^{T_{h,d_j}} x_{l_{r\prime},o_{i\prime},h_s,d_j}\right)$$

Finally, in maximizing the per-capita gain, we need to respect that outflows from the classroom of teacher $l_{r\prime}$ in origin school $o_{i\prime}$ cannot exceed the existing class size and must be non-negative.

Combining all these effects, we can write down the following Lagrangian:

$$\max_{x,\lambda}\ L = \alpha\ \sum_{j=1}^{n}\sum_{s=1}^{T_{s,d_j}}\sum_{i=1}^{m}\sum_{r=1}^{T_{l,o_i}} x_{l_r,o_i,h_s,d_j}\times\left(K_{h_s,d_j} - K_{l_r,o_i}\right) +$$

$$\sigma\ \sum_{j=1}^{n}\sum_{s=1}^{T_{s,d_j}}\sum_{i=1}^{m}\left(\log\left(\ \sum_{i=1}^{m}\sum_{r=1}^{T_{l,o_i}} x_{l_r,o_i,h_s,d_j}\ + PTR(d_j)\right) - \log\left(PTR(o_i)\right)\right)\times\sum_{r=1}^{T_{l,o_i}} x_{l_r,o_i,h_s,d_j}$$

$$+\sigma\ \sum_{j=1}^{n}\sum_{s=1}^{T_{s,d_j}}\left(\log\left(\sum_{i=1}^{m}\sum_{r=1}^{T_{l,o_i}} x_{l_r,o_i,h_s,d_j}\ + PTR(d_j)\right) -\ \log\left(PTR(d_j)\right)\right)\times\ PTR(d_j)$$

$$+\sigma\sum_{i=1}^{m}\sum_{r=1}^{T_{l,o_i}}\left(\log\left(PTR(o_i) - \sum_{j=1}^{n}\sum_{s=1}^{T_{h,d_j}} x_{l_r,o_i,h_s,d_j}\right) - \log\left(\ PTR(o_i)\right)\right)\times \left(PTR(o_i) - \sum_{j=1}^{n}\sum_{s=1}^{T_{h,d_j}} x_{l_r,o_i,h_s,d_j}\right)$$

$$-\sum_{i=1}^{m}\sum_{r=1}^{T_{l,o_i}}\lambda_{l_r,o_i}\left(\sum_{j=1}^{n}\sum_{s=1}^{T_{h,d_j}} x_{l_r,o_i,h_s,d_j}\ - PTR(o_i)\right)$$

Ignoring integer constraints, the first order condition with respect to $x_{l_{r\prime},o_{i\prime},h_{s\prime},d_{j\prime}}$ for an interior solution is

$$\text{(A21)}\qquad \alpha\times\left(K_{h_s,d_j} - K_{l_r,o_i}\right) + \sigma\left(\log\left(\ \sum_{i=1}^{m}\sum_{r=1}^{T_{l,o_i}} x_{l_r,o_i,h_{s'},d_{j'}}\ + PTR(d_{j'})\right) - \log PTR(d_{j'})\ - \log\left(PTR(o_{i'}) - \sum_{j=1}^{n}\sum_{s=1}^{T_{h,d_j}} x_{l_{r'},o_{i'},h_s,d_j}\right) - \log PTR(o_{i'})\right) = 0\ .$$

G. Projected impacts 2016-2050 of increasing teacher knowledge and time spent teaching

In 2009, it was estimated that there were almost 2.8 million primary school teachers in Sub Saharan Africa (UIS, 2011). Estimates for 2018 are not available but available evidence suggest that the number of teachers has increased to at least 3.0 million by 2015 (UNICEF, 2015). Assuming a constant growth rate over the period (2009-2018), we predict there to be 3.1 million teachers by 2018.

Using the SDI data, we derive an age distribution of teachers. 45% (or 1,395,000 teacher assuming there are 3.1 million primary school teachers in 2018) of the teachers were hired in the last decade. 697,500 teachers were hired in the decade before that; 558,000 teachers hired in the decade before that, and 449,500 teachers in the decade before that. We use these numbers and the assumption that teachers are working for 40 years on average to estimate the number of teachers that will retire in each 10-year period starting in 2019. We smooth the resulting data to derive annual retirement numbers for 2019-2050.

Population projections from UNICEF is used to predict the number of primary school aged children from 2019-2050. UNICEF (2014) report estimated and predicted numbers of births per year for Sub Sahara Africa for years 1980, 2015, 2030, and 2050. We extrapolate the number of births per year from 1980-2050 using the estimates and projections for 1980, 2015, 2030, and 2050.

The under-five mortality in the region fell from about 0.150 in 2000 to 0.080 in 2015 (World Bank. 2017). We assume that mortality will continue to fall to 0.050 in 2050. Combining data on number of births with under-five mortality rates, we can project the number of primary school-aged children from 2015-2050.

UIS (2018) report an out-of-school rate for primary school age children in Sub-Saharan Africa of 20.8 percent. Using that estimate and the number of teacher and number or primary school age children in 2018 (as described above), gives an estimated student-teacher ratio of 42.1, which is close to the average ratio, weighted by population, reported by the World Bank (2019). We assume the student-teacher ratio remains at 42.1, while the out-of-school rate falls continuously to 10 percent by 2050. With these assumptions we can estimate the number of teachers to be hired a given year t as the number of primary school age children in school at t divided by the student-teacher ratio at time t, minus the difference between the stock of teachers at the beginning of $t-1$ and the number of teachers that retire at the end of the same year.

We assume initially that all existing teachers are uniformly distributed over grades and that new teachers are added to replace retiring teacher and to ensure that the student-teacher ratio in each grade and year is equal to 42.1.

Table A1: The share of teachers in grade 3 and 4 who have switched since grade 1 and 2

	(1) Grade 3	(2) Grade 4
Panel A: Transition		
Grade	52%	52%
Class	32%	24%
Other	16%	25%
Panel B: Share who remain with their class		
s_g	37%	31%
N	1,126	2,462

Note: Panel A reports the share of teachers in grade 3 and grade 4 categorized as grade, class and "other". Panel B reports the resulting share of teachers who remained with their class between grade $g = 3, 4$ and $g' = 1, 2$.

Table A2: The correlation in teacher knowledge for teachers who are not the same

	IRT		*Effective years of education*	
	(1) Δx_3	(2) Δx_4	(3) Δx_3	(4) Δx_4
Panel A: $\rho_{g,g' \mid j(g)\neq j(g')}$				
Δx_1	0.10	0.15	0.23	0.10
Δx_2	0.34	0.18	0.08	0.07
Panel B: Observations				
Δx_1	59	88	34	73
Δx_2	44	63	22	38
Panel C:	$\rho_{g(\tau,h,m),g'(\tau',h',m')\mid j(g)\neq j(g')} = \bar{\rho}_{j(g)\neq j(g')}$			
χ^2-statistic	133.03		132	
Degrees of freedom	120		120	
p-value	[0.18]		[0.20]	

Note: The left-hand side of the table presents calculations for teacher knowledge transformed using Item Response Theory and then differenced across subjects. The right-hand side presents the same statistics for subject-differences in teacher knowledge, transformed into effective years of education and then differenced across subjects. Panel A reports the correlation in subject-differences in teacher knowledge for pairs of teachers where the grade $g' = 1,2$ teacher is a grade teacher. Panel B reports the number of observations in each category. Panel C tests for equality of $\rho_{g(\tau,h,m),g'(\tau',h',m')\mid j(g)\neq j(g')}$ across all possible values of $g(\tau, h, m)$ and $g'(\tau', h', m')$.

Table A3: Relationship between student and teacher content knowledge: CRE model

Dep. variable	(1) IRT test scores Mathematics	(2) IRT test scores Language	(3) Effective years of schooling Mathematics	(4) Effective years of schooling Language
Implied $\beta_{4,subject}$	0.052**	0.075**	0.022	0.035**
	(.023)	(.027)	(.016)	(.018)
Implied $\beta_{3,subject}$	0.029	0.051*	0.050***	0.034**
	(.023)	(.026)	(.015)	(.017)
$\chi^2(\beta_{4,k} = \beta_{4,k'})$	0.67		0.36	
	[.41]		[.55]	
$\chi^2(\beta_{4,k} = \beta_{4,k'})$	0.56		0.54	
	[.46]		[.46]	
$\chi^2(\pi_{4,k} = \pi_{4,k'})$	1.42		3.57*	
	[.23]		[.06]	
$\chi^2(\pi_{3,k} = \pi_{3,k'})$	0.10		0.08	
	[.75]		[.78]	
Observations	12,600		13,752	
Number of schools	939		1,024	

Note: Estimates from the correlated random effects model, equations (A15)-(A16). Dependent variable: Columns (1)-(2): student test score (IRT rescaled) in mathematics and language, respectively; Columns (3)-(4): student test score (Effective years of schooling) in mathematics and language, respectively. Regressions in the two subjects are estimated by seemingly unrelated regressions (SUR). Implied β is the effect of the teacher test score; i.e. $\beta_{t,k}^{cre} - \pi_{t,k} = \beta_{t,k'}^{cre} - \pi_{t,k'}$, for $t = 4, 3$ and $k = \{$*mathematics, language*$\}$ (see appendix E for details). χ^2s are the test statistics for the null hypotheses that the effects of teacher content knowledge is the same in the two subjects for a given grade ($\beta_{t,k} = \beta_{t,k'}$) and that ($\pi_{t,k} = \pi_{t,k'}$), with p-values in brackets. Regressions include controls for student gender, student age, teacher gender, teacher experience, teacher university degree, and school infrastructure. Clustered, by school, standard errors in parenthesis. *** 1% , ** 5% , * 10% significance.

Table A4: Variance ratio test

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Panel A: Teacher test score: IRT scaled				*Panel B: Teacher test score: Effective years of education*			
	Δx_1	Δx_2	Δx_3	Δx_4	Δx_1	Δx_2	Δx_3	Δx_4
Panel A. Summary statistics								
Mean	0.146	0.229	0.027	0.014	0.124	0.396	-0.217	0.027
Std dev	0.908	0.883	0.886	0.924	2.063	2.209	2.129	2.198
Obs	241	139	756	965	241	139	756	965
Panel B: Equality of variance test								
Δx_1								
Obs			997	1,206			997	1,206
F-test statistic			1.053	0.966			0.939	0.881
			[.611]	[.751]			[.563]	[.229]
Δx_2								
Obs			895	1,104			895	1,104
F-test statistic			0.995	0.913			1.076	1.009
			[.995]	[.507]			[.555]	[.917]
Δx_3								
Obs				1,721				1,721
F-test statistic				0.918				0.938
				[.213]				[.357]

Note: Panel A: Summary statistics of teacher content knowledge in grade 1 through 4. Panel B: Tests on the equality of standard deviations (variances) of $\Delta x_{g'}$ and Δx_g. F-test statistic is the test statistic for the for the null hypotheses that the ratio of the standard deviation (variance) of the current and prior teacher knowledge are the same, with p-values in brackets.

Table A5: Relationship between student and teacher knowledge (alternative knowledge measure)

	(1)	(2)	(3)	(4)	(5)
Dep. Variable	Effective year of schooling				
Effective years of education of current teacher	0.087***	0.031***	0.044***	0.027**	0.031*
	(.009)	(.008)	(.012)	(.013)	(.018)
Effective years of education of prior teacher				0.047***	0.046***
				(.012)	(.016)
Language	0.148***	0.165***	0.221***	0.167***	0.216***
	(.019)	(.018)	(.036)	(.026)	(.033)
Constant	2.062***	1.401***	1.119***	1.221***	1.029***
	(.065)	(.029)	(.036)	(.042)	(.042)
Lower bound (total effect)				0.074***	0.077***
				(.013)	(.018)
Observations	30,361	30,361	15,220	17,294	8,969
Adj. R-squared	0.136	0.031	0.024	0.048	0.034
Number of schools	1,974	1,974	905	1,503	626
Number of students	16,922	16,922	7,642	10,324	4,503
Country FE	x				
Student FE		x	x	x	x
Same teacher in both subjects			x		x

Note: Dependent variable: Effective years of schooling. Lower bound (total effect) is the test of the sum of the estimated coefficients on the effective years of education of current and prior teacher. Clustered, by school, standard errors in parenthesis. *** 1% , ** 5% , * 10% significance.

Table A6: Specification tests (alternative knowledge measure)

	(1)	(2)	(3)	(4)	(5)
Dep. variable	Effective years of schooling: Students				
Effective years of education of current teacher	0.031*	0.033*	0.030*	0.025*	0.021
	(.018)	(.018)	(.018)	(.013)	(.019)
Effective years of education of previous teacher	0.046***	0.039**	0.049***	0.048***	0.050***
	(.016)	(.017)	(.016)	(.012)	(.017)
Teacher pedagogy score				0.239	
				(.175)	
Constant	-0.216***	-0.216***	-0.216***	-0.165***	-0.186***
	(.033)	(.030)	(.033)	(.026)	(.036)
Lower bound (total effect)	0.077***	0.072***	0.079***	0.073***	0.071***
	(.018)	(.018)	(.018)	(.013)	(.019)
Specification/Sample	Main	District FE	Urban FE	Student FE	One gr. 4 class
Same teacher in both subjects	X	X	X		X
Observations	4,466	4,466	4,466	6,970	3,886
Number of schools	619	619	619	1037	534

Note: Dependent variable: Effective years of schooling. Lower bound (total effect) is the test of the sum of the estimated coefficients on the effective years of education of current and prior teacher. First difference (across subjects) specification, with clustered, by school, standard errors in parenthesis. Specification: (1) main specification with the sample of students with the same teacher in language and mathematics in a given year (grade 3 and 4); (2) main specification with subject variant district fixed effects; (3) main specification with subject variant urban fixed effects; (4) sample of all students with data on current and previous teacher score and current teacher pedagogy score; (5) main specification on sample of schools with one grade 4 classroom. *** 1% , ** 5% , * 10% significance.

Table A7: Placebo tests (alternative knowledge measure)

	(1)	(2)	(3)	(4)
Dep. variable	Effective years of schooling: Students			
Effective years of education of current teacher	0.015	0.021	0.067**	0.039***
	(0.021)	(0.020)	(0.026)	(0.013)
Effective years of education of previous teacher	0.082**	0.087***		
	(0.023)	(0.022)		
Effective years of education of higher grade teachers	0.022			
	(0.022)			
Constant	-0.373***	-0.374***	-0.124**	-0.250***
	(0.066)	(0.066)	(0.056)	(0.027)
Lower bound (total effect)	0.097***	0.108***		
	(0.028)	(0.025)		
Specification/Sample	Placebo	Comparison	Same class teacher in grade 3 & 4	Different class teachers in grade 3 & 4
Observations	1,669	1,669	1,814	5,764
Number of schools	217	217	283	746

Note: Dependent variable: Effective years of schooling. First difference (across subjects) specification, with clustered, by school, standard errors in parenthesis. Specification: (1) main specification controlling for teacher content knowledge (effective years of education) of higher grade teachers in school; (2) main specification using the same sample as in column (2); (3) Sample of students with the same teacher in both subjects in both years (grade 3 and grade 4); (4) Sample of students with a new teacher teaching both subjects in grade 4. *** 1% , ** 5%, * 10% significance.

Table A8: Correlation in teacher knowledge (subject-differenced) for student i's teachers in grade 3 and 4 and teachers in lower grades (alternative knowledge measure: effective years of education)

	(1)	(2)
$\rho_{g,g\prime}$	Grade 3	Grade 4
Grade 1	0.51	0.38
Grade 2	0.42	0.36
Grade 3	-	0.51

Note: The table reports $\rho_{g,g\prime}$, the correlation in teacher knowledge (subject-differenced) for teachers in grade $g = 3,4$ and $g' = 1,2,3$. For $g' = 1,2$, the correlation coefficient is estimated as $\rho_{g,g\prime} \approx s_g \times 1 + (1 - s_g) \times \rho_{g,g'|j(g)\neq j(g')}$ using the inputs in Table A1 and A2. For $g' = 3$, it is estimated from the variance-covariance matrix of the regressors in equation (5).