The Lost Human Capital

Teacher Knowledge and Student Achievement in Africa*

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

April 2018

In many low income countries, teachers do not master the subject they are teaching and children learn little from attending school. Using unique data from nationally representative surveys from seven Sub-Saharan African countries, we propose a methodology to assess the effect of teacher knowledge on student learning when panel data on students are not available. We show that data on test scores of the current and the previous year's teachers allows us to estimate a lower bound on the cumulative effect of teacher knowledge on student achievement. With further restrictions on the cumulative student achievement function we can also estimate bounds on both the contemporaneous effect of teacher content knowledge and the extent of fade out of the teachers' impact in earlier grades. We use these structural estimates to answer two questions. To what extent can shortfalls in teachers' content knowledge account for the large learning gap observed across countries? How much learning is lost because of misallocation?

* We are grateful to the many researchers, survey experts and enumerators who have worked on the country surveys that made this paper possible. 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, Oxford University, UPF, 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) 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, ezequielmolina@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. Four out of five students in Mozambique and Nigeria, for example, after more than three years of compulsory language teaching, cannot read a simple word of Portuguese and English, respectively. In India, only one in four fourth grade student manages tasks—such as basic subtraction—that is part of the curriculum for the second grade and roughly half of the students in Uganda, after three years of mathematics teaching, cannot order numbers between 0-100.[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. Less is known, however, about what specific dimensions of teacher quality matter.

In this paper we investigate the role of teachers' knowledge of the subject they teach, using unique cross sectional data collected from nationally representative surveys in Kenya, Nigeria, Mozambique, Senegal, Tanzania, Togo, and Uganda, which together represent close to 40 percent of the region's total population.

Our first contribution is primarily a methodological one. We show that data on test scores of the current and the previous year's teachers, under weak conditions, allows us to estimate a lower bound on the cumulative effect of teacher knowledge on student achievement, even when panel data on students are not available. With further restrictions on the cumulative student achievement function, we also show how one can estimate bounds on both the contemporaneous effect of teacher content knowledge and the extent of fade out of the teachers' impact in earlier grades.

Our second contributions is empirical. We show that teacher 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 0.07 SD in the short run (after one year of teaching). This implies that moving a student from the 5th to the 95th percentile of the teacher content knowledge distribution increases student learning by 0.23 SD in one year. These effects, however, fade out relatively fast over time, with approximately 50 percent of the short-run effect persisting between grades, yielding a persistence parameter similar to those estimated using value-added models (see Kane and Staiger 2008; Jacob, Lefgren, and Sims 2010;

[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).

Rothstein 2010; and Andrabi et al., 2011). As a result, being taught by a teacher with a 1 SD better content knowledge, over four years; roughly doubles the short run effect.

We use these structural estimates to answer two questions. First, to what extent can shortfalls in teachers' content knowledge account for the large learning gap observed across countries? Second, how much learning is lost because of misallocation, more precisely, that students are not allocated to the best teachers up to the point where the marginal gain from increased learning equals the marginal cost from increased class sizes? Finally, we consider the longer run implications of two complementary policy reforms.

Overall, our result highlights the huge shortcomings in teacher quality in Africa and the lost human capital as a consequence. We find that raising teacher content knowledge to the lower bar for primary teachers in Africa would in itself, holding teacher effort and pedagogical skills constant (and currently at low levels), reduce the observed effective education gap after four years by around one-third. Comparing the lowest performing countries in our sample in terms of student knowledge, to the highest performing, around 20 percent of the gap is explained by differences in teachers' knowledge. We further show that while the learning gains from reallocating students may not be so large, there are sizeable inefficiencies in terms of spending even when the class size effects are relatively large.

Our work is related to a growing literature on the impact of teacher quality. Several studies, primarily from the U.S, but more recently also from middle income countries (Ecuador and Pakistan), demonstrate the importance of teachers using a value-added approach, with effect sizes; i.e. the effect on student performance from a one standard deviation improvement in teacher value added, ranging from 0.1-0.2SD (Rockoff, 2004; Rivkin et al., 2005; Aaronson et al., 2007; Chetty et al., 2014b; Araujo et al., 2016; and Bau and Das, 2017). Our results confirm the importance of teachers 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 than in previous studies, and especially studies from the U.S. Our work also complements the literature in two 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 (a component of) teacher quality using cross-sectional data. While the availability of administrative databases that track individual student achievement over time, and link students to their teachers, is becoming increasingly more common in the US and other high income countries, and thus allows the estimation of, for

example, value-added models, the relative scarcity of such data in developing countries highlights the need also for less data demanding approaches.

We use within-student within-teacher variation to identify the structural estimates empirically. The within-student across subject comparisons was introduced by Dee (2005; 2007). The within-student within-teacher across subjects comparison was proposed by Metzler and Woessmann (2012). Specifically, Metzler and Woessmann (2012) use a contemporaneous specification relating grade 6 student achievement test score measure to contemporaneous teacher score measure using Peruvian data. Bietenbeck et al. (2017) use the same approach in a multi-country setting (13 Sub-Saharan African countries spanning years 2000 to 2007). As discussed in Todd and Wolpin (2003), in order to obtain consistent estimates of the impact of teacher content knowledge using as contemporaneous specification one needs to assume either that only contemporaneous teacher knowledge matter to the production of current achievement or that there is no correlation between the current teacher's knowledge and past teachers' knowledge. We can relax both of these assumptions by exploiting the fact that we have data for both the current and previous teacher. Thus, 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.

Our findings complement a growing experimental literature on the impact of teacher quality on student performance.[2] Here we focus on teacher content knowledge, noting that there is a dearth of well-identified studies on the impact of teacher knowledge on learning outcomes in developing countries (see Glewwe and Muralidharan, 2015).

To estimate the cumulative effect of teacher knowledge on student achievement, we use newly collected representative micro data from several Sub-Saharan African countries. 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) for details). 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 here is collected using one-on-one tests (for students) and test scores (for teachers) derived from mock student tests marked by teachers.

We proceed by first providing a short description of the data we use. In Section 3, we turn to providing summary statistics on both student and teacher content knowledge. In Section 4 we present a statistical model of cumulative knowledge acquisition, show how to

[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).

estimate the key structural parameters, and discuss how we identify these parameters empirically. Section 5 presents reduced form evidence followed by the main results. In addition, we present specification and placebo tests. Finally, Section 6 concludes with a short discussion of the implications of our findings.

2. Data and context

We use data from the Service Delivery Indicators (SDI)—an ongoing Africa-wide program with the aim of collecting 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, most clearly (and perhaps most damagingly) manifested at the school level, in fast-expanding systems of education.

To date, the SDI program has collected data, including from the two pilot countries, from a total of seven countries (eight surveys): Kenya (2012), Mozambique (2014), Nigeria (2013), Senegal (2010), Tanzania (2010, 2014), Togo (2013), and Uganda (2013). 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 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 in Sub-Saharan Africa.[3]

The surveys collected 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 were 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 core quality dimension: Time spent teaching, teachers' knowledge of the subject they are teaching, and teachers' pedagogical skills. Table 1 reports summary statistics on time spent teaching and teachers' pedagogical skills. Teachers, on average, are absent from class 44% of the time and about half of that classroom absence is

[3] See Bold et al. (2017) for details of the sample.

due to teachers not at all being at the school during regular teaching hours. 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 one in ten teachers deemed to have minimum pedagogy knowledge, and even fewer teachers are judged to properly manage to assess students learning progression and shortcoming.

In each school, ten students were 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 were selected for testing.[4] Overall, in more than 80 percent of the schools surveyed (and for 65 percent of the students), data was collected on both the current and previous teachers, i.e., the teacher 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 in order to build up a differentiated picture of students' cognitive skills; i.e., oral one-to-one testing allows one to 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 language test, which evaluated ability in English (Kenya, Nigeria, Tanzania, and Uganda), French (Senegal and Togo), or Portuguese (Mozambique), ranged from simple tasks that tested letter and word recognition to a more challenging reading comprehension test. The mathematics test ranged 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 spanned items from the first four years of the curriculum.[5]

[4] In five of eight surveys, teachers at higher grades were also sampled.

[5] 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.

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). Both the student and the teacher test have good reliability, with a reliability ratio (estimated by Cronbach's alpha) above 0.8 in both subjects on the student side and above 0.85 in both subjects on the teacher side.[6]

3. Teacher and student knowledge

The raw test scores were rescaled using item response analysis (IRT).[7] 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. We also report results using a complementary test score measure, labeled effective years of schooling for students (effective years of education for teachers). As with the IRT scores, this complementary test score

[6] 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.

[7] 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.

measure scales the raw scores by difficulty. The difference is that while the IRT methodology is essentially a data-driven approach which classifies a question as easy or difficult on the basis of how many teachers (or students) were able to answer it, the effective years of schooling (education) classify a question as easy or difficult based on where on the curriculum it is located.[8]

This alternative test score measure has some advantages. First, the transformed data points are informative in themselves, although since the tests were designed to maximize the precision over a smaller range of abilities in each grade, rather than assessing a wide range of skills associated with each grade level, the measure by construction should be viewed at best as a proxy measure of students/teachers grade level competencies. Second, the transformation allows us compare students and teachers using the same scale, although it is important then to keep in mind that the method of assessment differed. Finally, and maybe most importantly, it allows one to extrapolate beyond effective years of schooling observed in the sample in a meaningful way—an advantage we exploit for the structural simulations in section 5.4.

Table 2 (for students) and Table 3 (for teachers) shows how the scores constructed with item response theory line up with the curriculum-scaled years of schooling measure. There is a strong positive correlation between the two scores, above 90% in both cases. Table 2, and Figure 1, further show the percentage of students at each effective year of schooling level for language and mathematics. 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. Comparing across countries, the average student in Kenya (the top performer) has acquired two and a half years of effective years of schooling after three and a half years of schooling, while the average student in Mozambique (the bottom performer) has acquired only 0.4 years of curriculum-scaled years of schooling (see Figure 2).

Breaking down the summary score, a quarter of fourth grade students have not acquired any effective years of schooling and one-third have not acquired any effective years of schooling in mathematics. 15% and 21% have acquired one year of curriculum-scaled

[8] 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 not 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. Note also that our empirical specification used to identify the structural estimates uses variation across students.

years 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.

In Table 3 and Figure 3, we show the percentage of teachers at each quality adjusted year of education. On average, teachers have 3.8 years of quality adjusted education in language and 4.1 years of quality adjusted education in mathematics.[9] Again there are large differences across countries, with Kenyan mathematics teachers (on average) having 5.7 effective years of education (top performer) and mathematics teachers in Togo having only 1.9 years of quality adjusted education (see Figure 4).

4. A statistical model

4.1. Set-up

In this section we lay out a statistical model for cognitive achievement that assumes that children's achievement, as measured by test performance after t years of school, is the outcome of a cumulative process of knowledge acquisition.

We first present the model in a general form and show that with data on test scores of the current and the previous year's teachers, we can, under mild restrictions, estimate a lower bound on the cumulative effect of teacher knowledge on student achievement, even when panel data on students are not available. We then place further restrictions on the model in order to estimate 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 identify these parameters empirically and how we make inference about the structural parameters of interest.

Let $y_{ijt,k}$ be student i's achievement in school j after t years of schooling (or in grade t, in subject k. As in Todd and Wolpin (2003) we view knowledge acquisition as a production process in which current and past inputs are combined with an individual's innate ability (or motivation), denoted as ω_{ij}, to produce a cognitive outcome that can be measured. These inputs include school-supplied inputs (S), parent-supplied inputs (P), and teacher-

[9] 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 4.7 years. Nevertheless, we use the stricter measure here as it makes for a more transparent definition of the teacher test score.

supplied inputs (T), some of which could vary by subject and some of which do not. Specifically, let $S_{ijt,k} = \{s_{ijt,k}, \bar{s}_{ijt}\}$ and $P_{ijt,k} = \{p_{ijt,k}, \bar{p}_{ijt}\}$; i.e., each input vector consists of a vector of subject-specific and subject-invariant inputs, the latter indicated by a bar over the variable. Further, assume the vector of teacher-supplied inputs is $T_{ijt,k} = \{x_{ijt,k}, c_{ijt,k}, \bar{x}_{ijt}, \bar{c}_{ijt}\}$, where $x_{ijt,k}$ is the subject content knowledge of the teacher teaching student i in school j at t (or grade t) in subject k, $c_{ijt,k}$ is a vector of other subject-by-teacher characteristics/skills, and $\bar{x}_{ijt}$ and $\bar{c}_{ijt}$ are the corresponding subject-invariant terms.

Let $T_{ij,k}(t)$, $S_{ij,k}(t)$, $P_{ij,k}(t)$ denote input histories up to t years of schooling (or for students in grade t), then allowing for measurement error in test scores, denoted by $\varepsilon_{ijt,k}$, the cumulative student achievement function, after t years of schooling (or students in grade t) is

$$y_{ijt,k} = F\left[T_{ij,k}(t), S_{ij,k}(t), P_{ij,k}(t), \omega_{ij}, \varepsilon_{ijt,k}\right]. \quad (1)$$

Equation (1) highlights the two main problems we face in estimating the causal (cumulative) effect of teacher content knowledge on student learning. First, students' innate ability, and several school and parent-supplied inputs, are inherently unobservable and may be correlated with x_{ij}, for instance if better students sort into schools with better teachers. Second, student achievement in grade (or year) $t_n > 1$ is a function of the whole history of the quality of teaching; i.e., x_n, x_{n-1},…x_1. Below we describe how we deal with these two issues.

4.2. The cumulative effect of teacher knowledge on student learning

We start by linearizing the production function and rewriting it in first-differences. Implicitly we then assume that the coefficients on all inputs are subject-invariant; i.e., a one unit increase in teacher content knowledge in subjects k and k', has the same marginal effect on student test scores in subject k as in subject k'.[10] This yields the following specification for fourth year students,

$$\Delta y_{ij4} = y_{ij4,k} - y_{ij4,k'} = \beta_0 + \beta_4 \Delta x_{ij4} + \beta_3 \Delta x_{ij3} + \beta_2 \Delta x_{ij2} + \beta_1 \Delta x_{ij1} + \epsilon_{ij} \quad (2)$$

where $\Delta x_{ijt} = x_{ijt,k} - x_{ijt,k'}$ is the difference in teacher subject content knowledge across the two subjects, and where the error term ϵ_{ij} subsumes all the remaining student subject-specific unobservable inputs in the vectors S, P, and T.

We assume that the effect of teacher knowledge applied at time t on student achievement measured at time t' depends the size of the contemporaneous effect and the time

[10] As reported in section 5, we cannot reject the hypothesis that the effect is the same in the two subjects.

between the teaching was performed and the measurement of achievement. Specifically, let α_t denote the contemporaneous effect of teacher knowledge in grade t; i.e., as the effect of teacher content knowledge on student achievement at the end of grade t of one year of teaching in grade t. Further let $\gamma_{t,t'} \leq 1$ denote the degree of persistence of teaching taking place at time t and measured at t'. That is, if α_{t-1} is the marginal effect of teacher content knowledge in grade $t-1$ on student achivement at the end of the same grade, then $\gamma_{t-1,t}\alpha_{t-1}$ is the marginal effect of teacher content knowledge in grade $t-1$ on student achivement at the end of grade t.

With these assumptions, we can rewrite (2) as

$$\Delta y = \alpha_0 + \alpha_4 \Delta x_4 + \alpha_3 \gamma_{3,4} \Delta x_3 + \alpha_2 \gamma_{2,4} \Delta x_2 + \alpha_1 \gamma_{1,4} \Delta x_1 + \epsilon \quad (3)$$

where, for convenience, we have dropped all i and j subscripts.

If we have data of the students' teachers' knowledge over their full schooling history, and assuming $\text{cov}(\Delta x_t, \epsilon) = 0$, we could use OLS to estimate consistent (reduced form) estimates of the structural parameters in equation (3). Further, we could then also estimate the cumulative effect of teacher knowledge on student learning after four years as $\alpha_4 + \sum_{t=1}^{3} \alpha_t \gamma_{t,4}$. In the data, however, we do not observe Δx_2 and Δx_1. Thus, instead of equation (2), what we can estimate is

$$\Delta y = \beta_0 + \beta_4 \Delta x_4 + \beta_3 \Delta x_3 + \mu \, . \quad (4)$$

Clearly, even assuming $\text{cov}(\Delta x_t, \epsilon_{ij}) = 0$, we cannot recover the structural coefficients of interest from (4) as long as there is correlation between subject differences in teacher test scores across grades and their effect on current student scores does not decay completely. Nevertheless, under mild conditions, knowledge of current and previous test scores in grade 3 and 4 is sufficient to estimate a lower bound on the cumulative effect of teacher knowledge on student learning.

To see this, we start by assuming stationarity of the test score distribution over time: $\text{var}(\Delta x_t) = \text{var}(\Delta x_{t'}) = \text{var}(\Delta x)$.[11] The OLS estimator of β_4 and β_3 in (4) can then be written as (see derivation in the appendix):

$$\text{plim}\, \hat{\beta}_4 = \alpha_4 + \alpha_2 \gamma_{2,4} \left(\frac{\rho_{2,4} - \rho_{2,3}\rho}{1-\rho^2} \right) + \alpha_1 \gamma_{1,4} \left(\frac{\rho_{1,4} - \rho_{1,3}\rho}{1-\rho^2} \right), \quad (5)$$

and

$$\text{plim}\, \hat{\beta}_3 = \alpha_3 \gamma_{3,4} + \alpha_2 \gamma_{2,4} \left(\frac{\rho_{2,3} - \rho_{2,4}\rho}{1-\rho^2} \right) + \alpha_1 \gamma_{1,4} \left(\frac{\rho_{1,3} - \rho_{1,4}\rho}{1-\rho^2} \right), \quad (6)$$

[11] As reported in appendix, we cannot reject the hypothesis that the variances are the same for teachers' knowledge in grade 4 and grade 3.

where $\rho_{t,t'} = \frac{\text{cov}(\Delta x_t, \Delta x_{t'})}{\text{var}(\Delta x)}$ and where ρ is estimated from the linear projection of current test scores on previous test scores (differenced across subjects); i.e., from equation (7),

$$\Delta x_4 = \rho_0 + \rho_{43}\Delta x_3 + v_{4,3}\ , \tag{7}$$

where $\rho_{43} = \rho_{34} = \rho$ and that the regression coefficient equals the correlation coefficient follows from the stationarity assumption.

Second, we make some mild, and we would argue empirically valid, restrictions on the parameter space for the $\rho_{t,t'}$. Specifically, we assume that all $\rho_{t,t'} \geq 0$. Second, we assume that $\rho_{t,t'}$ is decreasing in $|t - t'|$. In other words, the further apart are any two sets of grades, the lower the correlation between teacher knowledge. Finally, we assume that $\rho_{t,t-1}$ is decreasing in t. Note that this implies that no $\rho_{t,t-1}$ is smaller than the estimated correlation between test scores (differenced across subjects) in grade 3 and 4; i.e., ρ.

These restrictions are motivated by the typical pattern of transitions of teachers and their grades through primary school coupled with the assumption that these transition patterns are the main (though not necessarily exclusive) drivers of correlations across grades. In particular, there is a substantial share of teachers who transition with their grade each year (restriction (i)), any grade t is more likely to be taught by their teacher in grade $t-1$ than teachers in earlier grades (restriction (ii)), and teachers in earlier grades are more likely to transition together with their grade than teachers in higher grades (restriction (iii)).[12]

We can then summarize the main implications one can draw from estimating the reduced form equation (4).[13] First, even with the restrictions of the parameter space for the $\rho_{t,t'}$, it is clear that neither $\hat{\beta}_4$ nor $\hat{\beta}_3$ bound the structural parameters. More specifically, $\hat{\beta}_4$ provides a biased estimate of α_4 with the bias depending on the size and sign of $\text{cov}(\Delta x_4, v_{t,3})$, where $v_{t,3}$ is the residual in a regression of Δx_t on Δx_3, with $t = 1,2$. If these covariances are zero, the coefficient of current teacher knowledge captures the contemporaneous effect (in grade 4), once controlling for prior teacher knowledge. As a reference, assuming $\text{cov}(\Delta x_4, v_{t,3}) = 0$ for $t = 1,2$ is not sufficient to estimate α_4 in a contemporaneous specification, i.e., in a specification in which student achievement is regressed on current teacher knowledge only. Specifically, for the contemporaneous

[12] Note that implicit in this argument is the assumption that teacher knowledge is essentially fixed or changes only very little year by year. In other words, in a context in which teacher in-service training is rare, measuring the knowledge of a teacher who taught in year $t-1$ in year t, provides a good measure of the teacher knowledge input in year $t-1$.

[13] See appendix for details.

specification to identify the contemporaneous effect (α_4 in this case), one needs to assume no correlation between subject differences in teacher test scores across all grades; i.e., $\text{cov}(\Delta x_4, \Delta x_t) = 0$ for $t < 4$, or that the contemporaneous effect decays completely; i.e. $\gamma = 0$.

Second, the sum of $\hat{\beta}_3$ and $\hat{\beta}_4$ provides a lower bound on the cumulative effect ($\alpha_4 + \sum_{t=1}^{3} \alpha_t \gamma_{t,4}$). Importantly, this holds although we have not placed any restrictions on the contemporaneous effects (they could thus be age or grade dependent) or the degree of persistence.

Finally, access to additional data on teachers' knowledge in earlier grades would primarily influence the precision of the estimates. In fact, even with data of teacher knowledge over the students' full schooling history, we would need to add additional structure to identify the structural parameters α_t and $\gamma_{t,t'}$.

4.3. Structurally estimating α and γ

We have shown that data on test scores of the current and the previous year's teachers allows the estimation of a lower bound on the cumulative effect of teacher knowledge on student achievement. We now discuss the additional structure on the cumulative student achievement function required to estimate (ranges for) the contemporaneous effect of teacher content knowledge on student achievement (α), as well as the extent of fade out of the teachers' impact in earlier grades (γ).

We make two assumptions.[14] First, we assume that the contemporaneous effect is independent of the age at which it is applied, implying that $\alpha_t = \alpha_j = \alpha$. Second, we assume that the effect of teacher content knowledge declines (geometrically) with distance, implying that if α is the marginal effect of teacher content knowledge in grade t on student achivement at the end of the same grade, then $\gamma\alpha$ is the marginal effect of teacher content knowledge in grade $t-1$ on student achivement at the end of grade t, where γ denotes the degree of persistence.

With these assumptions, we can rewrite (3) as

(3') $$\Delta y = \alpha_0 + \alpha\Delta x_4 + \alpha\gamma\Delta x_3 + \alpha\gamma^2\Delta x_2 + \alpha\gamma^3\Delta x_1 + \epsilon$$

and the OLS estimator β_4 and β_3 in (4) as

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

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

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

We are now left with three equations (5'), (6'), and (7) and seven unknowns (α, γ, ρ, $\rho_{2,4}$, $\rho_{2,3}$, $\rho_{1,4}$, $\rho_{1,3}$). Thus, to recover the structural parameters α and γ from equations (5'), (6'), and (7), we need to reduce the number of free parameters; i.e., we need to know how the observed (Δx_4, Δx_3) and the omitted (Δx_2, Δx_1) teacher scores are correlated across grades. We take an agnostic approach. Thus, we continue to assume that the correlation coefficients, $\rho_{t,t'}$, vary freely in the restricted empirical space discussed in section 4.2., and estimate the structural parameters for each point in this space (after discretizing). We then report the mean and the full distribution of possible cumulative effects of teacher knowledge on student learning, $\alpha \sum_{t=0}^{3} \gamma^t$ in this space, which is equivalent to a Bayesian approach with a uniform prior over each admissible combination of correlation coefficients, $\rho_{t,t'}$.[15]

Beyond the values for $\rho_{t,t'}$, there are three unknowns in equations (5'), (6'), and (7), namely α, γ, ρ, which can be uncovered from the estimated coefficients $\hat{\beta}_3$, $\hat{\beta}_4$ and $\hat{\rho}$.

4.4. Identifying the parameters using within-student within-teacher variation

In deriving the OLS estimator β_4 and β_3 in (4), we have (implicitly) assumed that the $\text{cov}(\Delta x_t, \epsilon) = 0$. As in any selection-on-observables model, the extent to which this assumption is empirically valid crucially hinges on the assumption that student assignments to teachers can be controlled for.

Our core identifying assumption is that sorting of students and teachers are subject-invariant; i.e., that sorting can be controlled for by exploiting variation within students and teachers across subjects. To better understand what this assumption implies, consider again the cumulative student achievement function (1). Linearizing the production function and taking the first-difference (across subjects) yields equation (2). In our sample, about half of the students are taught by a class teacher; i.e., a teacher teaching both subjects in a given grade in the current and previous year (although not necessarily the same teacher in both years). If we restrict attention to students who were taught by such class teachers, we can further remove any teacher-specific, subject-invariant heterogeneity in each of these years, implying that the error term ϵ_{ij} in equation (2) can be written as

(8) $$\epsilon_{ij} = \Delta\omega_{ij} + \sum_{t=1}^{4}\left[\theta_t^S \Delta s_{ijt} + \theta_t^P \Delta p_{ijt} + \theta_t^T \Delta c_{ijt|f_k=f_{k'}}\right] + \sum_{t=1}^{2} \theta_t^{T\prime} \Delta c_{ijt|f_k\neq f_{k'}} + \Delta\varepsilon_{ijt}$$

[15] In the estimation, we also adjust $\hat{\beta}_4$ 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.

where we distinguish between unobserved teacher skills and characteristics for students who were taught by the same teacher f in both subjects in year 1 and 2, $\theta_t^T \Delta c_{ijt|f_k=f_{k'}}$, and for those who were taught by different teachers in language and mathematics, $\sum_{t=1}^{2} \theta_t^{T\prime} \Delta C_{ijt|f_k \neq f_{k'}}$. In the former case, only subject-specific variation enters the error term, while in the latter both subject-specific and subject-invariant variation may matter; i.e., $C_{ijt} = [c_{ijt,k}, \bar{c}_{ijt}]$.

Consider next the terms in (8), starting with the differenced ability/motivation term $\Delta\omega_{ij}$. Note that $\Delta\omega_{ij} \neq 0$ only if students have subject specific abilities/motivations. If that is the case, our identifying assumption rules out that students systematically sort, based on these subject-specific abilities, into schools with subject-specific teacher knowledge. For example, our assumption would be invalid if students in lower primary with relatively higher motivation for mathematics sort into schools (or classrooms) with relatively more knowledgeable mathematics teachers. It also places some restrictions on what parents and schools do. For example, while our identifying assumption does allow for parents (or the school) to respond to their children's low mathematics aptitude by providing additional teaching (or hire a private tutor), they cannot do this to compensate for insufficient teacher mathematics knowledge.[16] More generally, while differential (across subjects) supply of school (Δs_{ijt}) and parental (Δp_{ijt}) inputs may occur across schools and students, 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. In the context of lower primary schooling in Africa, these assumptions appear reasonable and we provide additional evidence in support of them in section 5. However, the assumptions remain fundamentally untestable without complete data on inputs histories.

Turning to the unobserved variation in teacher skills and characteristics across subjects that enters the error term for students who were not taught by a class teacher in year 1 and 2, $\sum_{t=1}^{2} \theta_t^{T\prime} \Delta c_{ijt|f_k \neq f_{k'}}$, there is reason to believe that the potential bias, if any, will be small. Specifically, the structure of primary school in the countries we study is such that students tend to have one 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 students move to upper primary and

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

secondary. In other words, if a student is taught by a class teacher in both mathematics and language in grade t, it makes it likely, we would argue, that the student was taught by a class teacher who teaches both subjects also in $t-1$, while the opposite transition may not hold. This pattern is also clear from the data. Specifically, while 60% of the students taught by a teacher teaching both subjects in grade 3 are also taught by subject-specific teachers in grade 4, more than 90% of the students that are taught by a class teacher who teaches both subjects in grade 4 also had a class teacher in grade 3. Although we do not know who taught the students in grade 1 and 2, we argue that these data patterns suggest that our core sample (which is restricted to students who had class teachers in grade 3 and 4) is unlikely to contain many students who were taught by teachers specializing in their subject in grade 1 and 2. Indeed, if all students that have class teachers in grade 3 also have class teachers in grade 1 and 2 (though again not necessarily the same teacher as in grade 3), then the within student transformation of the data and the restriction of the sample to students who have only one teacher in both subjects in grade 3 (and in grade 4) is sufficient to remove the nuisance term $\sum_{t=1}^{2}\theta_t^{T\prime}\Delta c_{ijt|f_k\neq f_{k'}}$.

Finally, consider the variation in $\Delta c_{ijt|f_k=f_{k'}}$; i.e., in other (unobservable) characteristics or skills that vary by subject for the same teacher. For example, a teacher, teaching both subjects may be more motivated to teach a subject she masters relatively well, or possibly put more effort into teaching if she is less knowledgeable of the subject. To the extent these additional subject-specific traits are systematically correlated with teacher subject-specific content knowledge, α_t needs to be reinterpreted slightly more broadly; i.e. as the impact of teacher content knowledge and other unmeasured teacher subject-specific teaching traits correlated with it.

4.5. Inference

To make inference about the structural parameters of interest in equation (3'), $\widehat{\boldsymbol{\theta}} = \left\{\hat{\alpha}, \widehat{\alpha\gamma}, \widehat{\alpha\gamma^2}, \widehat{\alpha\gamma^3}\right\}$, we need to estimate their standard errors. The asymptotic variance-covariance matrix of $V = \sigma_\epsilon^2(E(\Delta\tilde{x}'\Delta\tilde{x}))^{-1}/N$, 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 across four years. Note that each of the two terms in the variance-covariance matrix depends on population moments for which we have no sample analogue. That is, we cannot replace, for example $E(\Delta\tilde{x}'\Delta\tilde{x})_{3,2} = \text{cov}(\Delta x_3, \Delta x_2)$, by its sample analogue simply because we do not observe Δx_2.

To circumvent this, we proceed in two steps. First, we show in Appendix C that under the assumptions above, each of the terms can be rewritten as a function of population moments

whose sample analogue we do observe. Second, we show, again in Appendix C, how to estimate each of these functions and therefore the asymptotic variance-covariance matrix by substituting the appropriate sample analogue for the population moment.[17]

5. Results

5.1. Relationship between student and teacher content knowledge

We start by providing empirical support for two of the assumptions we rely on to derive our core empirical specification (4), namely the assumption that the coefficients on all (teacher knowledge) inputs are subject-invariant and the assumption of stationarity of the test score distribution over time. 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 (4) as a correlated random effects model. As discussed in in appendix D, and reported in Table A1, 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 A2 in the appendix reports summary statistics on the content knowledge of the current and prior teachers and performs tests on the equality of standard deviations (variances) of the two distributions. As evident, we cannot reject the assumption of stationarity of the test score distribution over time, at least for the two periods (current and prior teacher knowledge) we have data for.

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) on the basis of subject invariant characteristics, as well as other unobserved student and teacher subject invariant, characteristics. In these specification, the effect of teacher knowledge on student test scores can only be driven by differences

[17] Note that applying the delta method to the non-linear functions of $\hat{\beta}_4$ and $\hat{\beta}_3$ is not possible here. First, with the exception of the limiting cases, 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.

between the two subjects. The results suggests that part of the association in column (1) is driven by better students sorting into better schools.[18,19]

In column (4) we include past teacher knowledge, resulting in a fall in the estimated coefficient on current teacher knowledge by about 30 percent.[20] 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 increases student learning by at least 0.10 standard deviations.

Finally, in column (5), our preferred specification, we also introduce teacher fixed effects now 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 more marginal (the lower bound on the cumulative effect falls by approximately 16 percent).

5.2. Robustness checks of the reduced form estimates

We argue that, under mild conditions, knowledge of current and previous test scores in grade 3 and 4 is sufficient to estimate a lower bound for the cumulative effect, and with additional restrictions on the parameter space, also ranges 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 effects to be interpreted as causal, the estimated reduced form coefficients of current and previous teachers content knowledge in Table 4, column (5), cannot be biased by omitted variables (of other variables than grade 1 and grade 2 teacher knowledge) or sorting.

To recap the identification assumptions stated in Section 4, two things have to be true for this to be the case: (i) there must not be other factors (at teacher level or otherwise) that

[18] 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.

[19] The specifications in columns (2) and (3), where contemporaneous student achievement is regressed on a contemporaneous teacher score measure, are identical to the core specifications in Metzler and Woessmann (2012) and Bietenbeck et al. (2017), both using data for grade 6 students and their teachers. The point estimates on current teacher knowledge in columns (2) and (3) are similar, but somewhat larger, than those reported in these earlier studies (standardized effects of 0.07-0.08 in columns (2) and (3), as compared to 0.026 in Bietenbeck et al. (2017) and 0.06 in mathematics but a small and insignificant effect for language in Metzler and Woessmann (2012)).

[20] 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.

drive both student and teacher subject differences in knowledge; (ii) there is no sorting by students and teachers on the basis of subject differences. In other words, students that are better in language than in mathematics are not systematically more likely to select into schools with teachers that 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 suggesting that neither of these assumptions is likely to be violated. 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). As can be seen, compared to the main specification reported in column (1), the estimates, including the lower bound for the cumulative effect, change only marginally.[21]

Similarly, other teacher behavior and skills that vary by subject might be correlated with teacher knowledge and affect learning. While we do not have any measure of teacher behavior that varies across subjects for a given teacher, if we also include students taught by different teachers in language and mathematics in the sample, we can test directly how teacher subject knowledge correlates with other teacher skills. Specifically, 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), Table 4.[22] 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 (cf. Table 4, column (4)). The lower bound on the cumulative effect is only marginally affected by the inclusion of pedagogy knowledge. Hence, we would argue that unmeasured differences in teacher skills—at least pedagogical skills—across subjects are unlikely to confound the coefficient of interest.

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

[22] Pedagogical knowledge is measured as the score on a lesson preparation exercise that was administered to all teachers. The assessment pedagogy knowledge and skills is described in Bold et. al. (2017).

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

To further bolster the causal interpretation, 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 purely a causal relationship between student achievement and teacher knowledge, then including the test scores of teachers in the same school who have not taught the student should not change the coefficients on current and previous teacher knowledge and should in itself have no significant impact on student test scores. The results in column (1) and (2) of Table 5 broadly conform to this pattern. 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).[23] Hence, the estimated effects are clearly 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 the table, 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 indeed matters.

5.3. Estimates of the cumulative effect of teacher content knowledge on student learning

We now turn to presenting 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, $\alpha \sum_{t=0}^{3} \gamma^t$. Figure 5 depicts the full range of these structural parameters, implied by the reduced form coefficients in Table 4, column (5), a $\hat{\rho} = 0.58$, and the restrictions imposed on the correlations between observed and unobserved subject differences in test scores as

[23] 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.

described in section 4.2. Table 7, column (1), summarizes the findings by reporting the estimates for α, γ and $\alpha \sum_{t=1}^{4} \gamma^{4-t}$ at the median cumulative effect across all the admissible correlation structures.[24]

The contemporaneous effect is estimated to lie between 0.056−0.080 (5th−95th percentile) with a mean of 0.069 standard deviations (Figure 5, Panel A). As a comparison, effect sizes in the value added literature, which focuses on estimating the impact of a broader measure of teacher quality, range from 0.1-0.2SD (Rockoff, 2004; Rivkin et al., 2005; Aaronson et al., 2007; Chetty et al., 2014b; Araujo et al., 2016; and Bau and Das, 2017).[25]

Equally, the range for the degree of persistence, 0.39−0.59 (5th–95th percentile) with a mean of 0.47 (Figure 5, Panel B), 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). This implies that approximately 50 percent of the short-run effect persists between grades, again using the mean of the estimated persistence parameters as a reference point.

Figure 5, Panel C plots the cumulative effect, which is estimated to lie between 0.112 and 0.136, with the average total effect across all parameterizations estimated to be 0.123. The minimum for the cumulative effect coincides with the scenario in which students have the same teacher throughout lower primary (grades 1-3) and then some students change teacher only once they reach grade 4, implying that $\rho_{t,t'} = 1 \; \forall t, t' < 4$, and $\rho_{4,t'} = \rho$.[26] On the other hand, 97% of the values lie below the cumulative effect estimated in the case where subject differences in test scores follow an AR(1) process (i.e. $\text{cov}\left(\Delta x_t, v_{t',t'-1}\right) = 0 \; \forall t, t' < t$). Thus, assuming an AR(1) process effectively provides an upper bound on the cumulative effect.

In sum, we have shown that knowledge of current and previous test scores allows us, under mild restrictions on the correlation patterns of omitted and included teacher knowledge, to estimate fairly tight bounds on the total cumulative effect of teacher knowledge on student

[24] Note that α and γ that solve (5'), (6'), and (7) are not independent, and hence $E(\alpha \sum_{t=1}^{4} \gamma^{4-t}) \neq E(\alpha) \sum_{t=1}^{4} \text{E}(\gamma)^{4-t})$. In Table 7, we report the parameter estimates for α and γ at the median cumulative effect.

[25] 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.

[26] In this case, effectively, there is no omitted variable problem. The coefficients on the teacher knowledge variables suffer only from collinearity in the regressors. The sum of $\hat{\beta}_3$ and $\hat{\beta}_4$ is therefore an unbiased estimate of the cumulative effect.

learning (the 10th to the 90th percentile range of the estimate is 0.118−0.129) and that the range of values is bounded by two intuitively appealing processes.

5.4. Structural simulations

We next use the structural parameters to answer three questions. First, following on from the development accounting approach (see Caselli (2005) for a detailed review), we exploit the structural estimates of the returns to teacher content knowledge in the aggregate production of education to calculate the extent to which shortfalls in this input account for the low levels of student achievement.

Second, and inspired by the misallocation literature (see Banerjee and Duflo, 2005; Hsieh and Klenow, 2009), we consider heterogeneity in the distribution of teacher knowledge and class sizes across schools. Specifically, we ask how much learning is lost because of misallocation, more precisely, that students are not allocated to the best teachers up to the point where the marginal gain from increased learning equals the marginal cost from increased class sizes.

Finally, we consider the impact of two complementary policy reforms—ensuring that teachers are properly trained and increasing teacher effort and reducing absenteeism—and especially the longer run impact 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 lower secondary education knowledge.

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 alternative measure of teacher knowledge, which transforms student test scores into "effective years of learning" and teacher test scores into "effective years of education" (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. [27]

Based on our estimates teachers' have an education gap of 6.1 years relative to the official requirement.[28] Consequently, being taught by teachers with minimum knowledge,

[27] 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.3-A.5 (reduced form estimates) and Table 7, column (2) (structural estimates).

[28] 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.

over four years, would increase effective years of schooling by 0.64−0.80 years, with the median of the full range of parameterizations just below three quarters of a year (Figure 6, panel A). This in turn implies that insufficient teacher content knowledge accounts for 28−35 percent of the observed schooling gap after four years, with the median equal to 30 percent.[29] Alternatively, comparing the gap between the lowest performing countries in the sample in terms of student knowledge, Mozambique and Togo, to the highest performing, Kenya, 21 percent of the gap is explained by differences in teachers' effective years of education.

Together, these results suggest that variation in teacher knowledge accounts for a sizeable chunk of the learning gap (either relative to the curriculum or across countries). The results, however, also make clear that the lion's share of the gap in learning outcomes (again either relative to the curriculum or across countries) is explained by other shortfalls, 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. In fact, reforms focusing purely on increases in teacher knowledge to close the learning gap would require teachers in Sub-Saharan Africa to complete effective education exceeding university level.[30,31]

Reducing misallocation

In the second experiment, 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 implies that teachers who know more about their subject should teach more students.[32] In the data, however, we find, a negative correlation between teacher knowledge and class size.[33] We now ask how much

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.

[29] 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).

[30] This result is arrived at by dividing students' shortfall in human capital after four years, 2.5, by the amount of learning acquired after four years if teachers increased their human capital by one year.

[31] 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.

[32] 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.

[33] 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.

student learning is lost because of this misallocation, and thus consider the effects of moving students from the worst performing teachers to those with relatively better content knowledge, for different assumptions about the class size effects and for different assumptions of the feasibility of reallocating students across space.

In appendix E, we present the full constrained maximization problem and solve for the optimal reallocation. We consider several effects that partly offset each other: (i) a positive effect on learning for those students who were transferred from a teacher with low subject content knowledge to a teacher with higher subject content knowledge; (ii) a 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 positive effect on learning for the students left behind with the low knowledge teacher arising from decreased class sizes.

To calibrate the learning gain (i), we use the median across the estimated structural parameters of the contemporaneous effect of teacher knowledge on student learning, $\alpha = 0.07$. To calibrate the class size effects (ii)-(iv), we draw on the existing literature. In particular, we assume that class size effects, denoted σ, are linear in log-changes and consider effect sizes ranging from zero, as estimated experimentally by Duflo, Dupas and Kremer (2015) in Kenya, to an upper bound of a 0.15 standard deviation in learning for a doubling of class size, somewhat lower than the estimate by Muralidharan and Sundararaman for India (2013).[34] We consider both the possibility of reallocating students taught by the worst performing teachers to better performing teachers within a given school/grade and the possibility to reallocate students across schools within a district.

Figure 7 illustrates the costs of misallocation in terms of foregone student learning. If there are no costs from larger class sizes and all students from below average teachers are reallocated, then the human capital lost is equivalent to increasing teacher knowledge by half a standard deviation (when considering reallocation at the school level) and three quarters of a standard deviation when considering reallocation at the district level (or 0.05 standard deviation of student learning). 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 0.016 of a standard deviations.

Although the learning gains from efficient allocation decrease as the cost of learning in larger classes increases, there are sizeable inefficiencies in terms of spending even when σ

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

is at its upper bound. To calculate these inefficiencies, we conduct a second and related simulation. In particular, we ask whether teachers with less subject knowledge than the median teacher could be dismissed (and students optimally reallocated at either the school or district level) without affecting student learning.[35]

Our estimates suggest that learning outcomes could be achieved at roughly half the cost if a doubling of class sizes reduces learning by 5% of a standard deviation and that budgets could be reduced by 15% even for the highest class size effect considered. Alternatively, these savings could be used to incentivize the remaining teachers: If one doubled salaries conditional on being present and teaching, then, given class absence rates of 50% across the continent, one could de facto reallocate students holding class size and the effective time of instruction constant, while at the same time increasing learning by 0.05 of a standard deviation.

Increasing teacher knowledge and time spent teaching: Projected long run effects

Finally, we consider the cumulative effect of increasing both teacher knowledge and effort. The provision of primary education has expanded greatly in the developing world in the last two decades, including in Sub Saharan Africa. In the survey data, we find that twice as many teachers have entered the profession in the last ten years than in the decade before. As the large majority of teachers are employed on permanent and pensionable civil service contracts, and as upgrading current teachers' knowledge probably is significantly more costly than raising the quality of new cohorts of teachers, this huge expansion of the teaching body with low knowledge represents a lost potential.[36] On the other hand, the expansion will likely continue. According to recent population projections, close to half the world population of children will live in Africa by the end of the 21st century (UNICEF, 2014). Specifically, the number of primary school age children in Sub-Saharan Africa is set to rise from about 170 million to 209 million in the next 10 years, reaching 280 million by the mid-century.[37] As reported in Figure 7, 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

[35] 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 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

[36] 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, though partly reflects an age and a cohort effect, as many contract teachers graduate to civil service status over time (Bold et al., 2017).

[37] See appendix for details for these projections and the policy simulation.

teacher, would require the hiring of 1.7 million new teachers by 2030 and close to five million by 2050.

We use the estimate of the cumulative effect to project the impact of a reform that would raise teachers' knowledge over time to the minimum entry requirement for teacher training; i.e., lower secondary education. We treat the knowledge of current teachers as fixed; i.e., average teacher content knowledge can only be raised by hiring new qualified teachers, and assume teachers work for 40 years. Further, we assume that new teachers are hired to start teaching first grade and then move one grade each year up to grade six when they again start teaching first grade students. We also project the effect 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 $(270/166)\alpha$. Finally we consider the effect of combining the two reforms. The results of the policy simulations are summarized in Figure 8.

Four findings stand out. First, even if all newly hired teachers have the required minimum knowledge—a stark difference from the current system—the reform still has relatively small effects even after 10 years: effective years of student learning are projected to increase from the current average of 1.78 to 2.06 in the first decade; i.e. by approximately 16 percent. This relatively small effect, in turn, is caused by the fact that it takes time to improve grade four outcomes, given our assumption of how the new teachers are allocated in schools and given that in the shorter run most students will still be taught teachers already hired before the reform. For example, in the first three years, there will be no change in average student achievement for fourth graders, simply because students that have benefitted from being taught by qualified teachers have yet to reach grade four, and of the grade four students in the first four years, only about 4 percent will have been taught by qualified teachers.

Second, the reform is projected to increase effective years of schooling by approximately 40 percent by mid-century. This is likely a lower bound, however, since higher subject content knowledge also raises aspects of teachers' pedagogical skills, for instance their ability to interpret data so as to monitor their students' progress. Such indirect effects are not accounted for here.

Third, and by assumption, increasing the time teachers teach raises student achievement directly, since the reform is assumed to influences all teachers.[38] However, even if this reform results in a large increase in the amount of teaching, the results reported in Figure 8 illustrates its limitation in an environment where teachers' content knowledge is low. That is, having teachers that know little teach more only marginally raises student achievement.

Fourth, the combined reforms lead to both large short and longer run impacts. After five years, for example, average student achievement increases by 25 percent and at the end of the sample period, effective years of schooling is projected to be close to three years. Figure 8 also illustrates the complementarity between reforms to increase teacher quality; i.e., reducing teacher absenteeism and thus increasing the instructional time of students has a larger marginal effect if teachers master the subject they teach.

6. Discussion

Recent estimates suggest that differences in (the quality of) human capital can explain a dominant share of world income differences (Jones, 2014; Malmberg, 2017). 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 paper, 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 (see Glewwe and Muralidharan, 2015). Here we show that teachers' content knowledge, or lack thereof, is an important explanation for why students in primary schools in Africa already after a few years of schooling are far behind their counterparts in most developed countries. Potential human capital for cohorts of students is consequently lost.

Our results have implications for both policy and research. Regarding the latter, there are few, if any, well-identified studies on how to effectively improve teacher knowledge and the impact thereof (Glewwe and Muralidharan, 2015). Our results strongly suggest that this evidence gap is important to address.

We also provide evidence that a reform ensuring that the next cohorts of teachers is better prepared to teach well, and have incentives to do so when deployed, can potentially go a long way to improve outcomes. The (aggregate) impact of such a reform, however, takes

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

time to materialize. For that reason it is also important to experiment and roll-out shorter term solutions. A number of such interventions have been shown to produce promising results, including programs to supplement current teachers with additional instructors, or automate certain aspects of teaching using computer-aided learning programs or scripted lesson plans, including when taken to scale (Banerjee, et al., 2017).[39]

Our nationally representative data reveals large variation both within and across countries, a fact that also has implications for policy. For example, in countries like Kenya where teacher effort (in terms of actual presence) is low but effective knowledge is relatively high, putting focus on increasing teacher effort is sensible. In countries like Togo and Mozambique, however, given the quality of teaching, such a reform is less likely to yield the same results. Overall, our main message here and in previous work (Bold et al., 2017) is the importance of complementarities between different dimensions of teacher quality: Increasing the effort of a teacher who is well-trained and educated will likely go a long way toward increasing human capital accumulation.

[39] See, for example, the reviews in Murnane and Ganimian (2014), Glewwe and Muralidharan (2015), and Evans and Popova (2016).

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. 2016. “Experimental Evidence on Scaling Up Education Reforms in Kenya.” Working Paper, IIES.

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.

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.

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.

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

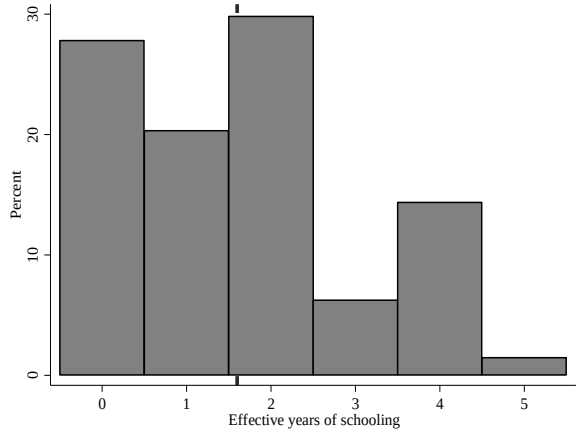
UIS. 2011. Financing Education in Sub-Saharan Africa: Meeting the Challenges of Expansion, Equity and Quality. UNESCO/UIS. Montreal.

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

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

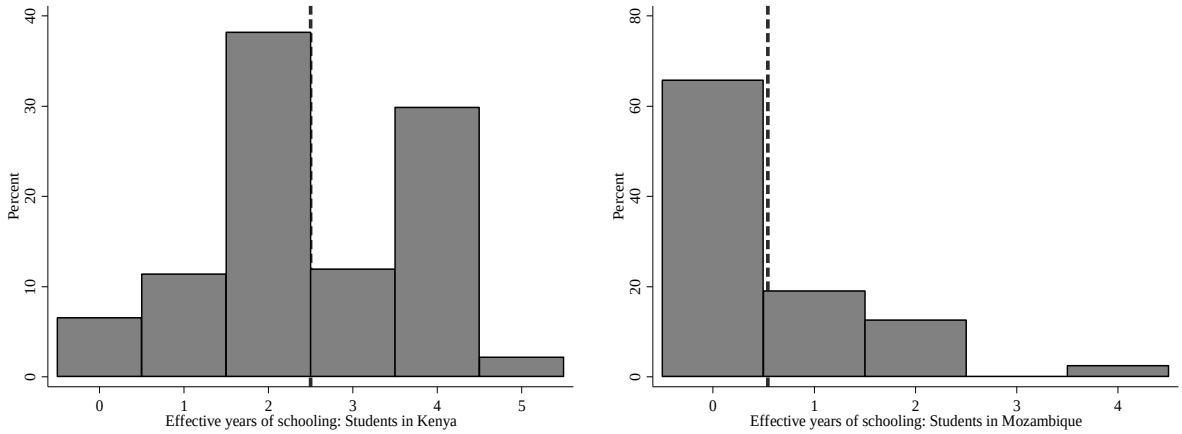
World Bank. 2017. *World Development Indicators*. Washington DC.

Figure 1: Effective years of schooling 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 after four years of primary education: Kenya and Mozambique

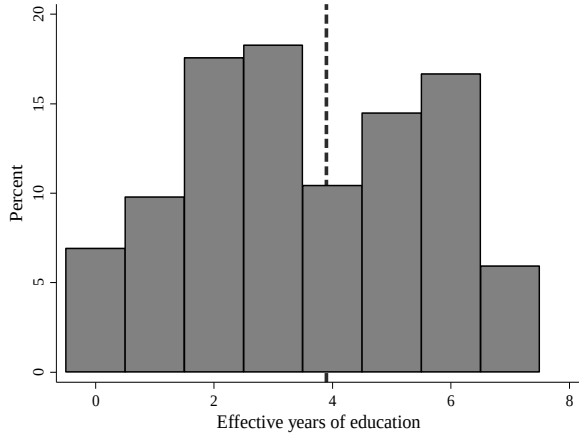


Panel A: Kenya

Panel B: Mozambique

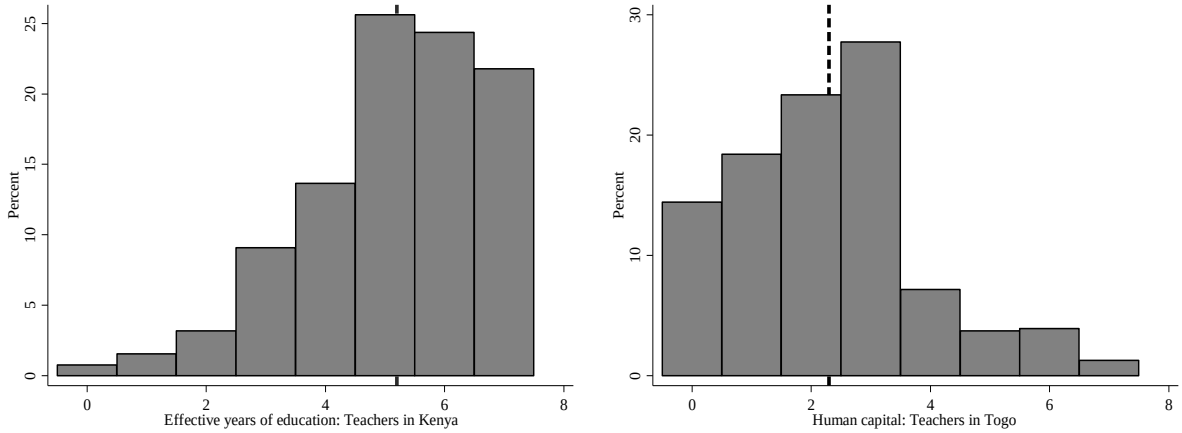
Note: Distribution of effective years of schooling for students in Kenya (Panel A) and Mozambique (Panel B). Pooled data across subjects for each country. Dashed vertical lines depict means.

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.

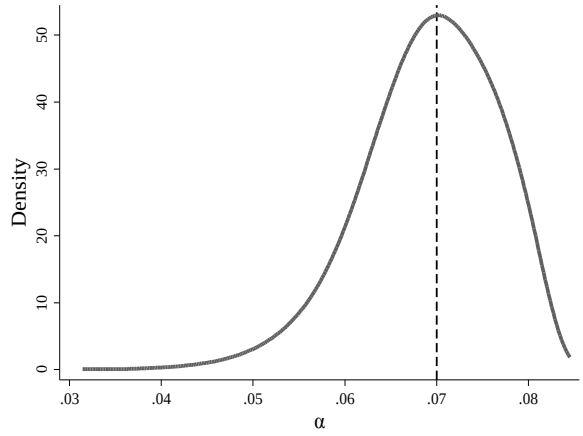
Figure 4: Effective years of education for primary school teachers: Kenya and Togo



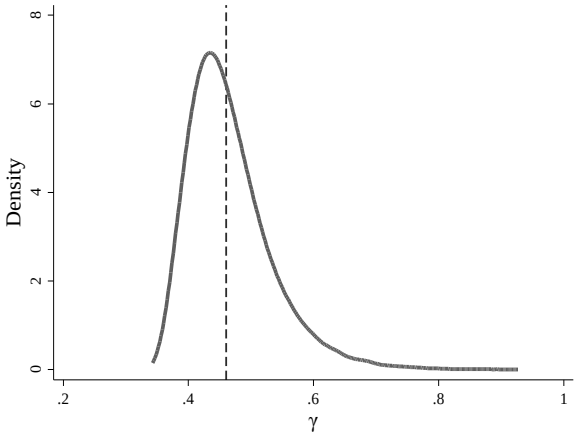
Panel A: Kenya Panel B: Togo

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.

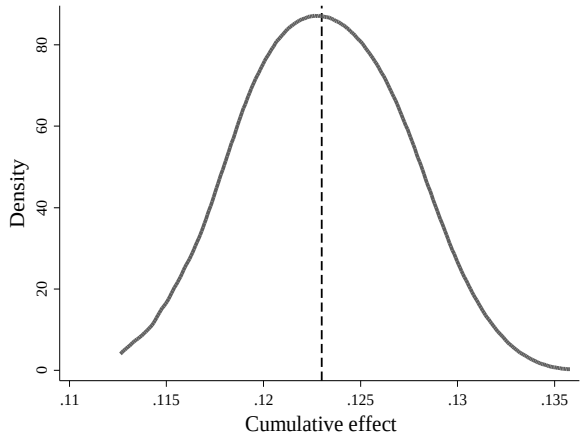
Figure 5: Probability density functions of the estimated α, γ, and the cumulative effects



Panel A: Contemporaneous effect



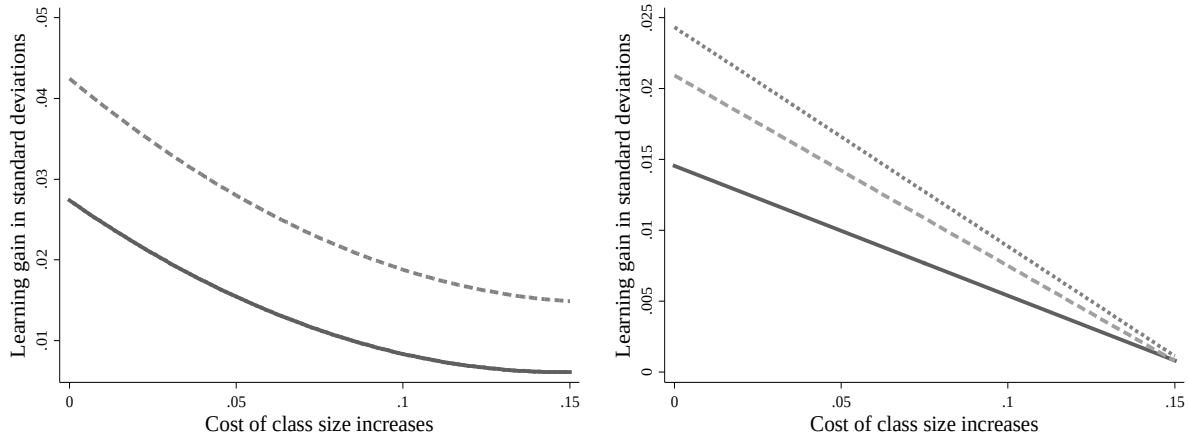
Panel B: Persistence



Panel C: Cumulative effect

Note: The figures show the distributions of the estimated contemporaneous effect of teacher knowledge on student learning, its persistence and the total cumulative effect after four years. Dashed vertical lines depict median values.

Figure 6: 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 25^{th} (50^{th}) 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 7: Teacher projections, 2016-2050

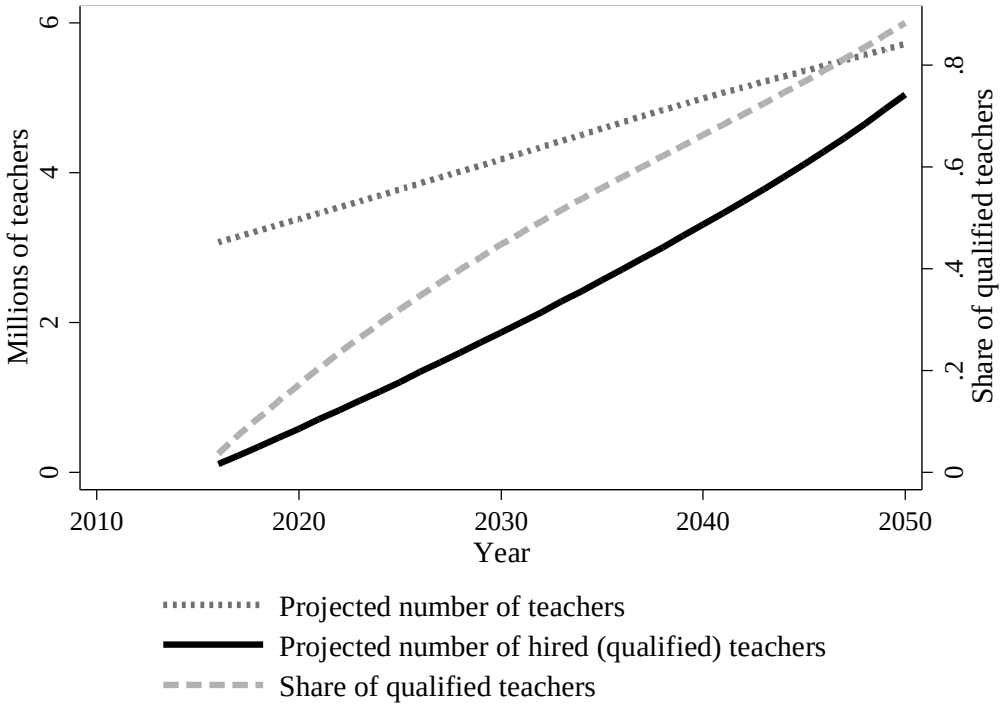


Figure 8: Projected impact of three policy reforms, 2016-2050

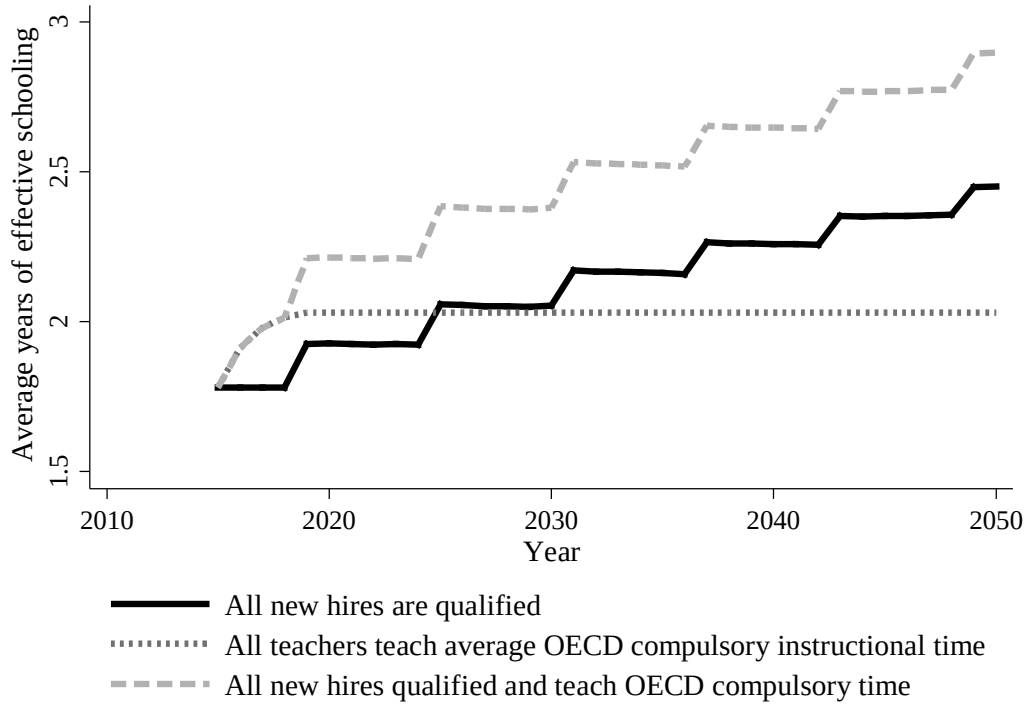


Table 1: Summary statistics

	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%	36%
1	-0.56	-0.06	15%	21%
2	0.15	0.49	45%	18%
3	0.98	0.92	5%	6%
4	1.48	0.94	8%	17%
5	-	1.57	n/a	3%
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.

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	-1.49	-2.08	5%	7%
1	-0.95	-0.93	4%	12%
2	-0.44	-0.41	12%	13%
3	-0.04	-0.008	19%	13%
4	0.36	0.53	19%	6%
5	0.91	0.84	33%	2%
6	1.28	0.66	2%	32%
7	1.65	1.07	6%	15%
N	5511	5772	5511	5772

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.

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

Dep. Variable	(1)	(2)	(3)	(4)	(5)
	Student test scores				
Content knowledge of current teacher	0.034	0.029	0.027	0.055**	0.029
	(.028)	(.023)	(.027)	(.022)	(.029)
Content knowledge of prior teacher	0.049*	0.056**	0.053**	0.038*	0.057***
	(.027)	(.023)	(.026)	(.021)	(.028)
Teacher pedagogy score				0.242**	
				(.123)	
Constant	-0.017	-0.017	-0.018	0.015	0.018
	(.021)	(.018)	(.026)	(.018)	(.022)
Lower bound (total effect)	0.083***	0.085***	0.080***	0.094***	0.086***
	(.024)	(.023)	(.024)	(.021)	(.024)
Specification/Sample	Main	District FE	Urban FE	Student FE	One gr. 4 class
Observations	4,393	4,393	4,393	6,382	3,831
Number of schools	605	605	605	951	524

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)
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)	0.067*	0.060**
	(0.0406)	(0.037)
Persistence (γ)	0.487	0.507
	(0.846)	(0.601)
Total effect after 4 years	0.123***	0.114***
	[0.001]	[0.000]

Note: Median estimates for α, γ and the cumulative effect across all correlation structures. Panel A reports estimates using IRT scores and Panel B uses effective years of education of teachers (explanatory variables) and effective years of learning 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.

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 (4) 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 $\Delta\tilde{y}$ from equation (5), divide each of the summed terms by N, the number of observations, and let N go to infinity. The resulting expressions are depicted in (5) and (6).

Note that

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

As discussed in section 4.2., we assume that all $\rho_{t,t'} \geq 0$; that $\rho_{t,t'}$ is decreasing in $|t - t'|$; and that $\rho_{t,t-1}$ is decreasing in t. This implies that $\rho_{2,3} \leq 1$; $\rho_{2,4} \leq \rho_{3,4} = \rho$; $\rho_{1,3} \leq 1$; $\rho_{1,4} \leq \rho$, which in turn implies that $\frac{\rho_{2,4}+\rho_{2,3}}{1+\rho} < 1$ and $\frac{\rho_{1,3}+\rho_{1,4}}{1+\rho} < 1$. That is, $\hat{\beta}_4 + \hat{\beta}_3$ provides a lower bound on the cumulative effect $\alpha_4 + \alpha_3\gamma_{3,4} + \alpha_2\gamma_{2,4} + \alpha_1\gamma_{1,4}$.

Finally consider the expression for the probability limit of $\hat{\beta}_4$. This can be rewritten as

(A4) $$\begin{aligned}\text{plim } \hat{\beta}_4 &= \alpha_4 + \alpha_2\gamma_{2,4}\left(\frac{\rho_{2,3}\rho + \frac{\text{cov}(\Delta x_4, v_{2,3})}{\text{var}(\Delta x)} - \rho_{2,3}\rho}{1-\rho^2}\right) + \alpha_1\gamma_{1,4}\left(\frac{\rho_{1,3}\rho + \frac{\text{cov}(\Delta x_4, v_{1,3})}{\text{var}(\Delta x)} - \rho_{1,3}\rho}{1-\rho^2}\right) \\ &= \alpha_4 + \alpha_2\gamma_{2,4}\left(\frac{\text{cov}(\Delta x_4, v_{2,3})}{\text{var}(\Delta x)(1-\rho^2)}\right) + \alpha_1\gamma_{1,4}\left(\frac{\text{cov}(\Delta x_4, v_{1,3})}{\text{var}(\Delta x)(1-\rho^2)}\right),\end{aligned}$$

That is $\hat{\beta}_4$ provides a biased estimate of α_4 with the bias depending on the size and sign of $\text{cov}(\Delta x_4, v_{t,3})$, where $v_{t,3}$ is the residual in a regression of Δx_t on Δx_3, with $t = 1,2$.

C. *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

(A5) $$V = \sigma_\epsilon^2 E(\Delta\tilde{x}'\Delta\tilde{x})^{-1}/N = \sigma_\epsilon^2/N \begin{pmatrix} \sigma^2_{\Delta x_4} & \rho_{4,3}\sigma^2_{\Delta x_3} & \rho_{4,2}\sigma^2_{\Delta x_2} & \rho_{4,1}\sigma^2_{\Delta x_1} \\ \rho_{4,3}\sigma^2_{\Delta x_3} & \sigma^2_{\Delta x_3} & \rho_{3,2}\sigma^2_{\Delta x_2} & \rho_{3,1}\sigma^2_{\Delta x_1} \\ \rho_{4,2}\sigma^2_{\Delta x_2} & \rho_{3,2}\sigma^2_{\Delta x_2} & \sigma^2_{\Delta x_2} & \rho_{2,1}\sigma^2_{\Delta x_1} \\ \rho_{4,1}\sigma^2_{\Delta x_1} & \rho_{3,1}\sigma^2_{\Delta x_1} & \rho_{2,1}\,\sigma^2_{\Delta x_1} & \sigma^2_{\Delta x_1} \end{pmatrix}^{-1},$$

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 $\sigma^2_{\Delta x_i}$ is the population variance of $\Delta x_t \; \forall t = 1, \ldots, 4$.

We now show that each of the terms in the product can be written as a function of known population moments. Consider first $E(\Delta\tilde{x}'\Delta\tilde{x})$: After imposing the usual stationarity assumption on the test score differences across subjects, we can write this as

(A6) $$E(\Delta\tilde{x}'\Delta\tilde{x}) = \sigma^2_{\Delta x_4} \begin{pmatrix} 1 & \rho_{4,3} & \rho_{4,2} & \rho_{4,1} \\ \rho_{4,3} & 1 & \rho_{3,2} & \rho_{3,1} \\ \rho_{4,2} & \rho_{3,2} & 1 & \rho_{2,1} \\ \rho_{4,1} & \rho_{3,1} & \rho_{2,1} & 1 \end{pmatrix},$$

which can be estimated by replacing $\sigma^2_{\Delta x_4}$ and $\rho_{4,3}$ by their sample analogues $\frac{\sum_i \Delta\tilde{x}_{i,4}{}^2}{N}$ and $\hat{\rho}$, the coefficient estimate from equation (8), and setting the remaining $\rho_{t,t'}$ equal to their values in the particular parameterization under consideration.

Next consider the error variance, σ_ϵ^2. If we could observe $\Delta\tilde{x}_2$ and $\Delta\tilde{x}_1$, 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 do not observe the test scores 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

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

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

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

Hence, we get

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

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 $B_4 = \alpha\gamma^2\left(\frac{\rho_{2,4}-\rho_{2,3}\rho}{1-\rho^2}\right) + \alpha\gamma^3\left(\frac{\rho_{1,4}-\rho_{1,3}\rho}{1-\rho^2}\right)$ and $B_3 = \alpha\gamma^2\left(\frac{\rho_{2,3}-\rho_{2,4}\rho}{1-\rho^2}\right) + \alpha\gamma^3\left(\frac{\rho_{1,3}-\rho_{1,4}\rho}{1-\rho^2}\right)$ are the asymptotic biases on the OLS estimates of $\hat{\beta}_4$ and $\hat{\beta}_3$.

Once again imposing the stationarity condition on the test score differences across subjects, the asymptotic error variance is,

(A10) $$\sigma_\epsilon^2 = \sigma_\mu^2 - \left(B_4^2 + B_3^2 + \alpha^2\gamma^4 + \alpha^2\gamma^6 + 2B_4B_3\rho_{4,3} - 2B_4\alpha\gamma^2\rho_{4,2} - 2B_4\alpha\gamma^3\rho_{4,1} - 2B_3\alpha\gamma^2\rho_{3,2} - 2B_3\alpha\gamma^3\rho_{3,1} + 2\alpha^2\gamma^5\rho_{2,1}\right)\sigma_{\Delta x_4}^2.$$

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, $\sigma_{\Delta x_4}^2$ with $\frac{\sum_i \Delta\tilde{x}_{i,4}{}^2}{N}$, $\rho_{4,3}$ with $\hat{\rho}$, and setting the remaining $\rho_{t,t'}$ equal to their values in the particular parameterization under consideration.

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).

D. *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 (A14) into equations (A12) and (A13) 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 (4), 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 (A15)-(A16). 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 A1, 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.

E. *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 $x_{l_{r'},o_{i'},h_{s'},d_{j'}}$ from teacher $l_{r'}$ in origin school $o_{i'}$ to teacher $h_{s'}$ in destination school $d_{j'}$ is

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

where $K_{h_{s'},d_{j'}} - K_{l_{r'},o_{i'}}$ is the difference in knowledge between the two teachers.
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 $h_{s'}$ in destination school $d_{j'}$ is $\sum_{i=1}^{m} \sum_{r=1}^{T_{l,o_i}} x_{l_r,o_i,h_{s'},d_{j'}}$ (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 $\sum_{i=1}^{m} \sum_{r=1}^{T_{l,o_i}} x_{l_r,o_i,h_{s'},d_{j'}} + PTR(d_{j'})$. For the teacher's existing students, learning is therefore reduced by

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

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

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

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 (either existing or new ones from the m origin schools) in the classroom of teacher $h_{s'}$ in destination school $d_{j'}$ is

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

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

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

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

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

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

The first order condition with respect to $x_{l_{r'},o_{i'},h_{s'},d_{j'}}$ for an interior solution is

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

F. 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 2015 are not available but available evidence suggest that the number of teachers has increased to at least 3.0 million (UNICEF, 2015).

Using the SDI data, we derive an age distribution of teachers. 45% (or 1,351,400 teacher assuming there are currently 3 million primary school teachers) was hired in the last decade. 675,700 teachers in the decade before that, 540,560 teachers in the decade before that, and 432,448 teachers in the decade before that. Assuming teachers are working for 40 years on average, this implies that in the next 10 years 432,450 teachers will retire, in the following 10 years 540,560 teachers will retire, in the following 10 years 675,700 teachers will retire, and in the following 10 years 1,351,400 teachers will retire. We smooth these data to derive annual retirement numbers for 2015-2050.

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

To derive projections for the number of teachers to be hired per year (starting from 2016), we also need to determine the student-teacher ratio for 2015-2050. We start by choosing a student-teacher ratio, 43, in 2015 to match the net-enrollment for that year as reported by World Bank (2017). We further assume it will stay at 43 for 5 years, then 42 in the next 5 years, then 41 in the next 5 years, and 40 for the remaining years. The number of teachers to be hired a given year t is then simply the sum of the teachers that retired that year and the increase in the size of the cohort (of primary school students) divided by the student-teacher ratio.

We assume all new teachers start teaching in the first grade and continue to follow the class (cohort of students) up to grade 6. Thereafter they start again teaching in grade 1. Thus if t_{2016} is the number of teachers hired in 2016, and all are qualified, there will be t_{2016} new teachers (qualified teachers) teaching in grade 1 in 2016, t_{2016} qualified teachers teaching grade 2 in 2017, …, t_{2016} qualified teachers teaching grade 6 in 2021. In 2022, there will be $t_{2016} + t_{2022}$ qualified teachers in grade 1, and so forth.

Table A1: Relationship between student and teacher content knowledge: CRE model

	(1)	(2)	(3)	(4)
Dep. variable	IRT test scores		Effective years of learning	
	Mathematics	Language	Mathematics	Language
Implied $\beta_{4,subject}$	0.052**	0.072**	0.023	0.035**
	(.024)	(.027)	(.016)	(.018)
Implied $\beta_{3,subject}$	0.029	0.049**	0.050***	0.034**
	(.024)	(.025)	(.015)	(.017)
$\chi^2(\beta_{4,k} = \beta_{4,k'})$	0.52		0.32	
	[.47]		[.57]	
$\chi^2(\beta_{4,k} = \beta_{4,k'})$	0.64		0.54	
	[.42]		[.46]	
$\chi^2(\pi_{4,k} = \pi_{4,k'})$	1.59		3.21*	
	[.21]		[.08]	
$\chi^2(\pi_{3,k} = \pi_{3,k'})$	0.16		0.08	
	[.19]		[.77]	
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 learning) 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 ($\pi_{t,k} = \pi_{t,k'}$). 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 A2: Variance ratio test: Current and prior teacher knowledge

	Panel A: Teacher test score: IRT scaled			
Sample	All teachers		Same teacher in both subjects	
Variable	Current teacher	Prior teacher	Current teacher	Prior teacher
Mean	-0.090	-0.045	-0.412	-0.341
Std dev	0.943	0.941	0.871	0.878
Obs	16,363		8,896	
F-test statistic	1.005		0.983	
Pr(F<f)	0.617		0.211	
Pr(F≠f)	0.766		0.422	
Pr(F>f)	0.383		0.789	
	Panel B: Teacher test score: Effective years of education			
Sample	All teachers		Same teacher in both subjects	
Variable	Current teacher	Prior teacher	Current teacher	Prior teacher
Mean	3.187	3.032	2.575	2.465
Std dev	1.920	1.915	1.714	1.717
Obs	17,294		8,969	
F-test statistic	1.005		0.996	
Pr(F<f)	0.632		0.434	
Pr(F≠f)	0.736		0.867	
Pr(F>f)	0.368		0.566	

Note: Summary statistics of current and prior teacher content knowledge and tests on the equality of standard deviations (variances) of the two knowledge variables. Sample: Unit of observation is a student. All teachers sample is sample with data for all teachers. Same teacher in both subjects sample is sample with the same teacher teaching both subjects. F-test statistic is the test statistic for testing the ratio of the standard deviation (variance) of the current and prior teacher knowledge, with p-values for the three alternative hypotheses.

Table A3: 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 A4: 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
Observations	4,466	4,466	4,466	6,970	3,895
Number of schools	619	619	619	1037	537

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 A5: 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)
Specification/Sample	Placebo	Comparison	Same class teacher in grade 3 & 4	Different class teachers in grade 3 & 4
Observations	1,669	1,6691	1,773	3,901
Number of schools	217	217	277	525

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.