The Lost Human Capital

Teacher Knowledge and Student Achievement in Africa*

Tessa Bold, Deon Filmer, Ezequiel Molina, Jakob Svensson

Draft June 2017

We use unique data from representative surveys from seven countries representing a substantial proportion of the population of sub-Saharan Africa to assess the effect of teacher content knowledge on student outcomes. After four years of schooling, the majority of students fail to master tasks covered in the second year curriculum. This result, adjusting for the cumulative process of knowledge acquisition, imperfect persistence in learning between grades and measurement error, can partly be explained by the fact that many teachers struggle with tasks that their students should master in lower primary. Had all students been taught by teachers deemed to master the lower secondary curriculum—a minimum official criterion in the countries in the sample—our estimates show that students would have acquired on average one more year of curriculum adjusted human capital after four years. The study highlights the huge shortcomings in teacher quality in Africa and the lost human capital as a consequence.

* 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 and David Strömberg for extensive comments and suggestions. 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 (ASER, 2013) and roughly half of the students in Uganda, after three years of mathematics teaching, cannot order numbers between 0-100.

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 and how teachers perform along these dimensions.

In Bold et al. (2017), we quantify teacher quality in Africa along three core quality dimension: Time spent teaching, teachers' knowledge of the subject they are teaching, and teachers' pedagogical skills. Here we take the next logical step and attempt to causally assess the impact on student learning of one of these core quality dimensions: teachers' subject content knowledge.

Using unique data of over 5,000 teachers and 20,000 students 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—we first document how far along the official curriculum children have progressed after almost four years of schooling; i.e., how many years of curriculum based, or quality adjusted, human capital they have acquired. We then present a simple statistical model of cumulative knowledge acquisition, accounting for imperfect persistence in learning between grades, measurement error and unobserved teacher effects, and exploit within-student within-teacher variation to estimate both 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.

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; i.e. in teachers' curriculum based human capital, increases students' quality adjusted human capital by 0.12 SD in the short run (after one year of schooling). This implies that moving a student from the 5th to the 95th percentile of the teacher human capital distribution increases students' human capital by 0.72 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; Rothstein 2010; and Andrabi et al., 2011).

Using the estimated structural parameters—the parameter capturing the contemporaneous effect of teacher content knowledge on student learning and the persistence in learning parameter—we then calculate the learning achievements of students in a series of counterfactual experiments, varying the assumptions on the correlation of teacher knowledge over time. Specifically, we quantify how much more human capital students would have acquired if they had been taught by teachers that master the lower secondary curriculum—a minimum official criterion in the countries in the sample—but also a level of knowledge few primary school teachers pass.

We find that had students been taught by teachers deemed to master the minimum official criterion, they would have acquired on average 1.25 year more of curriculum adjusted human capital, implying 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 at a low level), close the observed human capital gap after almost four years by a half.

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., 2014; Araujo et al., 2016; and Bau and Das, 2017). Our results confirm the importance of teachers, although using a different empirical methodology, and demonstrate the external validity of the value added findings to a low income environment where both the average quality of teacher 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 findings also complement a growing experimental literature on teacher quality, showing that that teacher effort, broadly defined, can be raised, by providing financial incentives tied either to attendance or student performance (Duflo et al., 2012; Muralidharan and Sundararaman, 2011), or by exploiting the operation of dynamic incentives; i.e., contract teacher programs (Bold et al., 2013; Muralidharan and Sundararaman, 2013; Duflo et al., 2015), leading to improved learning outcomes which can

be substantial.[1] Here we focus on teacher content knowledge, noting that there is an important evidence gap in the lack of well-identified studies on the impact of teacher knowledge on learning outcomes in developing countries (see Glewwe and Muralidharan, 2015).

Another common finding in the value added literature is that standard teacher characteristics—experience, education, among others—explain very little of the differences in teacher quality. This may not be so surprising, especially in a developing country setting, since these characteristics appear largely uncorrelated with key dimensions of teaching many would argue are important, such as the effort exerted, or motivation, of the teacher (Chaudhury et al., 2006), teacher's knowledge of the subject she is teaching and how well she teaches that subject (Bold et al., 2017). More recently, however, research on teacher effectiveness has examined characteristics that are typically not collected by school administrators, including content knowledge (Rockoff et al., 2011, Bau and Das, 2017) and measures of how teachers organize teaching and provide instructional support (Kane and Staiger, 2012; Araujo et al., 2016). Specifically, Bau and Das (2017) find that higher content knowledge is associated with significantly and quantitatively larger positive effect on teacher value added. Their preferred estimates imply that a 1 SD increase in teacher test scores raise student test scores by 0.07 SD. We find, exploiting within-student within-teacher variation to estimate the causal effect of student knowledge an effect size of similar magnitude (0.12 SD) in the short run.

Methodologically, our identification strategy extends the within student across subject comparisons exploited in Dee (2005; 2007) and the within-student within-teacher across subjects comparison proposed by Metzler and Woessmann (2012).[2] Unlike Metzler and Woessmann (2012), however, who use a "contemporaneous" specification relating grade six student achievement test score measure to contemporaneous teacher score measure, we exploit the fact that we have access to test score data for both the current teachers and previous year's teacher to estimate both the contemporaneous effect of teacher content knowledge on student achievement as well as the extent of fade out of the teachers' impact in

[1] There is also a growing literature on the provision of supplemental remedial education, and on automated teaching through computer-aided learning programs or scripted lesson plans, reviewed in, for example, Glewwe and Muralidharan (2015), showing large effects (in the former case), or more mixed results (in the latter case) on student learning outcomes, that at least indirectly speaks to the importance of teacher quality.

[2] Clotfelter et al. (2010) and Lavy (2015) also exploit within-student across subject variation to assess impact of teacher credentials and instruction time, respectively, on student achievement, while Bietenbeck et al. (2017) exploit within-student within-teacher comparison to assess the impact of teacher knowledge and textbook provision on student test scores.

earlier grades.[3] We then combine these estimates to assess the cumulative effect of teacher knowledge on student achievement. We also use newly collected representative micro data from several countries. 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, we use data collected using one-on-one tests (for students) and test scores (for teachers) derived from mock student tests marked by teachers. Finally, given the design of the two tests, we can map each teacher's (and student's) knowledge onto grade-specific (curriculum) knowledge and thus estimate how far students have progressed after four years of studies, and what level of the curriculum teachers master.

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; i.e. we quantify how far students have progressed after four years of schooling and what grade level their teachers master. In section 4 we present a statistical model of cumulative knowledge acquisition, describe how we attempt to estimate the causal effect of teacher knowledge on student performance, and discuss the identifying assumptions. Section 5 presents specification and placebo tests followed by the main results. Section 6, finally, 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

[3] As noted in Todd and Wolpin (2003), the underlying education production technology in the contemporaneous (within-student within-teacher) specifications is justified if students are taught by the same teacher throughout their schooling history or that the effect from previous years teaching completely fades out or that current teacher skills are unrelated to previous teacher skills. None of these justifications, however, are likely to empirically valid in most settings.

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 five 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 (see Bold et al., 2017, for details of the sample).

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. In Bold et al. (2017), we 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, as well as summary statistics on a set of school characteristics collected through the observational approach discussed above (see Bold et al., 2017, for details). Teachers, on average, are absent from class 44% of the time and about half of that classroom absence is due to teachers not at all being at the school during regular teaching ours. As a results, 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 49 minutes). Pedagogical knowledge is low, with one in ten teachers deemed to have minimum pedagogy knowledge, and even fever 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 be taught in the official language of their country (English in Nigeria, English and Swahili in Kenya, Tanzania and Uganda, French in Senegal

and Togo, and Portuguese in Mozambique). In each school, the students' current and previous language and mathematics teacher were selected for testing.[4]

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 the language of instruction (English, French, or Portuguese), 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]

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 normal activities—namely, marking student work. Second, by not testing teachers in the same way as students are tested, 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 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).

3. Teacher and student knowledge

[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.

Raw test score data, such as the fraction correct, are usually transformed into some specific test score measure. In many cases, including in well-known publicly available data set, this transformation is based on test specific, and largely arbitrary scales (see Jacob and Rothstein, 2016) and effect sizes are measured in standard deviations of these transformed data.[6] As normalizations of student test score removes some of the underlying information in the raw data, and as the standard deviation of a test score can be sensitive to the range of question difficulty on the test, we instead construct our main test score measure using the school curriculum as a yardstick.[7] That is, we use the raw scores to determine the grade level of proficiency of students and teachers and label the grade level of proficiency as effective human capital years acquired. Appendix A provides details on how these effective years of human capital are defined.

The transformation into human capital years, we argue, makes sense here since the test covered test items from the Grade 1 up to the end of primary school and since the samples by construction are nationally representative. The use of human capital years as test score measure has at least three advantages. First, the transformed data points are informative in themselves. Second, using human capital years as test score measure enables us to compare students and teachers using the same scale. Third, it allows one to extrapolate beyond human capital levels observed in the sample in a meaningful way. As a robustness test, however, we also report and compare the results using the curriculum based human capital measure with two other transformations of the raw scores; the fraction correct on the student and teacher test, standardized within each country in the case of the student test, and scores rescaled using item response analysis.

Table 2 and Figure 1 show the percentage of students at each human capital level for language and mathematics. On average, students have 1.5 years of quality-adjusted human capital 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 human

[6] The National Assessment of Educational Progress (NAEP) test program in the US, for example, which assesses students in grades 4, 8, and 12, reports scale scores, ranging from roughly 100 to 400 with standard deviations around 30 (Jacob and Rothstein, 2016). The NAEP also reports discrete proficiency categories (basic, proficient, and advanced). The Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) test scores of grade 6 students, with data collected in 1995, 2000, and 2007, report scale scores with a mean of 500 and standard deviation 100 across students participating in the second wave (2000).

[7] We use the Kenyan curriculum as the benchmark. This is likely to not cause any comparability concerns as the subject test items included in the tests were validated against a larger set of Sub-Saharan African primary curricula to ensure comparability.

capital in mathematics after three and a half years of schooling, while the average student in Mozambique (the bottom performer) has acquired only 0.4 years of human capital (see Figure 2).

Breaking down the summary score, 27% of fourth grade students have not acquired any human capital in language and one-third have not acquired any human capital in mathematics. 15% and 21% have acquired one year of human capital in language and mathematics, respectively. Less than a quarter of the students have three years or more of human capital in language and mathematics.

Table 2 shows how the average scores and scores constructed with item response theory line up with the quality adjusted human capital years. There is a strong positive correlation between the different definitions of the scores, above 90% in all cases. While the raw score does not take the difficulty of the questions into account, the IRT and the human capital adjusted years of education are two different methods to scale scores by difficulty. The IRT 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 human capital adjusted years of education classify a question as easy or difficult based on where on the curriculum it is located. It thus speaks to the validity of the adjusted human capital years scores to see a high correlation between the two measures.

In Table 3 and Figure 3, we show the percentage of teachers at each quality adjusted year of education. On average, teachers have 3.5 years of quality adjusted education in language and 3.7 years of quality adjusted education in mathematics. Again there are large differences across countries, with Kenyan mathematics teachers (on average) having 5.7 quality adjusted years of education (top performer) and mathematics teachers in Togo having only 1.9 years of quality adjusted education (see Figure 4).

Looking at the distribution, strikingly, just over a third of the teachers master at best the second grade curriculum and 90% of teachers have quality adjusted education at or below five years in language and at or below six years in mathematics. Also in the case of teacher knowledge, there is a high correlation between the different score aggregates (see Table 4).

4. A statistical model

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 then use it to highlight the assumptions we make, given the structure of our data, in order to estimate the causal effect of teacher quality on student learning.

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 a school-supplied inputs, parent-supplied inputs, and teacher-supplied inputs, some of which could vary by subject and some 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-invariants inputs. 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, again with a subject-specific component and a general component.

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 production 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 three main problems we face in estimating the causal 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. Finally, teacher content knowledge, like student achievement, is based on test scores, and it is well known that such test are inherently noisy measures of the teachers' true knowledge. We deal with these issues in three ways. First, we exploit variation in student and teacher knowledge within students and within teachers. Second, we add additional structure on the relationship between teacher content knowledge across grades and their effect on student test scores and utilize the fact that we have test score data both for students' current teachers and the teachers who taught them in previous year. Third, we

correct for measurement error analytically after estimating the internal reliability of each test based on item response theory. Below we describe these steps in more details.

4.1. Within-student within-teacher variation

Linearizing the production function and exploiting the fact that we observe test scores for two subjects, k and k', we can take the first-difference version of equation (1). If the coefficients on all inputs are subject-invariant; i.e., a one unit increase in, for example, teacher content knowledge has the same marginal effect on test scores in subject k as in subject k', this within-student transformation ensures that all subject-invariant unobserved heterogeneity at the school and parent level is removed. If we further restrict attention to students who were taught by the same teacher in both subjects (in a given year) over the last two years, we can further remove any teacher-specific, subject-invariant heterogeneity each of these years, yielding the following specification for fourth year students,

$$\Delta y_{ij4} = y_{ij4,k} - y_{ij4,k'} = \alpha_0 + \alpha_4 \Delta x_{ij4} + \alpha_3 \Delta x_{ij3} + \alpha_2 \Delta x_{ij2} + \alpha_1 \Delta x_{ij1} + \epsilon_{ij} \quad (2)$$

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 unobservable inputs; i.e.,

$$\epsilon_{ij} = \Delta\omega_{ij} + \sum_{t=1}^{4}\left[\theta_t^S \Delta s_{ijt} + \theta_t^P \Delta p_{ijt} + \theta_t^T \Delta c_{ijt|f_k=f_{k'}}\right] + \sum_{t=1}^{2} \theta_t^{T\prime} \Delta c_{ijt|f_k\neq f_{k'}} + \Delta\varepsilon_{ijt} \quad (3)$$

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. With access to complete content knowledge data on teachers, estimation of the within-student, within-teacher specification (2) with OLS would recover the causal effect of teacher knowledge on student performance if the variation in teacher knowledge across subjects (Δx_{ijt}) and the error term (ϵ_{ij}) are orthogonal. As this is our key identifying assumption we consider next what it implies.

Consider first the differenced ability/motivation term $\Delta\omega_{ij}$ in (3). 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 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.[8] 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.

There are reasons to believe that the potential bias, if any, due to the omission of variation in teacher skills and characteristics across subjects that arise from the student not being taught by the same teacher in language and mathematics in year 1 and 2, $\sum_{t=1}^{2} \theta_t^{T\prime} \Delta c_{ijt|f_k \neq f_{k'}}$, in equation (2), will be small. Specifically, the structure of primary school in the countries we study is such that students tend to have one class teacher, who teaches both language and mathematics in lower primary (grades 1-3) and tend to have subject teachers, who specialize in either language or mathematics, as they move to upper primary (grades 4-6). In other words, if a student is taught by a class teacher in both mathematics and language in grade three, it makes it likely, we would argue, that the student was taught by a class teacher who teaches both subjects also in year 1 and 2. In fact, in the data, 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. This estimate likely constitutes a lower bound on how common it is to be taught by a class teacher in grades 1 and 2 conditional on having a class teacher in grade 3 since subject specialization tends to increase from grade 4 onwards. If all students that have class teachers in grade 3 also have class teachers in grade 1 and 2 (though not necessarily the same teacher as in grade 3), the within student transformation of the data, and the restriction of the sample to students who have only one teacher in both

[8] Using data from kindergarten students in Ecuador—a middle income country—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 our identifying assumption here is even weaker. We assume parents do not respond to differential (across subjects) differences in the quality of the teacher.

subjects in grade 3 (and in grade 4) implies that the nuisance term $\sum_{t=1}^{2} \theta_t^{T\prime} \Delta c_{ijt|f_k \neq f_{k\prime}}$ disappears.

Finally, consider next the variation in $\Delta c_{ijt|f_k = f_{k\prime}}$; i.e., in other (unobservable) characteristics or skills that vary by subject (for a given 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, something we would argue is unlikely, 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.

4.2. Teacher and student knowledge over time and measurement error

In equation (2), the difference in student learning outcomes in grade 4, denoted by Δy after dropping student, school, and time subscripts, is a function of the students' teachers' knowledge over their full schooling history. In the data, however, we do not observe Δx_2 and Δx_1. Thus, instead of equation (2), what we can estimate with the data is

$$\Delta y = \beta_0 + \beta_4 \Delta x_4 + \beta_3 \Delta x_3 + \mu \,. \tag{3}$$

Clearly, even under the identifying assumption discussed above, i.e., $\text{cov}\left(\Delta x_{ijt}, \epsilon_{ij}\right) = 0$, the coefficients in (3) will be biased as long as there is correlation between subject differences in teacher test scores over time and their effect on current student scores does not decay completely. Moreover, our measure of teacher content knowledge is an inherently "noisy" measures of true teacher achievement, thus introducing an additional source of bias. Nevertheless, if we provide additional structure on how subject differences in test scores evolve over time, and the source of the measurement error, we can recover both 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 and combine these estimates to assess the cumulative effect of teacher knowledge on student achievement.

To see this, we start by introducing the possibility that student learning, as measured by test scores, partly fades out over time.[9] Specifically, we parameterize imperfect persistence of achievement in the statistical model by assuming that test scores decay

[9] Recent research, using data from both developed and developing countries, suggest that student learning, as measured by test scores, fades rapidly. Kane and Staiger (2008); Jacob, Lefgren, and Sims (2010); and Rothstein (2010), for example, show that teacher effects dissipate by between 50-80% over one year. Similar patterns are also observed in a number of education experiments (see Andrabi et al., 2011, for a discussion and references).

geometrically. We further assume that the contemporaneous effect of teacher knowledge is constant across years, $\alpha_j = \alpha$, that is, the effect of teacher knowledge is not affected by the age at which it is applied. Finally, we allow for the fact that teacher test scores are measured with classical measurement error; i.e.

$$x_{t,k} = x^*_{t,k} + \xi_{t,k}, \tag{4}$$

where $x^*_{t,k}$ is the teacher's true content knowledge, and where $\text{cov}(x^*_{t,k}, \xi_{t',k}) = 0 \; \forall t, t'$, and $\text{cov}(\xi_{t,k}, \xi_{t',k}) = 0 \; \forall t \neq t'$ if the student has different teachers in the two periods. With these assumptions, we can rewrite (2) as

$$\Delta y = \alpha_0 + \alpha \textstyle\sum_{t=1}^{4} \gamma^{4-t} (\Delta x_t - \Delta \xi_t) + \epsilon \tag{5}$$

where γ—the parameter that links achievement across periods—captures the degree of persistence, and where, for convenience, we have now dropped all i and j subscripts.

To recover the contemporaneous effect α and the decay parameter γ in equation (5), define the following linear projection of current teacher test scores on test scores of previous (or future) teachers:

$$\Delta x_t = \delta_0 + \; \delta_{t,t-a} \Delta x_{t-a} + v_t \; \forall t, a, \tag{6}$$

Below we will apply different assumptions to this process to overcome the omitted variable bias.

Finally, we make the following auxiliary assumptions that are not crucial, but simplify the algebra somewhat. Specifically, we assume stationarity of the test score distribution and the measurement error process over time. With these assumptions, $\text{var}(\Delta x_t) = \text{var}(\Delta x_{t'}) = \text{var}(\Delta x)$, and consequently $\delta_{ij} = \delta_{ji}$, and $\text{var}(\Delta \xi_t) = \text{var}(\Delta \xi_{t'}) = \text{var}(\Delta \xi)$. Thus the OLS estimator β_4 and β_3 in (3) can be written as (see derivation in the appendix):

$$\hat{\beta}_4 = \alpha \left(1 - \frac{\text{var}(\Delta\xi)}{\text{var}(\Delta x)}\right) + \alpha\gamma^2 \, \frac{\delta_{24} - \delta_{23}\delta}{1-\delta^2} \left(1 - \frac{\text{var}(\Delta\xi)}{\text{var}(\Delta x)}\right) + \; \alpha\gamma^3 \, \frac{\delta_{14} - \delta_{13}\delta}{1-\delta^2} \left(1 - \frac{\text{var}(\Delta\xi)}{\text{var}(\Delta x)}\right), \tag{7}$$

and

$$\hat{\beta}_3 = \alpha\gamma \left(1 - \frac{\text{var}(\Delta\xi)}{\text{var}(\Delta x)}\right) + \alpha\gamma^2 \, \frac{\delta_{23} - \delta_{24}\delta}{1-\delta^2} \left(1 - \frac{\text{var}(\Delta\xi)}{\text{var}(\Delta x)}\right) + \; \alpha\gamma^3 \, \frac{\delta_{13} - \delta_{14}\delta}{1-\delta^2} \left(1 - \frac{\text{var}(\Delta\xi)}{\text{var}(\Delta x)}\right). \tag{8}$$

where $\delta = \delta_{34} = \delta_{43}$.

To make progress, will now consider the expressions in (7) and (8) under three different scenarios: Two polar cases in which there is either no correlation or perfect correlation between teacher knowledge in grade 3 and 4 and previous teachers' knowledge, and an intermediate, and in our view most natural, case in which there is some, but not

perfect, correlation between the observed and the unobserved test scores. The polar cases enable us to provide an upper and a lower bound on the effect of teacher knowledge on student knowledge after four years of primary school.

The first polar case corresponds to a situation in which all teachers change after year 2 and there is no correlation between differences in knowledge across subjects for any two teachers that are not the same. In that case $\delta_{ij} = \delta_{ji} = 0 \; \forall i \leq 2$ and $\forall j \geq 3$ and $\delta = \delta_{34} = \delta_{43}$ as before. While we do not deem this case to be a realistic description of the data generating process, it is nevertheless useful for estimating an upper bound. In that case, the estimated coefficients do not suffer from omitted variable bias by assumption and are related to the structural parameters of interest as follows:

(9a) $$\hat{\beta}_4 = \alpha \left(1 - \frac{\text{var}(\Delta\xi)}{\text{var}(\Delta x)}\right)$$

(10a) $$\hat{\beta}_3 = \alpha\gamma \left(1 - \frac{\text{var}(\Delta\xi)}{\text{var}(\Delta x)}\right)$$

The second polar case corresponds to a situation which seems to describe the structure of schooling in lower primary in some countries, namely, students are taught by the same teacher in lower primary (i.e. grade 1, 2 and 3) and then some students change teacher only once they reach grade 4. In that case $\delta_{ij} = \delta_{ji} = 1 \; \forall i,j < 4$ and $\delta = \delta_{4,3} = \delta_{3,4}$. This provides a lower bound on the contemporaneous and persistent effect of teaching because it maximizes the "bias" on $\hat{\beta}_3$. Again, under this assumption, the coefficients on the teacher knowledge variables do not suffer from omitted variable bias, only from collinearity in the regressors. All that is necessary for solving for the coefficients of interest are the functional form assumptions about the education production function (and of course the identification assumptions). In this case:

(9b) $$\hat{\beta}_4 = \alpha \left(1 - \frac{\text{var}(\Delta\xi)}{\text{var}(\Delta x)}\right)$$

(10b) $$\hat{\beta}_3 = \alpha\gamma \left(1 - \frac{\text{var}(\Delta\xi)}{\text{var}(\Delta x)}\right) + \alpha\gamma^2 \left(1 - \frac{\text{var}(\Delta\xi)}{\text{var}(\Delta x)}\right) + \alpha\gamma^3 \left(1 - \frac{\text{var}(\Delta\xi)}{\text{var}(\Delta x)}\right)$$

Finally, we explore an intermediate and in our view most reasonable case, namely, that every year a fraction of teachers is reallocated to different students, and that the fraction of teachers that is reallocated is the same in each year. Together that implies that if a share z of students stay with the same teacher as they move from year 3 to year 4, then z^2 students have the same teacher in grade 4 as they had in grade 2 and so on. If in addition, we assume that the unobservables of those teachers who are reallocated are not correlated with subject

differences in test scores of previous (or future) teachers we can write $\delta_{ij} = \delta_{ji} = \delta^{|i-j|}$ and the two expressions above simplify to:

(9c) $$\hat{\beta}_4 = \alpha\left(1 - \frac{\text{var}(\Delta\xi)}{\text{var}(\Delta x)}\right)$$

(10c) $$\hat{\beta}_3 = \alpha\gamma\left(1 - \frac{\text{var}(\Delta\xi)}{\text{var}(\Delta x)}\right) + \alpha\gamma^2\,\delta\left(1 - \frac{\text{var}(\Delta\xi)}{\text{var}(\Delta x)}\right) + \alpha\gamma^3\,\delta^2\left(1 - \frac{\text{var}(\Delta\xi)}{\text{var}(\Delta x)}\right)$$

Hence, in each of the three cases, including a measure of past teacher test scores is sufficient to recover the contemporaneous effect of teaching (after adjusting for measurement error) because the coefficient on past teacher test scores captures all the omitted variable bias.

There are four unknowns in the three (variants of) equations (6), (9) and (10); α, γ, δ, and $\text{var}(\Delta\xi)$. While it is usually impossible to obtain estimates of the variance of the measurement error, $\text{var}(\Delta\xi)$, in the case of test scores, there exists an underlying theory of measurement, item response theory, that tells us how the measured test score and the latent trait of interest, namely teacher knowledge, are related and from which we can recover an estimate of the variance of the measurement error. Specifically, we use the inverse of the test information function as an estimate of the variance of measurement error (see appendix).[10]

Thus, given $\text{var}(\Delta\xi)$, we are left with three parameters, α, γ, and δ, that can be uncovered from the estimated coefficients $\hat{\beta}_3$, $\hat{\beta}_4$, and $\hat{\delta}$. In particular, α can be solved from the equation for $\hat{\beta}_4$ alone, and is thus independent of the precise assumption about the correlation between included and omitted teacher knowledge variables.[11] Second, note that the ratio of $\hat{\beta}_4/\hat{\beta}_3$ results in a cubic equation for γ that is independent of measurement error, but does depend on the nature of the omitted variable bias. In sum, controlling for past teacher knowledge, the contemporaneous effect of teacher knowledge is robust to the various forms of omitted variable bias and collinearity considered here, but its cumulative effect is not.

4.3. Inference

To make inference about the structural parameters of interest, we need to construct their standard errors, given by the square root of the diagonal of the following matrix

(11) $$\sum \frac{\varepsilon_i\varepsilon_i}{N}\,(\Delta x^{*\prime}\Delta x^{*})^{-1} \times (1 + (n-1)\rho)\,,$$

[10] Our main outcome variable is years of curriculum based human capital, which in essence is a transformation of tests scores using Item Response Theory (IRT). Thus, we transform the variance of the measurement error using the relationship $\text{var}(Y) = \text{var}(a + bX) = b^2\text{var}(X)$, where Y is years of human capital and X is tests scored using IRT.

[11] In the empirical estimation, we also adjust $\hat{\beta}_4$ for the fact that test scores are measured half way through year 4, while the structural model measures the impact of teaching after a full year of teaching in each grade.

where the first term is an estimate of the error variance, the vectors in the $(\Delta x^{*\prime}\Delta x^{*})^{-1}$ matrix are given by $\Delta x^{*} = \{\Delta x_4^*, \Delta x_3^*, \Delta x_2^*, \Delta x_1^*\}$, with x_t^* being the teacher's true content knowledge (demeaned, so we can disregard the constant) defined in equation (4), and the last term is a parametric adjustment for clustering of the standard errors at the school level (the Moulton factor), with ρ measuring the intraclass correlation of human capital years for the students. In the appendix we describe how to construct each of these terms.

5. Results

In Table 6, we begin to explore the relationship between teacher and student knowledge as measured by their scores on the respective tests. In column (1) of the table, we regress student achievement (quality adjusted years of education or human capital years) on teacher subject knowledge (quality adjusted years of education or human capital years) controlling only for a set of country fixed effects and find a large positive association. In column (2), we also include past teacher knowledge (also in human capital years), resulting in a fall in the estimated coefficient on current teacher knowledge by 30 percent (as can be seen by comparing with the findings in column (3) which reports the results of the contemporaneous specification on the sample with both current and past teachers; i.e. the sample used in column (2)).

In column (4), we introduce student fixed effects to control for sorting of students to schools (or teachers) on the basis of subject invariant characteristics, as well as other unobserved student, and subject invariant, characteristics. In this specification, the effect of teacher knowledge on student test scores can only be driven by differences between the two subjects. The results indicate that part of the large association in column (1) is indeed driven by better students sorting into better schools, but the point estimate is still large and significant.

Note that the effect of teacher content knowledge depends on both β_4 and β_3; i.e., on both the current and previous teachers' content knowledge. For example, in the first scenario consider above, where all teachers are assumed to change after year 2 and there is no correlation between differences in knowledge across subjects for any two teachers that are not the same, the effect of two years of teaching, abstracting from measurement errors, is simply $\beta_4 + \beta_3$. We therefore also report the estimate of the sum of these two coefficients in the table.

It is still possible that the association reported in column (4) is driven by teaching activities that vary across subjects because students are not taught by the same teacher. Hence, in column (5), our preferred specification, we also introduce teacher fixed effects for the current and previous teacher. The results change only slightly, suggesting that other activities and skills

of teachers that do not vary across subjects are not correlated with the variation in teacher subject knowledge in a way that would affect the estimate.

To recap the identification assumptions stated in Section 5, two things have to be true for our preferred specification to be interpreted as causal: (i) there must not be other factors (at teacher level or otherwise) that drive both student and teacher subject differences in knowledge; (ii) there is no sorting by students and teachers 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 7 suggesting that neither of these assumptions is likely to be violated. In column (1), we start by reporting our main specification; i.e. the specification also reported in column (5), Table 6. In columns (2)-(4), we then examine whether the difference between language and mathematics scores might be driven by a common underlying factor. 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) in the first differenced specification. As can be seen, compared to the main specification reported in column (1), the estimates change only marginally.[12]

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 vary across subjects for a given teacher, we have already seen that results remain basically unchanged when we restrict the sample to those students who have the same teacher in both subjects (Table 6, columns (4) and (5)). That in itself suggests that unobserved teacher behavior and skills are unlikely to confound the estimates, and it also presents us with an opportunity to test this more directly. Specifically, in column (4) of Table 7, we repeat the student fixed effects specification reported in column (4), Table 5, thus including also students taught by different teachers in language and mathematics in the sample, and add teachers' score on a lesson preparation exercise that was administered to all teachers as an additional

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

explanatory variable.[13] While pedagogy knowledge has a positive effect on student learning, the coefficients of interest are only marginally affected by the inclusion of this variable. Hence, we would argue that unmeasured differences in teacher skills—at least pedagogical skills—for the same teacher are unlikely to matter given that they do not even matter when teachers are different across subjects.

To further test for sorting across (and within) schools we report the results of two specifications where we constrain the sample to only include schools in rural areas—where the choice of schools to attend for students are more limited (column 5)—and schools with only one classroom (column 6); thus effectively ruling out sorting into different classes within schools. While the estimate on current teacher knowledge falls somewhat and the estimate on previous teacher slightly increases, the joint effects remain largely unchanged in these smaller samples.

To further bolster the causal interpretation, we present a set of placebo test in line with Chetty et al. (2014). Column (7) uses the subject differences of test scores of teachers in higher grades as an additional control. The argument is as follows: if there is purely a sorting relationship between subject differences of student and teacher test scores in the school, then the teacher test scores in other grades should also be correlated with student test scores. Hence, including such test scores should change the coefficient on current and previous teacher test scores if sorting is taking place, but not, if the effect is causal. As seen by comparing columns (7) and (8) of Table 7, both using the same sample of students, the coefficients on current and previous teacher test score are unchanged by the inclusion of teacher test scores in higher grades, which themselves have an insignificant and essentially zero effect. Second, if the relationship between teacher and student knowledge is purely due to sorting, than the length of exposure to a given teacher should not matter. We test this in columns (9) and (10), where we compare the coefficient on current teacher knowledge for those who have kept their grade 3 teachers in grade 4 and those who changed teacher. The coefficient is almost twice as large in the first case, implying that length of exposure indeed matters.

Table 8 presents the main findings based on the reduced form specification reported in Table 7, column (1). Unadjusted (not factoring in measurement errors) and measurement adjusted estimates of the contemporaneous effect of teacher content knowledge on student learning (α) and the persistence in learning parameter (γ) are reported under the three different scenarios in panel A and panel B, respectively. Under our preferred specification; i.e., when we

[13] The assessment of pedagogy knowledge and skills as part of the Service Delivery Indicators is described in Bold et. al. (2017).

assume that each year a fraction of teachers is reallocated to different students, and that the fraction of teachers that is reallocated is the same in each year, the unadjusted contemporaneous effect is 0.054, implying that being taught by a teacher with one more year of human capital would increase student learning by roughly half a month after one year.[14] Adjusted for measurement error, the estimate of the contemporaneous effect increase by a forty percent to 0.09. To put this number in context with other findings, a 1 SD increase in teacher human capital years increases student learning by 0.12 standard deviations.

The persistence of learning (γ) across years can be backed out by taking the ratio of β_4 and β_3 and finding the positive root of the resulting cubic equation in γ. Depending on the specification, persistence is estimated to lie between 0.5 and 0.88, with the lower values obtained assuming either perfect correlation (implying persistence of 0.5) or an AR (1) process (persistence of 0.6). These numbers are consistent, albeit on the higher end, of what has been reported using data from Pakistan and the US (see Kane and Staiger 2008; Jacob, Lefgren, and Sims 2010; Rothstein 2010; and Andrabi et al., 2011).

As shown in Section 4, the estimate of the contemporaneous effect of teacher knowledge on student learning is robust to the various assumptions made about the correlation of teacher knowledge over time, though it is sensitive to measurement error. In contrast, the cumulative effect after four years of primary school changes depending on the nature of the correlation between the included and excluded teacher test scores: this is the case because the more the coefficient on past teacher test scores, $\hat{\beta}_3$, is confounded by omitted variable bias (or collinearity between the test score variables), the lower the resulting estimate of persistence (for a given $\hat{\beta}_4$), thus reducing the total cumulative effect.

Given that students lag behind 2.5 years already after four years of primary school and their teachers do not master the primary curriculum, what do these results imply for policy reforms designed to combat the learning crisis?

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

[14] The contemporaneous effect is larger than the reduced form effect in Table 7, column (1), because we adjust for the fact that α measures the effect of an input after one year, while we test students half way through the year.

We now ask how many years of human capital students would accumulate after four years if teachers' years of quality adjusted education rose to the lower secondary level, the minimum entry requirement for teaching—and thus equaled the number of years most of them spent in school. This policy experiment is equivalent to an increase of 6.5 years of teachers' human capital relative to the current average of 3.5 years. Extrapolating the quasi-experimental results in our preferred specification (Table 9, Panel B, column (1)) this would lead to an increase of students' human capital of a year and a quarter after four years, implying that raising teacher content knowledge to the lower bar for primary teachers in Africa would in itself, holding teacher effort and pedagogical skills constant, close the observed human capital gap after almost four years by a half.[15] Even in our most conservative specification, students would still have accumulated one more year of human capital, reducing their knowledge shortfall by 40%.

These results apply holding other dimensions of teacher quality, such as effort and skill, constant. As we show in Bold et al. (2017), there are also shortfalls along these dimensions: for example teachers are absent from classroom roughly half of the scheduled teaching time, and addressing them would be an important part of any policy reform. In contrast, reforms that focus purely on teacher knowledge and training would require teachers in Sub-Saharan Africa to complete 16 years of education, similar to or even exceeding the requirements for teacher training in most high-income countries.[16]

Finally, in Panel C of Table 8, we also present robustness of our findings with respect to the definition of the student and teacher knowledge variables. In particular, we present the contemporaneous and cumulative effect of teacher knowledge on student learning when constructing the dependent and independent variables using item response theory rather than the human capital adjustment. The reduced form coefficients for the effect of current and past teacher knowledge are 0.03 and 0.5 respectively, almost identical to the effects estimated when using human capital years of education. As a result, the point estimates for the cumulative effect of education after four years are rather similar, though the actual magnitudes of the IRT results are about a third smaller when expressing both sets of effects in standard deviations.

[15] This result is arrived at by multiplying the cumulative effect of four years of teaching in the third row of Panel B in Table 9 by the number of human capital years required to increase from the current average (3.5 years) to the minimum requirement (10 years).

[16] 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.

6. Discussion

Recent estimates suggest that differences in (the quality) of human capital can explain a dominant share of the world income differences (Caselli and Coleman, 2006; 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 in 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 obvious implications for both policy and research. On the later, there are few, if any, well-identified studies on how to effectively improve teacher knowledge and skills and the impact thereof (Glewwe and Muralidharan, 2015). Our results strongly suggest that this evidence gap is important to address.

Overall, our findings highlight the importance of improving teacher quality in low income countries. Importantly, the continued rapid expansion of new teachers—two million new teachers are anticipated to be hired in the next 15 years in Sub-Saharan Africa alone—ought to provide amble opportunities to do so. Thus, although it may be costly, and difficult, to systematically and significantly raise the quality of existing stock of teachers, a focus on how to ensure that the next cohort of teachers is better prepared to teach well, and rewarded for doing so when deployed, can potentially go a long way to improve outcomes. Related interventions, that either supplements the current teachers with additional instructors, or automate certain aspects of teaching using computer-aided learning programs or scripted lesson plans, also show promising results (Murnane and Ganimian, 2014; Glewwe and Muralidharan, 2015; 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.

Bau Natalee and Jishnu Das. 2016. “The Misallocation of Pay and Productivity in the Public Sector: Evidence from the Labor Market for Teachers.” Working Paper, 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* (forthcoming).

Caselli, Francesco, and Wilbur John Coleman. 2006. “The World Technology Frontier.” American Economic Review, 96(3): 499-522.

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. 2014. “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.

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.

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.

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.

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

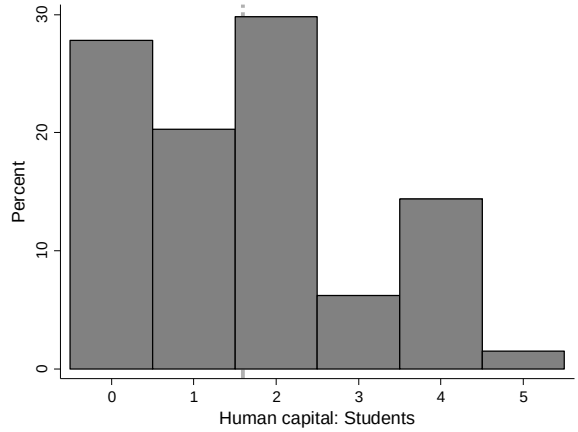
Murnane, Richard J. and Alejandro J. Ganimian. 2014. “Improving Educational Outcomes in Developing Countries: Lessons from Rigorous Evaluations.” NBER Working Paper No. 20284.

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.

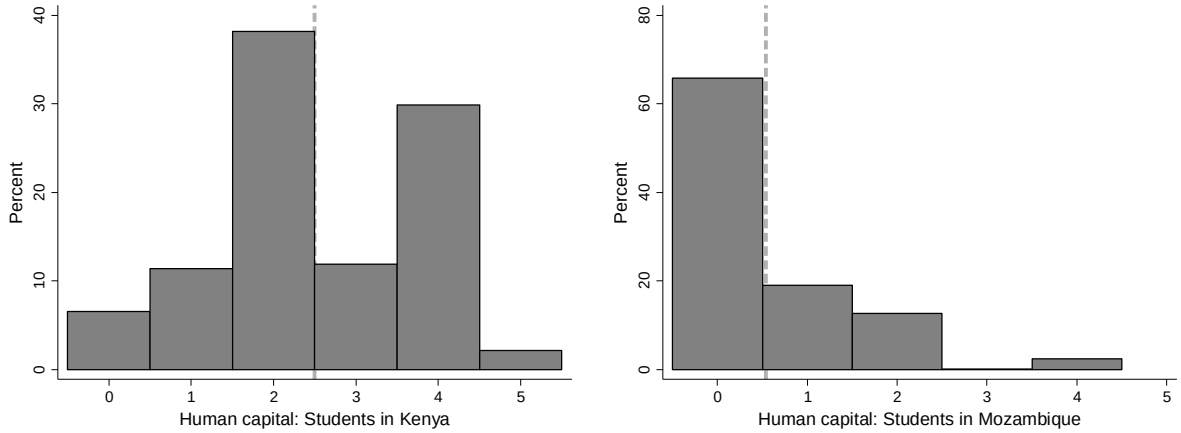
Petra E. Todd and Kenneth I. Wolpin, 2003, “On the Specification and Estimation of the Production Function for Cognitive Achievment”, *Economic Journal*, 113, pp.F3-F33.

Figure 1: Years of human capital after four years of schooling



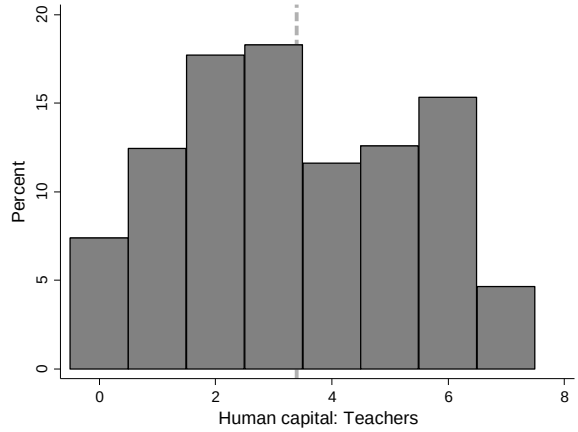
Note: Distribution of quality-adjusted human capital years of education for students (pooled data across countries and subjects). Dashed vertical line depicts mean.

Figure 2: Years of human capital after four years of schooling: Kenya and Mozambique



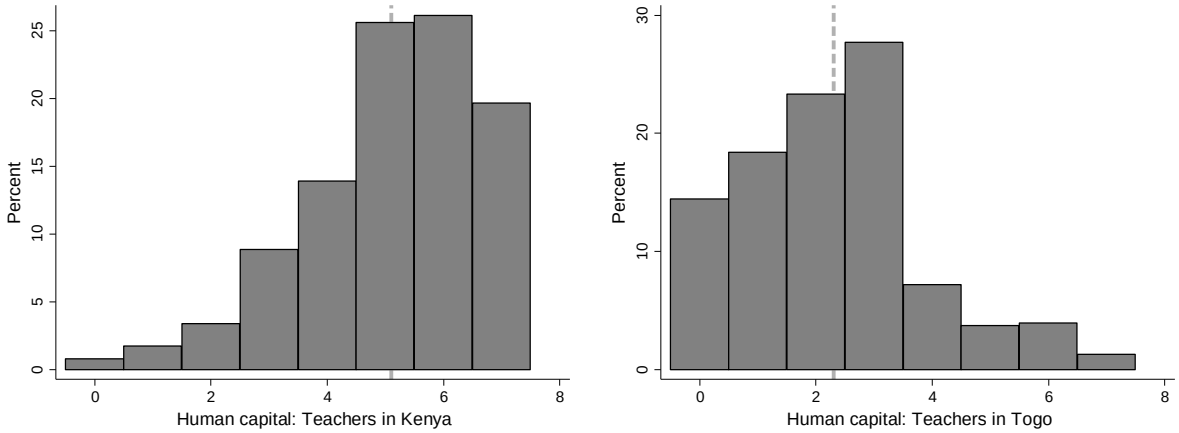
Note: Distribution of quality-adjusted human capital years of education for students in Kenya and Mozambique (pooled data across subjects for each country). Dashed vertical lines depict means.

Figure 3: Teachers' years of human capital



Note: Distribution of quality-adjusted human capital years of education for teachers (pooled data across countries and subjects). Dashed vertical line depicts mean.

Figure 4: Teachers' years of human capital: Kenya and Togo



Note: Distribution of quality-adjusted human capital years of education for teachers in Kenya and Togo (pooled data across subjects for each country). Dashed vertical lines depict means.

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 49mins
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: Distribution of quality-adjusted years of education for students

Human capital years (curriculum based)	Language	Mathematics
0	27%	36%
1	15%	21%
2	45%	18%
3	5%	6%
4	8%	17%
5	n/a	3%
N	23,884	23,016

Note: Distribution of quality-adjusted years (Human capital years) of education for students. 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.

Table 3: Quality-adjusted years of education, Item Response Scores (IRT), and average scores for students

	Language		Mathematics	
Human capital years (curriculum based)	Average	IRT	Average	IRT
0	12%	-1.19	24%	-0.99
1	23%	-0.56	43%	-0.06
2	58%	0.15	55%	0.49
3	93%	0.98	65%	0.92
4	98%	1.48	67%	0.94
5	-	-	83%	1.57
N	23,884	23,884	23,016	23,016

Note: Item Response Scores (IRT) and average scores for students, conditional on quality-adjusted years of education (Human capital years). 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.

Table 4: Distribution of quality-adjusted years of education for teachers

Human capital years (curriculum based)	Language	Mathematics
0	8%	8%
1	7%	16%
2	19%	14%
3	22%	14%
4	18%	6%
5	23%	2%
6	2%	28%
7	3%	12%
N	4755	4970

Note: Distribution of quality-adjusted years (Human capital years) of education for teachers. 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.

Table 5: Quality-adjusted years of education, average scores and Item Response Scores (IRT) for teachers

	Language		Mathematics	
Human capital years (curriculum based)	Raw	IRT	Raw	IRT
0	20%	-1.57	4%	-1.99
1	33%	-0.94	26%	-0.77
2	42%	-0.38	40%	-0.31
3	49%	-0.02	49%	0.08
4	54%	0.33	68%	0.57
5	65%	0.88	78%	0.92
6	72%	1.10	91%	0.74
7	82%	1.51	89%	1.13
N	4755	4755	4970	4970

Note: Item Response Scores (IRT) and average scores for teachers, conditional on quality-adjusted years of education (Human capital years). 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.

Table 6: Relationship between student and teacher content knowledge

	(1)	(2)	(3)	(4)	(5)
Dep. variable	Human capital years of education: Students				
Human capital years of current teacher	0.087***	0.067***	0.094***	0.027**	0.031*
	(.009)	(.013)	(.012)	(.013)	(.018)
Human capital years of previous year's teacher		0.049***		0.047***	0.046***
		(.012)		(.012)	(.016)
Language	0.148***	0.148***	0.124***	0.167***	0.216***
	(.019)	(.026)	(.027)	(.012)	(.033)
Constant	2.062***	1.927***	2.033***	1.221***	1.029***
	(.065)	(.088)	(.083)	(.042)	(.042)
Observations	30,361	17,294	17,294	17,294	8,969
Adj. R-squared	0.136	0.132	0.129	0.497	0.524
Number of schools	1,974	1,503	1,503	1,503	626
Number of students	16,922	10,324	10,324	10,324	4,503
Joint effect		0.116		0.074	0.077
		[.000]		[.000]	[.000]
Country FE	X	X	X		
Student FE				X	X
Same teacher in language and mathemathics					X

Note: Fixed effects specifications with clustered, by school, standard errors in parenthesis. Joint effect is the test of the sum of the coefficients on human capital years of current and previous teacher, with p-value in brackets. *** 1% , ** 5% , * 10% significance.

Table 7: Specification tests

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Dep. variable	Human capital years of education: Students									
Human capital years of current teacher	0.031* (.018)	0.033* (0.018)	0.030* (0.018)	0.025* (0.013)	0.017 (0.019)	0.021 (0.019)	0.046 (0.034)	0.047 (0.034)	0.072*** (0.027)	0.043*** (0.016)
Human capital years of previous teacher	0.046*** (.016)	0.039** (0.017)	0.049*** (0.016)	0.048*** (0.012)	0.050*** (0.017)	0.050*** (0.017)	0.072** (0.030)	0.074** (0.030)		
Teacher pedagogy Score				0.239 (0.175)						
Human capital years of higher grade teachers							0.005 (0.031)			
Constant		-0.216*** (0.030)	-0.216*** (0.033)	-0.165*** (0.026)	-0.181*** (0.035)	-0.186*** (0.036)	-0.277*** (0.066)	-0.275*** (0.066)	-0.117** (0.057)	-0.296*** (0.031)
Joint effect	0.077	0.072	0.079	0.073	0.068	0.071	0.118	0.121	-	-
p-value	[.000]	[.000]	[.000]	[.000]	[.001]	[.000]	[.001]	[.001]		
Specification:	Main	District FE	Urban FE	Student FE	Rural	One grade 4 class	Placebo	Comparison	Grade 3 and 4	Grade 4 only
Observations	4,466	4,466	4,466	6,970	3,626	3,895	1,201	1,201	1,773	3,901
Number of schools	619	619	619	1037	501	537	167	167	277	525

Note: 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 rural schools; (6) main specification on sample of schools with one grade 4 classroom; (7) main specification controlling for teacher content knowledge (human capital years) of higher grade teachers in school; (8) main specification using the same sample as in column (7); (9) Sample of students with the same teacher in both subjects in both years (grade 3 and grade 4); (10) Sample of students with a new teacher teaching both subject in grade 4. Joint effect is the test of the sum of the coefficients on human capital years of current and previous teacher, with p-value in brackets. *** 1% , ** 5% , * 10% significance.

Table 8: Structural parameters

	(1)	(2)	(3)
Panel A: Effect of teacher test score on student test score (human capital, unadjusted)			
Contemporaneous effect (α)	0.054*	0.054*	0.054***
	(0.030)	(0.032)	(0.019)
Persistence (γ)	0.627	0.501	0.879
	(0.562)	(0.451)	(0.616)
Total effect after4 years	0.122	0.101	0.180
	[0.000]	[0.000]	[0.000]
Panel B: Effect of teacher knowledge on student knowledge (human capital, adjusted)			
Contemporaneous effect (α)	0.088***	0.088***	0.088***
	(0.024)	(0.024)	(0.024)
Persistence (γ)	0.627*	0.501	0.879*
	(0.331)	(0.356)	(0.477)
Total effect after 4 years	0.198	0.165	0.291
	[0.000]	[0.000]	[0.000]
Panel C: Effect of teacher knowledge on student knowledge (IRT scores, adjusted)			
Contemporaneous effect (α)	0.086***	0.086***	0.086***
	(0.025)	(0.025)	(0.025)
Persistence (γ)	0.653	0.544	0.998
	(0.491)	(0.432)	(0.571)
Total effect after 4 years	0.202	0.171	0.343
	[0.000]	[0.000]	[0.000]
Specification			
Fraction replaced each year	X		
Same teacher (1-3)		X	
No correlation with previous			X

Note: Estimates of the contemporaneous effect (α) and persistence parameter (γ) under three different scenarios (varying assumptions about the correlation between teacher knowledge in grade 3 and 4 and previous teachers' knowledge). Panel A report unadjusted estimates while panel B report estimates adjusted for measurement error. In both cases using human capital years of teachers (explanatory variables) and students (dependent variable). Panel C: Adjusted for measurement errors, IRT scores. Specifications: Fraction replaced each year: The correlations between the differenced teacher knowledge between grades 3 and 2 and 2 and 1 are estimated, using equation (6), based on data from grades 3 and 4; Same teacher (1-3): Students are assumed to taught by the same teacher in lower primary (i.e. grade 1, 2 and 3); No correlation: No correlation between differences in knowledge across subjects for any two teachers that are not the same in grades 1 and 2. Total effect after 4 years is $\alpha \sum_{t=1}^{4} \gamma^{t-1}$.

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 half of an unprompted Cloze passage, as six years if they could correct more than half 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 in equation (3)

The OLS estimator β_4 in equation (3) is

$$\text{(A1)} \qquad \hat{\beta}_4 = \alpha + \alpha\gamma^2 \frac{\delta_{24} - \delta_{23}\delta_{34}}{1-\delta_{34}\delta_{43}} + \alpha\gamma^3 \frac{\delta_{14} - \delta_{13}\delta_{34}}{1-\delta_{34}\delta_{43}}$$
$$- \frac{\alpha}{1-\delta_{34}\delta_{43}} \frac{\text{var}(\Delta\xi_4)}{\text{var}(\Delta x_4)} + \frac{\alpha\delta_{34}\delta_{43}}{1-\delta_{34}\delta_{43}} \frac{\text{var}(\Delta\xi_4)}{\text{var}(\Delta x_3)} - \frac{\alpha\gamma\delta_{43}}{1-\delta_{34}\delta_{43}} \frac{\text{var}(\Delta\xi_3)}{\text{var}(\Delta x_4)} + \frac{\alpha\gamma\,\delta_{34}}{1-\delta_{34}\delta_{43}} \frac{\text{var}(\Delta\xi_3)}{\text{var}(\Delta x_3)} -$$
$$- \frac{\alpha\gamma^2\delta_{42}}{1-\delta_{34}\delta_{43}} \frac{\text{var}(\Delta\xi_2)}{\text{var}(\Delta x_4)} + \frac{\alpha\gamma^2\,\delta_{34}\,\delta_{32}}{1-\delta_{34}\delta_{43}} \frac{\text{var}(\Delta\xi_2)}{\text{var}(\Delta x_3)} - \frac{\alpha\gamma^3\delta_{41}}{1-\delta_{34}\delta_{43}} \frac{\text{var}(\Delta\xi_1)}{\text{var}(\Delta x_4)} + \frac{\alpha\gamma^3\,\delta_{34}\,\delta_{31}}{1-\delta_{34}\delta_{43}} \frac{\text{var}(\Delta\xi_1)}{\text{var}(\Delta x_3)}$$

,

which given our auxiliary assumptions becomes

$$\text{(A2)} \qquad \hat{\beta}_4 = \alpha\,(1 - \frac{\text{var}(\Delta\xi)}{\text{var}(\Delta x)}) + \alpha\gamma^2 \frac{\delta_{24} - \delta_{23}\delta}{1-\delta^2}(1 - \frac{\text{var}(\Delta\xi)}{\text{var}(\Delta x)}) + \alpha\gamma^3 \frac{\delta_{14} - \delta_{13}\delta}{1-\delta^2}(1 - \frac{\text{var}(\Delta\xi)}{\text{var}(\Delta x)}).$$

The OLS estimator β_3 is

$$\text{(A3)} \qquad \hat{\beta}_3 = \alpha\gamma + \alpha\gamma^2 \frac{\delta_{23} - \delta_{24}\delta_{43}}{1-\delta_{34}\delta_{43}} + \alpha\gamma^3 \frac{\delta_{13} - \delta_{14}\delta_{43}}{1-\delta_{34}\delta_{43}} - \frac{\alpha\,\delta_{34}}{1-\delta_{34}\delta_{43}} \frac{\text{var}(\Delta\xi_4)}{\text{var}(\Delta x_3)} + \frac{\alpha\delta_{43}}{1-\delta_{34}\delta_{43}} \frac{\text{var}(\Delta\xi_4)}{\text{var}(\Delta 4)}$$
$$- \frac{\alpha\,\gamma}{1-\delta_{34}\delta_{43}} \frac{\text{var}(\Delta\xi_3)}{\text{var}(\Delta x_3)} + \frac{\alpha\gamma\delta_{43}\delta_{43}}{1-\delta_{34}\delta_{43}} \frac{\text{var}(\Delta\xi_4)}{\text{var}(\Delta 4)} - \frac{\alpha\,\gamma^2\delta_{32}}{1-\delta_{34}\delta_{43}} \frac{\text{var}(\Delta\xi_2)}{\text{var}(\Delta x_3)} + \frac{\alpha\gamma^2\delta_{43}\delta_{42}}{1-\delta_{34}\delta_{43}} \frac{\text{var}(\Delta\xi_2)}{\text{var}(\Delta 4)} -$$
$$\frac{\alpha\,\gamma^3\delta_{31}}{1-\delta_{34}\delta_{43}} \frac{\text{var}(\Delta\xi_1)}{\text{var}(\Delta x_3)} + \frac{\alpha\gamma^3\delta_{43}\delta_{41}}{1-\delta_{34}\delta_{43}} \frac{\text{var}(\Delta\xi_1)}{\text{var}(\Delta 4)}$$

which, given the auxiliary assumptions becomes

$$\hat{\beta}_3 = \alpha\gamma(1 - \frac{\text{var}(\Delta\xi)}{\text{var}(\Delta x)}) + \alpha\gamma^2 \frac{\delta_{23} - \delta_{24}\delta}{1-\delta^2}(1 - \frac{\text{var}(\Delta\xi)}{\text{var}(\Delta x)}) + \alpha\gamma^3 \frac{\delta_{13} - \delta_{14}\delta}{1-\delta^2}(1 - \frac{\text{var}(\Delta\xi)}{\text{var}(\Delta x)})$$

C. *Estimate of the variance of the measurement error*

We estimate the underlying subject knowledge of each teacher (or their latent trait), φ, using a partial credit model (to account for the fact that not all items on the teacher test are scored as 0/1. The test information function is then given by the negative of the expectation of the second derivative of the log-likelihood function:

$$I(\varphi) = -E\left\{\frac{\partial^2}{\partial\varphi^2}\log L(B)\right\}$$

where B is the set of parameters of the partial credit model. The variance of the measurement error is given by the inverse of this function (StataCorp, 2015) and is estimated to be 0.214 for language and 0.391 for mathematics.

D. *Inference*

To estimate the standard errors of the structural parameters of interest, we first relate the squared sum of residuals estimated by the model to the `true' error variance. The residual from the estimated model is:

(A4) $$\hat{\varepsilon} = (\Delta y - \beta_4 \Delta x_4 - \beta_3 \Delta x_3)$$

Adding and subtracting the true model error, gives

(A5) $$\hat{\varepsilon} = \varepsilon - (\Delta y - \ \alpha \sum_{t=1}^{4} \gamma^{4-t} [\Delta x_t - \Delta \xi_t]) - \ (\Delta y - \beta_4 \Delta x_4 - \beta_3 \Delta x_3),$$

Where all terms are demeaned so we can disregard the constant. The sum of squared residuals is then equal to

$$\sum \hat{\varepsilon}_i \hat{\varepsilon}_i = \sum \varepsilon_i \varepsilon_i + (\alpha \lambda \Delta x_4 + (\alpha\gamma\lambda - \alpha\gamma^2\delta(1-\lambda) - \alpha\gamma^3\delta^2(1-\lambda))\Delta x_3 + \alpha\gamma^2 \Delta x_2 + \alpha\gamma^3 \Delta x_1 - \alpha \sum_{t=1}^{4} \gamma^{t-1} \Delta \xi_t)^2,$$

where $\lambda = \frac{\text{var}(\Delta\xi)}{\text{var}(\Delta x)}$. That is, the variance of the true error equals the estimate of the residual variance plus an adjustment term, which is a function of the structural parameters and the cross-products $\text{cov}(\Delta x_t, \Delta x_{t'})$ and $\text{cov}(\Delta\xi_t, \Delta\xi_{t'})$. Given our assumptions on how test scores and measurement error evolve through time; i.e., with constant variance, we can easily estimate the adjustment term and therefore back out the variance of the error term.

Secondly, we note that the elements of

$$(\Delta x^{*\prime} \Delta x^*) = N \times \begin{bmatrix} \text{var}(\Delta x_4^*) & \text{cov}(\Delta x_4^*, \Delta x_3^*) & \text{cov}(\Delta x_4^*, \Delta x_2^*) & \text{cov}(\Delta x_4^*, \Delta x_1^*) \\ \text{cov}(\Delta x_3^*, \Delta x_4^*) & \text{var}(\Delta x_3^*) & \text{cov}(\Delta x_3^*, \Delta x_2^*) & \text{cov}(\Delta x_3^*, \Delta x_1^*) \\ \text{cov}(\Delta x_2^*, \Delta x_4^*) & \text{cov}(\Delta x_2^*, \Delta x_3^*) & \text{var}(\Delta x_2^*) & \text{cov}(\Delta x_2^*, \Delta x_1^*) \\ \text{cov}(\Delta x_1^*, \Delta x_4^*) & \text{cov}(\Delta x_1^*, \Delta x_3^*) & \text{cov}(\Delta x_1^*, \Delta x_2^*) & \text{var}(\Delta x_1^*) \end{bmatrix}$$

$$= N \times \begin{bmatrix} \text{var}(\Delta x_4) & \delta\text{var}(\Delta x_4) & \delta^2\text{var}(\Delta x_4) & \delta^3\text{var}(\Delta x_4) \\ \delta\text{var}(\Delta x_3) & \text{var}(\Delta x_3) & \delta\text{var}(\Delta x_3) & \delta^2\text{var}(\Delta x_3) \\ \delta^2\text{var}(\Delta x_2) & \delta\text{var}(\Delta x_2) & \text{var}(\Delta x_2) & \delta\text{var}(\Delta x_2) \\ \delta^3\text{var}(\Delta x_1) & \delta^2\text{var}(\Delta x_1) & \delta\text{var}(\Delta x_1) & \text{var}(\Delta x_1) \end{bmatrix} - N \times$$

$$\begin{bmatrix} \text{var}(\Delta\xi_4) & \delta\text{var}(\Delta\xi_4) & \delta^2\text{var}(\Delta\xi_4) & \delta^3\text{var}(\Delta\xi_4) \\ \delta\text{var}(\Delta\xi_3) & \text{var}(\Delta\xi_3) & \delta\text{var}(\Delta\xi_3) & \delta^2\text{var}(\Delta\xi_3) \\ \delta^2\text{var}(\Delta\xi_2) & \delta\text{var}(\Delta\xi_2) & \text{var}(\Delta\xi_2) & \delta\text{var}(\Delta\xi_2) \\ \delta^3\text{var}(\Delta\xi_1) & \delta^2\text{var}(\Delta\xi_1) & \delta\text{var}(\Delta\xi_1) & \text{var}(\Delta\xi_1) \end{bmatrix},$$

which we again estimate directly given the assumptions about the evolution of teacher scores and measurement error over time. Finally, the Moulton factor is estimated to be 2.35.